Optimal estimation of Gaussian mixtures via denoised method of moments

Yihong Wu and Pengkun Yang*

Working paper: March 15, 2018

Abstract

The Method of Moments is one of the most widely used methods in statistics for parameter estimation, obtained by 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 analysis of method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

In this talk I will present some recent results for estimating Gaussians location mixtures with known or unknown variance. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space before executing the method of moments. 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 Wasserstein distance via polynomial interpolation and marjorization, we establish the statistical guarantees and optimality of the proposed procedure. These results can also be viewed as provable algorithms for Generalized Method of Moments which involves non-convex optimization and lacks theoretical guarantees.

Contents

*Y. Wu is with the Department of Statistics and Data Science, Yale University, New Haven, CT, yihong.wu@yale.edu. P. Yang is with the Department of Electrical and Computer Engineering and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL, pyang14@illinois.edu.

1 Introduction

1.1 Gaussian mixture model

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

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

Here w_i is the mixing weight, μ_i is the location (center) and σ is the common standard deviation of each Gaussian component. In a Gaussian mixture model, we can equivalently write the distribution of an observation X as a convolution

$$X = U + \sigma Z, \tag{2}$$

where U is a discrete latent random variable taking value μ_i with probability w_i, and Z is independently standard normal. The goal of this paper is to learn Gaussian mixture models using n independent samples.

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 is further divided into the case 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 are more flexibility in designing the estimator but less interpretability.

- *Clustering*: estimate the latent label of each sample point (i.e. U_i if the ith sample is represented as $X_i = U_i + Z_i$) with small misclassification rate.

It is clear that to ensure the possibility of clustering it is necessary to impose certain separation conditions 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 models, that is, how many clusters there are. However, in the present literature, most of the estimation procedures are either classification-based, or rely on separation conditions in the analysis. Bridging this conceptual divide is one of the main motivations of the present paper.

The most popular heuristic procedure to estimate Gaussian mixture model parameters is the *Expectation-Maximization* (EM) algorithm [DLR77], an iterative procedure which is expected to converge the *Maximum Likelihood Estimate* (MLE), but no theoretical guarantee is known in general. Besides likelihood-based algorithms, there are other moment-based algorithms, such as the classical *method of moments* [Pea94] and generalizations. A detailed review of those algorithms in Gaussian mixture models is given in Section 1.5. Motivated by the failure of the vanilla method of moments as explained next, this paper provides an efficient moment-based algorithm as well as provable optimality guarantees.

1.2 Failure of the vanilla method of moments

The method of moments, commonly attributed to Pearson [Pea94], produces an estimator by equating the population moments to the sample moments and solving the system of moment equations. This method is conceptually simple but suffers from the following problems especially in the context of mixture models:

- *Solubility*: the method of moments entails solving a multivariate polynomial system, in which one frequently encounters non-existence or non-uniqueness statistically meaningful solutions.

- *Computation*: solving moment equations can be computationally intensive. For instance, for k-component Gaussian mixture models, the moments equations consist of a multivariate polynomial system of $2k-1$ equations and $2k-1$ variables.

- *Accuracy*: existing statistical literature on the method of moments [VdV00, Han82] either shows only consistency under weak assumptions, or proves central limit theorem assuming very strong regularity conditions (so that 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. [VdV00, Theorem 5.52] and [Kos07, 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, consider the following simple example:

Example 1. Consider a Gaussian mixture model with two unit variance components. 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:

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

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. With finite samples, there is always a non-zero chance of failure; even with infinite samples, when two components completely overlap, the failure happens with constant probability. To see this, note that the first two equations are equivalent to $\mathbb{E}_n[X] = \mathbb{E}[U]$ and $\mathbb{E}_n[X^2] - 1 = \mathbb{E}[U^2]$, and the vanilla method of moments fails as long as

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

which disobeys the Cauchy-Schwarz inequality. When two components completely overlap, the above equation is equivalent to $n(\mathbb{E}_n[Z^2] - \mathbb{E}_n^2[Z]) \le n$, whose left-hand side follows the χ^2-distribution with $n-1$ degrees of freedom. The failure happens with probability approaching to 0.5 as n diverges, according to the central limit theorem.

The main observation is that individually each moment estimate is $\frac{1}{\sqrt{n}}$-consistent, jointly they are not the moments of any distribution. The failure easily happens when the population moments are estimated with finite samples; even with infinite samples, it also provably happens with constant probability when the true model has fewer components than postulated, or equivalently, the population moments are on the boundary of moment space (the equivalence is established in Lemma 30).

1.3 Main results

In this paper, we propose the *denoised method of moments*, where the moments are jointly denoised through a projection step. The main procedure is simple: we first compute noisy estimates of moments, then project them into the moment space, and finally invoke the vanilla method of moments. This procedure resolves the three issues of the vanilla method of moments, as pointed out in Section 1.2, simultaneously:

- solubility: there uniquely exists a statistically meaningful solution;
- computability: the solution can be found through an efficient algorithm;
- accuracy: the solution is provably optimal, and the optimality is adaptive.

The denoising (projection) step is either explicitly carried out via convex optimization in Section 4.1, or implicitly used in analyzing Lindsay's algorithm [Lin89] in Section 4.2. After denoising, the solubility is guaranteed by classical theory on moments, as summarized in Section 2.1 and 2.1. The accuracy and optimality of our algorithm are introduced next.

The framework in this paper is different from most previous parameter estimation in mixture models: rather than individually recover parameters of each component, our goal is to jointly recover the latent discrete distribution consisting of those parameters. The main benefits of this framework include:

- Assumption-free: the task of recovering individual components require assumptions to ensure identifiability, such as strictly positive mixing weight and distinct components. In our framework, those unidentifiable instances have the same latent distribution to be estimated.
- Inference on the number of components: this framework allows us to deal with misspecified model and estimate the order of finite mixture.

This framework can be equivalently viewed as deconvolution from (2): the goal is to recover the distribution of U using observations following the convolved distribution.

In this framework, it turns out a useful and flexible loss function is the 1-Wasserstein distance (see Section 1.4 for a justification in the context of mixture models), defined by

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

which, in one dimension, coincides with the ℓ_1 distance between the cumulative distribution functions [Vil03]. Denote the underlying model as a convolution of ν and $N(0,\sigma^2)$, and the estimated model by $\hat{\nu}$ and $N(0,\hat{\sigma}^2)$. Our main results are the following Theorem 1. Moreover, those results are all minimax rate-optimal for constant k as shown in Section 5.

Theorem 1 (Optimal rates)**.** *Suppose $|\mu_i| \le M$ and σ is bounded for each component, and both k and M are given. If σ is also given, then, with probability at least $1-\delta$,*

$$W_1(\nu,\hat{\nu}) \le O_k\left(M\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k-2}}\right); \tag{3}$$

if σ is unknown, then

$$W_1(\nu,\hat{\nu}) \le O_k\left(M\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{4k}}\right), \tag{4}$$

and

$$|\sigma^2-\hat{\sigma}^2| \le O_k\left(M\left(\frac{n}{\log(1/\delta)}\right)^{-\frac{1}{2k}}\right), \tag{5}$$

where all hidden constants in O_k are at most polynomial in k.

Given separation conditions on the Gaussian mixture model, it is reasonable to expect a better rate of accuracy. For instance, when components are well-separated and each generated sufficient samples, a parametric rate is achievable. A weaker condition is that components form $k_0 \le k$ well-separated clusters, where $k-k_0$ quantifies the degree of overfitting [Che95, HK15]. Our estimator is *adaptive* to those benign situations, with accuracy given in Theorem 2. Additionally, the rate in (6) is also minimax optimal when k, k_0 and γ are constants, by the lower bounds in [HK15], and a proof by a simple extension of our method is given in Remark 4.

Theorem 2 (Adaptive rate)**.** *Under the conditions of Theorem 1, suppose k component centers form k_0 different clusters, where intercluster atoms are separated by at least γ and total weights of each cluster is at least $\epsilon/(0.1\gamma)$, with ϵ denoting the W_1 guarantee of Theorem 1 (or Theorem 6). If σ is given, with probability at least $1-\delta$,*

$$W_1(\nu,\hat{\nu}) \le O_k\left(M\gamma^{-\frac{2k_0-2}{2(k-k_0)+1}}\left(\frac{n}{\log(k/\delta)}\right)^{-\frac{1}{4(k-k_0)+2}}\right); \tag{6}$$

if σ is unknown, then

$$\sqrt{|\sigma^2 - \hat{\sigma}^2|},\ W_1(\nu, \hat{\nu}) \le O_k\left(M\gamma^{-\frac{k_0-1}{k-k_0+1}}\left(\frac{n}{\log(k/\delta)}\right)^{-\frac{1}{4(k-k_0+1)}}\right). \tag{7}$$

In density estimation, the goal is to recover the density function of the unknown mixture model. We accurately recovered the latent distribution by $\hat{\nu}$, so the density function $\hat{f}$ as the convolution of $\hat{\nu}$ and $N(0, \sigma^2)$ serves as a natural estimate. The estimated density $\hat{f}$ is evaluated in Theorem 3 under the χ^2-divergence, defined by $\chi^2(f\|g) = \int \frac{(f-g)^2}{g}$, which immediately implies other common distance measures such as Kullback-Leibler divergence, total variation, and Hellinger distance.

Theorem 3 (Density estimation)**.** *Under the conditions of Theorem 1, denote the density of the underlying model by f. If σ is given, with probability $1-\delta$,*

$$\chi^2(\hat{f}\|f), \chi^2(f\|\hat{f}) \le O_k(\log(1/\delta)/n).$$

1.4 Why Wasserstein distance?

Instead of evaluating the parameter estimates in Euclidean space (or in L_p space), we consider the loss function given by the Wasserstein distance between probability distributions. This is a natural criterion for recovering the latent discrete distribution, which is not too stringent to give trivial result (e.g. the Kolmogorov-Smirnov distance) and, at the same time, strong enough to yield guarantees on the atoms and weights under the usual assumptions. In fact, the commonly used criterion $\min_\pi \|\theta_i - \hat{\theta}_{\pi(i)}|$ is precisely the Wasserstein distance between two equally weighted distributions [Vil03].

Accurate parameters in ℓ_∞-norm. An accurate latent distribution under the Wasserstein distance implies accurate locations and mixing weights under additional necessary assumptions. This is discussed in detail in Section 2.3. As a concrete application, we demonstrate the following Lemma 1, which provides an accuracy of parameters up to a permutation under typical assumptions on Gaussian mixture models, including separations between components and lower bounds on mixing weights [Das99, KMV10, HP15].

Lemma 1. *Suppose a distribution ν has k atoms that are separated by at least ϵ_1 and each has at least ϵ_2 probability. If $\hat{\nu}$ has at most k atoms, and $W_1(\nu, \hat{\nu}) < \epsilon$ with $\epsilon < 0.1\epsilon_1\epsilon_2$, then $\hat{\nu}$ must have exactly k atoms, and there is a permutation π such that*

$$\|x - \pi\hat{x}\|_\infty < \epsilon/\epsilon_2, \quad \|w - \pi\hat{w}\|_\infty < O(\epsilon/\epsilon_1),$$

where (x, w) and $(\hat{x}, \hat{w})$ denote atoms and weights of ν and $\hat{\nu}$, respectively.

1.5 Related work

There exists a vast literature on mixture models, in particular Gaussian mixtures, and method of moments, from various perspectives and for a comprehensive review see []. 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 statistical inference, and it is also applicable in mixture models. In parameter estimation, under strong restrictions on the parameter space, the MLE in mixture models is shown to be consistent and asymptotically normal [RW84], following from the initial work of Cramér [Cra46] and a generalization by [Cha54]. Certainly, the restrictions implicitly require that the parameters are identifiable: only one instance is allowed in the parameter space among nonidentifiable parameters. In the case of nonidentifiable mixtures, the likelihood function is still almost surely maximized in a neighborhood of some parameter with an equal mixture density function [Red81, FM96], so the density of the mixture distribution obtained from the MLE is conceivably consistent. Indeed, a rate of convergence of the mixture density from the MLE is established in general mixture models in [VDG96] by analyzing the metric entropy despite without discussion on Gaussian mixture models, whose rate is obtained using the same methods in [GW00, GVDV01]; a rate is also obtained differently from the perspective of approximation by [LB00]. Without assuming a specific parametric form, nonparametric maximum likelihood estimation (NPMLE) in mixture models has also been studied, and its characterizations and properties are obtained in [Lai78, Lin81, Lin95], but few accuracy guarantees are available. Despite those nice properties, there are many problems to be solved in practice, for example, the likelihood is possibly unbounded and no maximizer exists, or the likelihood has multiple maxima; the estimate is usually obtained from zeros of likelihood equations, which only attains local maximum. Moreover, the maxima of the likelihood function, or zeros of the likelihood equation, are often computationally infeasible to find out. The computational burden of MLE is alleviated by the brilliant invention of the Expectation-Maximization algorithm, the most popular method by far for estimation of a latent distribution.

Expectation-Maximization (EM) algorithm [DLR77] is an iterative method to find the maximum of the likelihood function, and has been extensively studied in the application of Gaussian mixture models [RW84, XJ96]. This method converges to a local maxima of the likelihood function in general [RW84], but possibly suffers poor quality, and no guarantee exists that it converges to the global optima. In practice we need to employ different heuristic selections of initial guess [KX03] and stopping criterion [SMA00], and possibly different data augmentation techniques [PL01]. Computationally, its slow convergence rate is widely observed through empirical experiments [RW84, KX03]: it requires large number of iterations, especially when two components have large overlap. Additionally, instead of first summarizing into moments, this method accesses all raw data in each iteration, which is particularly expensive in the presence of large scale problems and large data sets these days.

The Gaussian mixture model also received considerable attention in Bayesian estimation, where a further prior is postulated on the latent distribution, and popular selections include Dirichelet process priors [Fer83, Esc94, EW95] and hierarchical or independent priors on means, variances, and mixing weights [DR94, RG97, RW97]. The practical performance depends on the prior and the heuristic choice of parameters of the prior. Moreover, the posterior distribution is often difficult to evaluate even with moderate sample sizes, and is usually approximated by Monte Carlo methods. Aforementioned literatures on Bayesian estimation aim for a suitable prior and a good simulation procedure for the posterior distribution. For specific priors, the posterior distribution on the latent distributions, as well as its consistency and rate of concentration, is analyzed in [IJS01, IZ02, Ngu13].

Moment-based methods. The most representative moment-based method is the vanilla method of moments (MM) [Pea94] as introduced in Section 1.2. The failure of the vanilla method motivates various modifications including minimum chi-square [Ber80] and the *generalized method of moments* [Han82].

Generalized method of moments (GMM), proposed in [Han82], instead of exactly solving the set of moment equations, aims to minimize the distance between the sample moments and the moments of the estimated model. This generic method is widely applied in many practical statistical problems such as econometrics [Hal05]. It enjoys various nice asymptotic properties, including asymptotic consistency and normality, and the asymptotic variance can be optimized by a suitable choice of weighting matrix [Han82]. Despite nice theoretical properties of the minimizer, the solution is found through generic optimization procedures, and thus is not guaranteed to converges to the global optimum in Gaussian mixture models, where the moment conditions are non-convex. Our denoised method of moments (DMM) proposed in Section 4.1, essentially under the same formulation as GMM, efficiently finds the global optimal parameters of Gaussian mixture models, and thus enjoys theoretical properties of GMM.

There are some more recent works studying the polynomial learnability of Gaussian mixture models [MV10, KMV10, BS10] and the optimal rate of accuracy [HP15] using moment-based methods. A key observation is that, under necessary assumptions for identifiability, a Gaussian mixture model with finite number of components is uniquely determined (up to a permutation of components) by its finite number of moments, so it is expected that estimated parameters are accurate if the model moments are close to estimated moments. For Gaussian mixtures in some fixed dimensions, [MV10, KMV10, BS10] rely on the exhaustive search over a sufficiently fine grid of parameters and the comparison with estimated moments, and it turns out a grid of polynomial size suffices, but this method is still too expensive in practice; on the real line, [HP15] follows a similar approach as [Pea94] by carefully analyzing a polynomial equation obtained from the usual method of moments, and shows the optimal rate $\Theta(n^{-1/12})$ in the case of two components, but this approach is difficult to generalize in the case of more components. The problems in higher dimensions are reduced to lower dimensions by many projections, with an extra step to match components estimated in each projection.

Other methods. From a frequentist perspective, the minimum distance estimation is studied by [Che95, HK15], where they consider the Gaussian mixture model that is closest to the empirical distribution in Kolmogorov-Smirnov (KS) distance. Since empirical distributions are $\sqrt{n}$-consistent in KS distance, the same rate of consistency holds for the minimum distance estimator. An upper bound of the W_1 distance between latent distributions using the KS distance between mixture distributions is obtained in [HK15] (which corrects the previous bound in [Che95]), and it yields their estimation accuracy. In particular, they studied overfitted mixture models, where a k-component mixture model is postulated while the true model has $k_0 \le k$ components. The local uniform rate of accuracy is $n^{-\frac{1}{4(k-k_0)+2}}$ asymptotically as n diverges around any fixed k_0-component Gaussian mixture model. However, their estimator is in general computationally infeasible. Our denoised method of moments is computationally efficient and adaptively achieves the same rate of accuracy as discussed later in Section 4.3.

1.6 Notations

A discrete distribution supported on k atoms is called a k-atomic. The r^{th} moment of a measure μ is denoted by $m_r(\mu)$. The moment matrix associated with a sequence of moments $m_0, m_1, \dots, m_{2r}$

is a Hankel matrix of size $r+1$ using those moments:

$$\mathbf{M}_r = \begin{bmatrix} m_0 & m_1 & \cdots & m_r \\ m_1 & m_2 & \cdots & m_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_r & m_{r+1} & \cdots & m_{2r} \end{bmatrix}. \tag{8}$$

The r^{th} moment matrix of a probability measure μ is the moment matrix associated with the moments of μ up to degree $2r$. For a random variable X, we also write $m_r(X)$ and $\mathbf{M}_r(X)$ to denote the r^{th} moment and the r^{th} moment matrix, respectively, of the distribution of X. The χ^2-divergence between two probability measures μ and ν is denoted by $\chi^2(\mu\|\nu) = \int \frac{(\mathrm{d}\mu-\mathrm{d}\nu)^2}{\mathrm{d}\nu}$, and for two random variables X and Y we also write $\chi^2(X\|Y)$ to denote the χ^2-divergence between their distributions. 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 two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, let $\langle x, y\rangle \triangleq \sum_i x_i y_i$. Unless stated otherwise, we always refer Z to an independent standard normal random variable.

2 Preliminaries

2.1 Moment space, SDP characterization, and Gauss quadrature

The study of moments has played a key role in the development of analysis, probability, statistics, and optimization. For a thorough treatment on this topic we refer to the readers to the classical monographs [ST43, KS53] and the recent works [Las09, Sch17]. Below, we briefly discuss a few results from the theory of moments that are essential throughout this paper.

The r^{th} moment space consists of vectors of the first r moments of all probability distributions supported on a given set. It is a convex body and satisfies many inequalities such as the Cauchy-Schwarz inequality. This convex set allows efficient characterization using moment matrices. Specifically, the space of first r moments of all probability distributions on an interval $[a, b]$ is completely characterized by two positive semidefinite conditions [ST43, Theorem 3.1] (see also [KS53] and the recent monograph [Las09]):

$$\begin{cases} \mathbf{M}_{0,r} \succeq 0, \quad (a+b)\mathbf{M}_{1,r-1} \succeq ab\mathbf{M}_{0,r-2} + \mathbf{M}_{2,r}, & \text{for } r \text{ even}, \\ b\mathbf{M}_{0,r-1} \succeq \mathbf{M}_{1,r} \succeq a\mathbf{M}_{0,r-1}, & \text{for } r \text{ odd}, \end{cases} \tag{9}$$

where $\mathbf{M}_{i,j}$ for $i \le j$ denotes the moment matrix associated with the sequence of moments $m_i, m_{i+1}, \ldots, m_j$ constructed by (8).

Example 2 (Moment space on $[0,1]$)**.** The space of the first two moments of distributions on $[0,1]$ is simply characterized by $m_1 \ge m_2 \ge 0$ and $m_2 \ge m_1^2$. The space of the first three moments of all distributions on $[0,1]$ is characterized by

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

Since positive semidefinite condition is equivalent to non-negativeness of all principal minors, we have the following specific inequalities for the first three moments:

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

Those constraints have the following natural interpretations: the first two conditions follow immediately from the range of distribution $[0,1]$, and the last two conditions follow from the Cauchy-Schwarz inequality. These necessary conditions for m_1, m_2, m_3 to be the moments of some distribution on $[0,1]$ turn out to be sufficient.

For a discrete distribution, its moments reveal more properties of the underlying distribution and have a stronger structure. The number of atoms of a discrete distribution is characterized by determinants of moment matrices, as presented in the following Theorem 4 (see [Usp37, p. 362] or [Lin89, Theorem 2A]). Moreover, a k-atomic distribution itself is uniquely determined by its first $2k-1$ moments. Though the moment space of k-atomic distributions is not necessarily convex, the quadrature rule introduced next shows that the space of first $2k-1$ moments of k-atomic distributions coincides with the $(2k-1)^{\text{th}}$ moment space of all distributions, and thus is convex and is also fully characterized by (9).

Theorem 4. *A sequence $m_1, \ldots, m_{2r}$ is the moments of a distribution with exactly r points of support if and only if* $\det(\mathbf{M}_{r-1}) > 0$ *and* $\det(\mathbf{M}_r) = 0$.

The quadrature rule finds a k-atomic representation for the first $2k-1$ moments of any distribution. Specifically, given a measure ν, its k-point quadrature rule consists a sequence of atoms $x_1, \ldots, x_k$ and the corresponding weights $w_1, \ldots, w_k$ such that

$$\int P(x)\mathrm{d}\nu(x) = \sum_{i=1}^{k} w_i P(x_i), \tag{10}$$

for any polynomial P of degree no greater than $2k-1$. It is known that, due to Christoffel [Chr77], such quadrature rule uniquely exists. The quadrature rule is completely determined by the first $2k-1$ moments of ν, as summarized in Algorithm 5, and improved algorithms include [Rut62] using quotient-difference, [GW69] using QR decomposition, [Gau68] through a suitable discretization, and [SD71, Gau70] from modified moments.

2.2 Polynomial interpolation, majorization, and the Neville diagram

Polynomial interpolation is a basic primitive in numerical analysis to approximate a given function. Given a function f and a set of $n+1$ distinct supports $x_1, \ldots, x_{n+1}$, there exists a unique polynomial of degree n that coincides with that function on those $n+1$ supports. The interpolating polynomial can be easily expressed by the Lagrange interpolation formula, which, however, is difficult to analyze since properties of the original function are hidden in the combination of individual Lagrange basis. An alternative approach is by Newton's interpolation formula using Neville's algorithm in terms of divided differences, which are closely related to derivatives:

$$P(x) = \sum_{i=1}^{n+1} f[x_1, \ldots, x_i] \prod_{j=1}^{i-1} (x - x_j), \tag{11}$$

where $f[x_i, \ldots, x_j]$ is the notation for divided difference that can be recursively obtained by $f[x_i, \ldots, x_j] = \frac{f[x_{i+1}, \ldots, x_j] - f[x_i, \ldots, x_{j-1}]}{x_j - x_i}$ with $f[x_i] = f(x_i)$.

A generalization of the above interpolation is the Hermite interpolation, which not only coincides with the function values but also the derivatives. Given a function f, there exists a unique polynomial P of degree n satisfying $P^{(k)}(x_i) = f^{(k)}(x_i)$ for $i = 1, \ldots, m+1$ and $k = 0, \ldots, n_i - 1$, where $n+1 = \sum_i n_i$. Analogous to the Lagrange interpolation formula, the polynomial can be expressed

by the generalized Lagrange basis, whose explicit formulas are given in [SB02]. The Newton's interpolation formula in (11) still applies after the following modifications: replace x_i with n_i copies of itself, and let $f[x_i, \dots, x_j]$ denote the generalized divided difference that $f[x_i, \dots, x_{i+k}] = \frac{1}{k!} f^{(k)}(x_i)$ when those points are identical that $x_i = \cdots = x_{i+k}$.

Example 3 (Neville's algorithm in Hermite interpolation)**.** Suppose we want to interpolate a step function $f(x) = \mathbf{1}_{\{x \le 0\}}$ with a polynomial satisfying

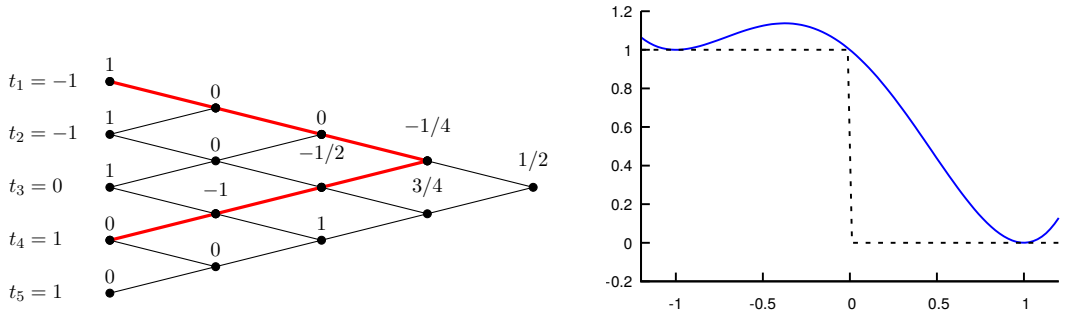
$$\begin{aligned} &P(-1) = 1, P'(-1) = 0, \\ &P(0) = 1, \\ &P(1) = 0, P'(1) = 0 \end{aligned}$$

Replacing each x_i by n_i copies of itself gives a sequence $-1, -1, 0, 1, 1$. Applying (11) yields that

$$\begin{aligned} P(x) = f[-1] + f[-1,-1](x+1) + f[-1,-1,0](x+1)^2 \\ + f[-1,-1,0,1]x(x+1)^2 + f[-1,-1,0,1,1]x(x+1)^2(x-1). \end{aligned}$$

Those divided differences can be obtained from the Neville's diagram in Fig. 1(a). The polynomial is $P(x) = 1 - \frac{1}{4}x(x+1)^2 + \frac{1}{2}x(x+1)^2(x-1)$, shown in Fig. 1(b).

In this example, the above Hermite interpolation gives a polynomial majorant to the step function. To see this, we note that $P'(\xi) = 0$ for some $\xi \in (-1, 0)$ by Rolle's theorem. Since $P'(-1) = P'(1) = 0$, the polynomial P has no other stationary point than $-1, \xi, 1$, and thus decreases monotonically in $(\xi, 1)$. Hence, it follows that $-1, 1$ are the only local minimum points of P, and $P \ge f$ everywhere.



(a) Neville's diagram.

(b) Hermite interpolation.

Figure 1: Neville's diagram and Hermite interpolation. In (a), each node is the divided difference of the two left nodes connecting to it. For example, the red thick line shows that $f[-1,-1,0,1] = -1/4$, which is the divided difference of $f[-1,0,1]$ and $f[-1,-1,0]$ and is calculated by $\frac{-1/2-0}{1-(-1)}$.

2.3 Wasserstein distance

Wasserstein distance measures the cost of optimal transportation, whose development has broad impact in economics and motivates the study of linear programming, and it also metrizes weak convergence in probability theory [Vil03]. Given the cost $c(x, y)$ to transport unit mass from x to y, the optimal transference plan between two probability measures μ and ν is a coupling π that minimizes the total cost $\mathbb{E}_\pi[c(X, Y)]$. In a particular case that $c(x, y) = \|x - y\|_p^p$ for $p \ge 1$,

Wasserstein distance is defined as $W_p^p(\mu,\nu) = \inf\{\mathbb{E}[\|X-Y\|_p^p] : X \sim \mu, Y \sim \nu\}$. In this paper, we mainly focus on the W_1 distance, which can be equivalently formulated, through the Kantorovich duality (see, e.g., [Vil08]), as

$$W_1(\mu,\nu) = \sup\{\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi] : \varphi \text{ is 1-Lipschitz}\}. \tag{12}$$

The optimal transportation allows many equivalent characterization [Vil08] but is often difficult to solve analytically. A special case is on the real line, where the optimal transference plan is simple to find out and is common for various cost functions. Specifically, W_1 distance coincides with the L_1 distance between cumulative distribution functions:

$$W_1(\mu,\nu) = \int |F_\mu(t) - F_\nu(t)| \mathrm{d}t. \tag{13}$$

Wasserstein distance is a convenient metric for weak convergence, and it also implies other evaluation measures under additional assumptions. The accuracy of atoms is given in Lemma 2 under the Hausdorff distance, which measures the closeness between two sets and is defined by:

$$d_H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}|a-b|, \sup_{b\in B}\inf_{a\in A}|a-b|\right\}. \tag{14}$$

The accuracy of weights is described by Lemma 3. As a concrete application of those results, we prove Lemma 1, which recovers each components up to a permutation.

Lemma 2. *Let X and Y be two discrete random variables whose atom takes values in S_X and S_Y, respectively, and the minimum weight of each atom has at least ϵ probability. Then*

$$W_1(X,Y) \geq \epsilon \cdot d_H(S_X, S_Y).$$

Proof. For any coupling between X be Y,

$$\mathbb{E}d(X,Y) = \sum_x \mathbb{P}[X=x]\,\mathbb{E}[d(X,Y)|X=x] \geq \sum_x \epsilon \cdot \inf_{y\in S_Y} d(x,y) \geq \epsilon \cdot \sup_{x\in S_X}\inf_{y\in S_Y} d(x,y).$$

Interchanging X and Y completes the proof. □

Lemma 3. *For any Borel set B and any $\delta \geq 0$,*

$$\mu(B) \leq \nu(B^\delta) + W_1(\mu,\nu)/\delta \text{ and } \nu(B) \leq \mu(B^\delta) + W_1(\mu,\nu)/\delta,$$

where $B^\delta \triangleq \{x : \inf_{y\in B} d(x,y) \leq \delta\}$ denotes the δ-fattening of B.

Proof. Let X and Y be distributed according the optimal coupling of μ and ν, respectively. Then, applying Markov inequality yields that

$$\mathbb{P}[d(X,Y) > \delta] \leq \mathbb{E}[d(X,Y)]/\delta = W_1(\mu,\nu)/\delta.$$

By Strassen's theorem [Str65] (see, also, [Hub05]), for any Borel set B, we have $\mu(B) \leq \nu(B^\delta) + W_1(\mu,\nu)/\delta$ and $\nu(B) \leq \mu(B^\delta) + W_1(\mu,\nu)/\delta$. □

Proof of Lemma 1. Any atom x_i of U must be close to some atom of $\hat{U}$ within ϵ/ϵ_2, and otherwise $W_1(U,\hat{U}) \geq w_i(\epsilon/\epsilon_2) \geq \epsilon$. Note that $\epsilon/\epsilon_2 < 0.1\epsilon_1$ and atoms of U are separated by at least ϵ_1. Since $\hat{U}$ is k-atomic, then it must have exactly k atoms, and there is a permutation π such that $\|x - \pi\hat{x}\|_\infty < \epsilon/\epsilon_2$. Applying Lemma 3 with $B = \{x_i\}$ and $\delta = \epsilon_1/4$ yields that $w_i \leq (\pi\hat{w})_i + 4\epsilon/\epsilon_1$, and with $B = \{(\pi\hat{x})_i\}$ yields that $(\pi\hat{w})_i \leq w_i + 4\epsilon/\epsilon_1$. Therefore, $\|w - \pi\hat{w}\|_\infty \leq O(\epsilon/\epsilon_1)$ under the same permutation π. □

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.

1. If X and Y both have at most k atoms, then X and Y have the same distribution if and only if $m_i(X) = m_i(Y)$ for $i = 1, ..., 2k-1$.

2. If X has at most k atoms and Y is arbitrary, then X and Y have the same distribution if and only if $m_i(X) = m_i(Y)$ for $i = 1, ..., 2k$.

In the context of statistical inference, we only have access to samples and the noisy estimates of moments. To solve the inverse problems from moments to distributions, our theory relies on the following moments comparison results, a stable version of identifiability, which show that closeness of moments implies closeness of distributions in Wasserstein distance:

Proposition 1. *Let X and Y be k-atomic random variables taking values in $[-1, 1]$. Then*

$$\max_{i\in[2k-1]} |m_i(X) - m_i(Y)| \geq (W_1(X, Y)/O(k))^{2k-1}.$$

Proposition 2. *Let X be a k-atomic random variable taking values in $[-1, 1]$. Let Y be an arbitrary random variable. Then*

$$\max_{i\in[2k]} |m_i(X) - m_i(Y)| \geq (W_1(X, Y)/O(k))^{2k}.$$

Remark 1. The exponents in Proposition 1 and 2 cannot be improved. To see this, we first note that the number of moments in the identifiability results cannot be improved:

1. Given any $2k$ distinct points, there are two distributions that are supported on disjoint k-point subsets with identical first $2k-2$ moments (see the proof of Lemma 20).

2. Given any continuous distribution, its k-point quadrature is supported on k points and have identical first $2k-1$ moments.

By the first result, there exists two k-atomic random variables U and V whose $2k-1$ moments differ by $\Theta_k(1)$, while the W_1 distance between them is also $\Theta_k(1)$. Then we have $W_1(\epsilon U, \epsilon V) = \Theta_k(\epsilon)$ and $\max_{i\in[2k-1]} |m_i(X) - m_i(Y)| = (\Theta_k(\epsilon))^{2k-1}$, and thus the exponent in Proposition 1 is not improvable. Similarly, the conclusion for the exponent in Proposition 2 follows from the second result.

Remark 2. Classical moments comparison theorems aim to show convergence of distributions by comparing a growing number of moments. For example, Chebyshev's theorem states that if a probability distribution has the same first r moments as those of the standard normal, then its CDF deviates from the standard normal CDF by at most $\sqrt{\frac{\pi}{2r}}$ uniformly, i.e., in Kolmogorov-Smirnov distance (see [Dia87, Theorem 2]); for compactly supported distributions, this estimate can be improved to $O(\frac{\log r}{r})$ [Kra32]. More general comparison theorems can be obtained from the Chebyshev-Markov-Stieltjes inequalities (see [Akh65]). In contrast, in the context of estimating finite mixtures we are dealing with discrete distribution(s), in which case a *fixed* number of moments suffice to identify the distribution. However, with finitely many samples, it is impossible to exactly determine the moments, and measuring the error in Kolmogorov-Smirnov distance is too much to

ask.[1] As shown in Proposition 1 and Proposition 2, 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 two CDFs as opposed the uniform error (L_∞ distance).

4 Estimators and statistical guarantees

4.1 Known variance

This subsection describes the denoised method of moments to estimate Gaussian location mixture models whose component variance σ^2 is known. First, individual moments of the latent discrete distribution are estimated by generalized Hermite polynomials. Then, the vector of estimated moments are jointly denoised by semidefinite programming. Finally, the estimated discrete latent distribution is given by the quadrature rule.

The unbiased estimators for moments of the latent distribution in a Gaussian location mixture model are well-known: they are generalized Hermite polynomials $\gamma_r(x, \sigma)$ orthogonal with respect to the density function of $N(0, \sigma^2)$, given by

$$\gamma_r(x, \sigma) = r! \sum_{j=0}^{\lfloor r/2 \rfloor} \frac{(-1/2)^j}{j!(r-2j)!} \sigma^{2j} x^{r-2j}. \tag{15}$$

The polynomial $\gamma_r(\cdot, \sigma)$ is of degree r, and

$$\mathbb{E}[\gamma_r(U + \sigma Z, \sigma)] = \mathbb{E}[U^r].$$

To recover a k-atomic discrete latent distribution, by Proposition 1, moment estimates of degree up to $2k-1$ suffice. The estimate for each moment is $\sqrt{n}$-consistent:

Lemma 4. *Given n observations from (2), there is an estimate $\tilde{m}_r$ for $m_r(U)$ with*

$$|\tilde{m}_r - m_r(U)| < \sqrt{\frac{(O(r))^r \log(1/\delta)}{n}}$$

with probability at least $1 - \delta$.

As observed in the failure of vanilla method of moment in Section 1.2, one major difficulty lies in the possibility that the estimated moments do not constitute a legitimate moment sequence, despite each one is consistent. This observation inspires us to project the vector of moments back to the moment space. The projection is efficiently solvable since the space of the first $2k-1$ moments of all k-atomic distributions is convex and is completely characterized by positive semidefinite conditions (9). Therefore, through semidefinite programming (SDP), the estimated moments are jointly denoised by[2]

$$\min\{\|\tilde{m} - \hat{m}\|_2 : \hat{m} \text{ satisfies (9)}\}, \tag{16}$$

where $\tilde{m}$ denotes the vector of estimated moments.

Now that $\hat{m}$ is indeed a valid moment sequence, we proceed to the inverse moment procedure of finding a discrete latent distribution with those moments. This inverse procedure is known as the

[1] For example, consider two point masses δ_0 and δ_ϵ with arbitrarily small ϵ.

[2] The optimization problem formulated in (16) can already be implemented in popular modeling languages for convex optimization problem such as CVXPY [DB16]. A formulation in the standard form of SDP is given in Appendix A.

quadrature rule in numerical analysis [Akh65, CF91]. One basic algorithm to find the quadrature rule is given in Algorithm 5, and more efficient algorithms are introduced in Section 2.1.

We discuss the connection of our denoised method of moments to the generalized method of moments. The formulation of generalized method of moment minimizes the difference between estimated moments and moments of the estimated model [Han82, Hal05]:

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

where θ is the model parameters and W is a positive semidefinite weighting matrix. This formulation has pros and cons: it introduces a weighting matrix W that provides extra freedom to improve the accuracy, but the objective function $Q(\theta)$ in general can be non-convex in θ, notably under the Gaussian mixture model, and thus cannot be efficiently optimized. Our denoised method of moments resolves its difficulty, and can also be adapted to its advantage. We first discuss how this non-convex problem is efficiently solved. Through a change of variables, the model parameters θ are replaced by the first $2k-1$ moments of the latent distribution, and a key observation from Section 2.1 is that this moment space is convex and allows positive semidefinite characterizations, so the optimal auxiliary variables can be efficiently obtained through semidefinite programming. The inverse problem from auxiliary variables to original model parameters is still non-convex, and is typically solved through a high degree polynomial equation. Surprisingly, recovering a k-atomic distribution from its moments is the well-studied quadrature problem, and efficient solvers are readily available. Therefore, our denoised method of moments efficiently solved the non-convex optimization problem when applying generalized method of moments to Gaussian mixture models. Now we discuss how can our denoised method of moments be adapted to the optimal weighting matrix. The denoised method of moments described above is under the identity weighting matrix, and can be modified to incorporate a weighting matrix by a different objective function, which can also be efficiently solved. The optimal choice of weighting matrix is inverse of the limits of the covariance matrix of $\sqrt{n}(\hat{m} - m(\theta_0))$ as n diverges, where θ_0 is the unknown parameters to be estimated. It involves unknown model parameters and thus also need to be estimated from data. A popular approach is a two-step estimator [Hal05]: on the first step a suboptimal weighting matrix is used to obtain a consistent preliminary estimate of parameters and then a consistent estimate $\hat{W}$ of the optimal weighting matrix; on the second step the parameters are re-estimated using weighting matrix $\hat{W}$. This approach can be similarly applied in our denoised method of moments.

We close this subsection by a complete proof of the estimate accuracy (3), the Wasserstein distance guarantee in the main Theorem 1.

Proof of (3) of Theorem 1. We first note that it suffices to consider the case $M = 1$. For a Gaussian mixture model $X = U + \sigma Z$ with U bounded by M, we first divide all samples by M and consider the model $\frac{U}{M} + \frac{\sigma}{M} Z$, and then the centers are bounded by one. If we can recover the distribution of $\frac{U}{M}$ within Wasserstein distance ϵ, so is the distribution of U within ϵM. In the following, we assume $M = 1$. Denote the estimated moments of the mixture distribution using n samples by $\tilde{m}' = (\tilde{m}'_1, \ldots, \tilde{m}'_{2k-1})$. It follows from Lemma 27 and the union bound that $|\tilde{m}'_r - m_r(X)| < \epsilon \triangleq \sqrt{\frac{(O(k))^{2k} \log(k/\delta)}{n}}$ for all $r = 1, \ldots, 2k-1$ with probability at least $1 - \delta$. Then the unbiased estimate of moments of U, denoted by $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_{2k-1})$, using Hermite polynomials (15), has accuracy $|\tilde{m}_j - m_j(U)| < (O(\sqrt{k}))^{2k-1}\epsilon$ by Lemma 18, and the error is at most $(O(k))^k \epsilon$ in ℓ_2-norm. Denote the denoised moments after projection (16) onto the moments space of distributions on $[-1, 1]$ by $\hat{m}$. Since the moments of the underlying latent distribution $m = (m_1(U), \ldots, m_{2k-1}(U))$ satisfies (9), the projection (16) incurs an error at most $(O(k))^k \epsilon$ in ℓ_2-norm, and thus $\|\hat{m} - m\|_2 \le (O(k))^k \epsilon$. Note that $\hat{m}$ is the moments of the estimated latent variable $\hat{U}$, and then applying Proposition 1 yields that $W_1(U, \hat{U}) \le O(k^2 (n/\log \frac{1}{\delta})^{-\frac{1}{4k-2}})$ with probability at least $1 - \delta$. □

4.2 Unknown variance

In this subsection, it is assumed that the Gaussian variance parameter is unknown. The optimal rate in Section 4.1 is obtained by an unbiased estimate on the moments of the latent distribution. Unfortunately, when the variance is unknown, such unbiased estimator no longer exists, since the unique unbiased estimator under each fixed variance is the Hermite polynomial (15), which is not universal across different variances. In fact, it is impossible to obtain an unbiased estimator using any finite number of samples, as proved in Lemma 25. In hindsight, the non-existence of unbiased estimators can be seen from the lower bound on the estimation error of the variance in Proposition 7. Consider the estimate on the second moment. Under the model $X = U + \sigma Z$, we have $m_2(X) = m_2(U) + \sigma^2$, where $m_2(X)$ is $\sqrt{n}$-consistent. Since estimate accuracy on σ^2 is worse than $\Omega(n^{-\frac{1}{2k}})$ in the worst case by Proposition 7, any estimate on $m_2(U)$ is impossible to be more accurate.

The problem formulation in this subsection can still be understood from the perspective of deconvolution, but now the Gaussian noise is of unknown variance which also needs to be estimated from data. It is not difficult to consistently estimate the unknown variance, for example, by $\hat{\sigma} = \max\{X_1, \dots, X_n\}/\sqrt{2\log n}$, which satisfies $|\sigma - \hat{\sigma}| = O_P(1/\sqrt{\log n})$, and substitute into the denoised method of moments developed in Section 4.1; however, the convergence rate is too slow to be useful for achieving the optimal rate in Theorem 1. In fact, it is impossible to have an estimate for the variance more accurate than $O(n^{-\frac{1}{2k}})$ in the worst case, obtained in Proposition 7. Then, after deconvolution, the estimate on the moments of the latent distribution is also only of accuracy at most $O(n^{-\frac{1}{2k}})$. Analogous analysis to Section 4.1 using Proposition 1 yields an accuracy no better than $O_k(n^{-\frac{1}{2k(2k-1)}})$. It is unclear how to improve the accuracy guarantee without a dedicated estimate on the variance. A joint estimate on both variance and means is crucial to achieve the optimal rate.

The following Proposition 3 provides an accuracy guarantee of any joint estimate of both latent distribution and variance in terms of its moments accuracy. One procedure to find a joint estimate is by Lindsay's estimator [Lin89]. In this subsection we describe Lindsay's estimator and prove that the usual method of moments is exactly solved with probability one. Therefore, all formulations of the usual method of moments, the generalized method of moments, and our denoised method of moments, as discussed in Section 4.1, are all equivalent.

Proposition 3. *Let $X = U + \sigma Z$ and $X' = U' + \sigma' Z$, where U and U' are bounded k-atomic random variables, σ, σ' are bounded and positive, and Z is independent standard normal. Suppose $|m_i(X) - m_i(X')| \le \delta$ for $i = 1, \dots, 2k$. Then, $|\sigma^2 - \sigma'^2| \le O(k^2 \delta^{\frac{1}{k}})$, and $W_1(U, U') \le O(k^{3/2} \delta^{\frac{1}{2k}})$.*

In the following, we describe the main procedure of Lindsay's estimator [Lin89], prove a few key properties, and conclude that the usual method of moments allows a solution with probability one as long as mixture model has a density. Then its accuracy immediately follows from the accuracy of moments estimate and Proposition 3. The first step is to estimate the Gaussian scale σ. On the population level, the sequence $\mathbb{E}[\gamma_i(X, \sigma)]$ for $i = 1, \dots, 2k$, with γ_i defined in (15), is the first $2k$ moments of U, which is a k-atomic random variable, whose k^{th} moment matrix is degenerate according to Theorem 4. For any $\sigma' < \sigma$, the sequence $\mathbb{E}[\gamma_i(X, \sigma')]$ equals the first $2k$ moments of $U + \tau Z$, whose k^{th} moment matrix is non-degenerate, also by Theorem 4, where $\tau = \sqrt{\sigma^2 - \sigma'^2}$. Let $\sigma \mapsto d_k(\sigma)$ denote the determinant of the moment matrix associated with the moment sequence $\mathbb{E}[\gamma_i(X, \sigma)]$ for $i \le 2k$, which is a polynomial of degree $\frac{k(k+1)}{2}$ in σ^2. Then σ can be identified by the smallest non-negative root of the equation $d_k(\sigma) = 0$. When only samples from the mixture model are available, each $\mathbb{E}[\gamma_j(X, \sigma)]$ in the k^{th} moment matrix, as a function of σ, can be estimated by the empirical moments of $\gamma_j(X, \sigma)$, denoted by $\hat{\gamma}_j(\sigma)$, which is a linear combination of the empirical

moments of X of degree up to j. Denote the determinant of this estimated k^{th} moment matrix with by $\hat{d}_k(\sigma)$. The estimate of the standard deviation, is defined as

$$\hat{\sigma} = \min\{\sigma \geq 0 : \hat{d}_k(\sigma) = 0\}. \tag{17}$$

The following result guarantees that $\hat{d}_k(\sigma) = 0$ necessarily has a non-negative root.

Lemma 5. *Assume that the number of samples $n > k$ and the distribution of X has a density. Then, almost surely, the estimate $\hat{\sigma}$ exists, and $0 < \hat{\sigma} \leq s$, where $s^2 = \overline{X_i^2} - (\overline{X_i})^2$ is the sample variance.*

The next step is to estimate the latent distribution. On population level, the j^{th} moment of the latent distribution equals $\mathbb{E}[\gamma_j(X, \sigma)]$. It is estimated by the empirical moment $\hat{\gamma}_j(\hat{\sigma})$, using the estimate of standard deviation $\hat{\sigma}$ defined above, for each $j = 1, \ldots, 2k$. One might be concerned about whether the vector of individual estimated moment $\hat{\gamma}_j(\hat{\sigma})$ is in the moment space of k-atomic distributions, but it turns out this vector indeed is with probability one, as shown in the following Lemma 6. Then, analogous to the known variance case in Section 4.1, the estimate on the latent distribution is defined by the unique k-atomic distribution whose j^{th} moment coincides with the corresponding estimated moment $\hat{\gamma}_j(\hat{\sigma})$, for all $j = 1, \ldots, 2k$. The distribution is obtained by a procedure that calculates a quadrature representation from moments, as introduced in Section 2.1 and summarized in Algorithm 5. Equivalently, the moments of the estimated Gaussian mixture agree with the empirical moments exactly:

$$m_j(\hat{U} + \hat{\sigma}Z) = \bar{m}_j = m_j(\bar{\pi}), \quad j = 1, \ldots, 2k, \tag{18}$$

where $\bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ denotes the empirical distribution. The complete procedure is summarized in Algorithm 1.

Lemma 6. *Assume that the number of samples $n \geq 2k-1$ and the distribution of X has a density. Then, almost surely, the empirical moments $\hat{\gamma}_j(\hat{\sigma})$ for $j \leq 2k$ coincides with the corresponding moments of a discrete distribution with exactly k points of support.*

Algorithm 1 Lindsay's estimator for normal mixtures with an unknown common variance

Input: n samples $X_1, \ldots, X_n$ drawn i.i.d. from the mixing distribution, $n \geq 2k - 1$.
Output: estimated mixture model $\hat{U} + \hat{\sigma}Z$.

```
for r = 1 to 2k do
    m̂_r = (1/n) Σ_i X_i^r
    γ̂_r(σ) = Σ_{i=0}^{⌊r/2⌋} r! / ((−2)^i i! (r−2i)!) m̂_{r−2i} σ^{2i}
end for
Let d̂_k(σ) be the determinant of the matrix {γ̂_{i+j}(σ)}_{i,j=0}^{k}.
Let σ̂ be the smallest non-negative root of d̂_k(σ) = 0.
for r = 1 to 2k do
    γ_r = γ̂_r(σ̂)
end for
Let (x_1, ..., x_k) be nodes and (w_1, ..., w_k) be weights of the quadrature rule of (γ_1, ..., γ_{2k−1}).
Let the distribution of Û be P[Û = x_i] = w_i.
```

In summary, Algorithm 1 produces an estimate of the k-component mixture model $\hat{X} = \hat{U} + \hat{\sigma}Z$ whose r^{th} moment equals the empirical moment $\hat{m}_r$ for $r \leq 2k$, where $\hat{U}$ is a k-atomic random

variable. Those empirical moment estimates are accurate within an additive error of $O_k(\frac{1}{\sqrt{n}})$ with high probability, when U and σ are both bounded. By Lemma 5, the estimate $\hat{\sigma}$ is also bounded. The estimate $\hat{U}$, as the outcome of the quadrature rule, is not necessarily to be bounded. Nevertheless, the tail probability of $\hat{U}$ in Corollary 1 below allows us to apply a threshold on $\hat{U}$ while only slightly change the moments, and thus $\hat{U}$ can be effectively treated as bounded. Then we are ready to apply Proposition 3 to obtain a provable accuracy.

Lemma 7. *Given two random variables X and Y, and one of them takes at most k values. If $|X| \le 1$, then, for any $t \ge 2$,*

$$\mathbb{P}[|Y| \ge t] \le \frac{\max_{i\in[2k]} |m_i(X) - m_i(Y)|}{(\Omega(t))^{2k}}.$$

Corollary 1. *Let $X = U + \sigma Z$ and $\hat{X} = \hat{U} + \hat{\sigma} Z$, where $U, \hat{U}$ each takes at most k values, and $U, \sigma, \hat{\sigma}$ are all bounded. Then, for any $t \ge O(\sqrt{k})$,*

$$\mathbb{P}[|\hat{U}| \ge t] \le \frac{\max_{i\in[2k]} |m_i(X) - m_i(\hat{X})|}{(t/O(\sqrt{k}))^{2k}}.$$

This subsection closes with Theorem 5 on the estimation accuracy of Algorithm 1, which implies (4) in the main Theorem 1 for $M = 1$. Analogous to the proof of (3) of Theorem 1 (in the end of Section 4.1), the result for general M follows from a simple scaling argument.

Theorem 5. *Let $(\hat{U}, \hat{\sigma})$ be obtained by Algorithm 1 using $n' = n/\log\frac{1}{\delta}$ i.i.d. samples from the mixture model $X = U + \sigma Z$, where $|U|, \sigma \le 1$. Then, with probability at least $1 - \delta$, the estimate satisfies $|\sigma^2 - \hat{\sigma}^2| \le O_k(n'^{-\frac{1}{2k}})$ and $W_1(U, \hat{U}) \le O_k(n'^{-\frac{1}{4k}})$ for any $k \le O(\frac{\log n'}{\log\log n'})$, where the hidden constants in O_k are at most polynomial in k.*

4.3 Adaptive to the cluster structure

In this subsection, we demonstrate the adaptive rates of our denoised method of moments. To show an improvement on the accuracy of our denoised method of moments, a crucial step is the rates of our moments comparison theorems in Section 3 under separation assumptions. There are various formulations on separations, for example, opposite to the case of no separation, another extreme case is when all atoms of each discrete distribution are pairwise well-separated. We assume a milder separation condition that each atom is possibly close to a few atoms but is far away from others. The following Proposition 4 and 5 are generalizations of Proposition 1 and 2, respectively, in which each atom is possibly close to all other atoms.

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

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

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

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

Now we proceed to the adaptive rates of our denoised method of moments under separation assumptions. Suppose k component centers in the unknown Gaussian mixture model $U + \sigma Z$ can be grouped into k_0 different clusters, and we assume separations between intercluster atoms but impose no assumption on intracluster atoms. In this case each atom of either U or $\hat{U}$ is close to at most $2k - k_0$ atoms, and when σ is known, the adaptive rate follows from Proposition 4. The result is given in Theorem 6, a generalization of (3), where one cluster contains all atoms. Note that the upper bound in (4) in the case of unknown σ is not improvable since the worst scenario present in Section 5 is still legitimate and the same lower bound continues to hold.

Theorem 6. *Under the conditions of Theorem 1, suppose k component centers form k_0 different clusters and intercluster atoms are separated by at least γ. If σ is given, then with probability at least $1 - \delta$,*

$$W_1(\nu, \hat{\nu}) \leq M(O(k))^{\frac{4k-k_0+1}{2k-k_0}} \gamma^{-\frac{k_0-1}{2k-k_0}} \left(\frac{n}{\log(k/\delta)} \right)^{-\frac{1}{4k-2k_0}}. \tag{19}$$

Proof. Note that U and $\hat{U}$ are both supported on a set of $2k$ atoms, and the largest cluster of U is of size at most $k - k_0 + 1$. Since different clusters of U are separated by γ, then each atom of either U and $\hat{U}$ is at least $\gamma/2$ away from all but $2k - k_0$ atoms. From the proof of (3), we have $|m_r(\hat{U}) - m_r(U)| < (O(k))^{2k} \sqrt{\frac{\log(k/\delta)}{n}}$. The conclusion follows from Proposition 4. □

The analytical accuracy of our denoised method of moments can be further improved if, in addition to the cluster separation conditions, the total mixing weights of each of the k_0 clusters of U are not superficial. This situation is considered in [HK15], and therein the optimal rate of accuracy is shown to be $n^{-\frac{1}{4(k-k_0)+2}}$ when σ is given, achieved by minimum distance estimation, the best fit of empirical distribution under Kolmogorov-Smirnov distance, which is in general computationally infeasible. Our denoised method of moments is adaptive to this condition and achieves the optimal rate. In this case, each cluster of U has at least one atom of $\hat{U}$ nearby due to the total mixing weights, so atoms of $\hat{U}$ also form k_0 clusters. When σ is given, each atom of either U or $\hat{U}$ is close to at most $2(k - k_0) + 1$ other atoms, and the adaptive rate follows again from Proposition 4; when σ is unknown, the rate follows from Proposition 4 and an extension of Proposition 3. The results are given in Theorem 2 and are proved here.

Proof of Theorem 2. By the lower bound on the total mixing weights of each cluster of U, there is at least one atom of $\hat{U}$ within a distance of 0.1γ, due to the assumption that $W_1(U, \hat{U}) \leq \epsilon$. Since different clusters of U are separated by γ, so atoms of $\hat{U}$ also form k_0 clusters with intercluster atoms separated by $\Omega(\gamma)$. When σ is given, each atom of either U and $\hat{U}$ is at least $\Omega(\gamma)$ away from all but $2(k - k_0) + 1$ atoms, and the conclusion in (6) follows analogous to the proof of Theorem 6. When σ is unknown, the conclusion in (7) follows from a similar to the proof of Theorem 5 with Proposition 3 replaced by Proposition 9. □

4.4 Unbounded means

In both Section 4.1 and 4.2, we assume that centers in the Gaussian mixture model are bounded. The boundedness assumption is necessary to reach a Wasserstein distance guarantee on the latent distribution, and the necessity can be seen from the following situation: Consider a noise-free model $\sigma = 0$. It is impossible to reliably distinguish the two-atomic distribution $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_M$ from a slightly perturbed distribution $\frac{1+\epsilon}{2}\delta_0 + \frac{1-\epsilon}{2}\delta_M$ with $O(1/\epsilon^2)$ samples. When only n samples are available, let $\epsilon = 1/\sqrt{n}$ and the worst-case loss in W_1 distance for any estimator is $\Omega(M/\sqrt{n})$, which grows linearly in M. Therefore, the worst-case risk in the Wasserstein distance is infinity without any

bound on the centers. Though impossible to recover the latent distribution in Wasserstein distance, due to the accuracy limitation on mixing weights and the wide range of centers, it is still possible to recovers the centers of components in the Hausdorff distance given lower bounds on mixing weights, as defined in (14), without any boundedness assumption. This subsection provides an algorithm that recovers centers with a slightly worse performance as compared to the bounded centers case.

In the unbounded centers case, blindly applying the moment-based methods proposed in Section 4.1 and 4.2 incurs a large error, since the moments estimate can have high variance, again due to the wide range of centers. To resolve this problem, when there are only a small number of components, the idea is to first roughly cluster the samples and the mixing components into a small number of groups, and narrow the range of centers within each group. Applying previous methods to each group of samples yields accuracy guarantees for the corresponding group of centers. Provided that centers in each group are accurate in the Hausdorff distance, the union of all centers inherits the same accuracy.

Next we describe our algorithm, a two-step procedure, in details. In the first step, we cluster the samples into disjoint groups. The range of each group is obtained via a small number of independent test samples, denoted by $\tilde{X}_1, \ldots, \tilde{X}_{n'}$, where n' is selected only to ensures that each component generated at least one sample in the test samples. The ranges are delimited by a union of disjoint intervals, denoted by $I_1, \ldots, I_s$, given by

$$I_1 \cup \cdots \cup I_s = \cup_i [\tilde{X}_i \pm \Theta_k(\sqrt{\log n})],$$

where s is the number of groups. The intuition for this clustering is that n realizations of Gaussian noise are at most $O(\sqrt{\log n})$ in magnitude with high probability, and, conditioned on that happens, this procedure precisely clusters all samples by their latent centers: a sample is in one interval if and only if its latent center is in the same interval. Moreover, each cluster is on a provably moderate range. The procedure is given in Algorithm 2 and its properties are shown in Lemma 8. This is a conservative yet simple clustering with provable guarantees, and in practice various clustering algorithms for specific applications are available, such as the popular Lloyd's algorithm [Llo82] for k-means clustering.

Algorithm 2 Cluster samples from a Gaussian mixture model.

Input: n samples $X_1, \ldots, X_n$ to cluster, and n' test samples $\tilde{X}_1, \ldots, \tilde{X}_{n'}$, both drawn i.i.d. from a Gaussian mixture model, and a length parameter $L = \Omega(\sqrt{\log n})$.

Output: disjoint intervals $I_1, I_2, \ldots$, and disjoint groups of samples $\mathcal{S}_j \subseteq \{X_1, \ldots, X_n\}$ on I_j.

Merge overlapping intervals $[\tilde{X}_i \pm L]$ into disjoint ones $I_1, \ldots, I_s$, given by,

$$I_1 \cup \cdots \cup I_s = [\tilde{X}_i \pm L].$$

Report $\mathcal{S}_j = \{X_i \in I_j : i \in [n]\}$ for each j.

Lemma 8. *Given n samples to cluster from a k-component Gaussian location mixture model with mixing weights at least ϵ and bounded Gaussian scale, using n' extra independent test samples, with probability at least $1 - \delta$, Algorithm 2 with $L = \Omega(\sqrt{\log n}) + \sqrt{2\log(1/\epsilon')}$ groups n samples correctly: a sample X_i is in cluster j if and only if the latent center is in the same cluster, where $\delta = n^{-\Omega(1)} + n'^{-\Omega(1)} + ke^{-n'(\epsilon-\epsilon')}$. Moreover, the range of each interval only depends on the test samples and is of length $O(k(L + \sqrt{\log n'}))$.*

In the second step, we focus on each group of samples separately, partitioned by those s disjoint intervals $\{X_i \in I_j : i \in [n]\}$. We shall first shift all samples in each interval to the origin $\{X_i - c_j : X_i \in I_j\}$ where c_j is the center of interval I_j, and apply our denoised method of moments to each set of shifted samples, which produces a set of centers $\{\hat{\mu}_i'\}$, and then shift the centers back to the original interval by $\{\hat{\mu}_i = \hat{\mu}_i' + c_j\}$. The final estimate on the centers is the union of centers over all intervals. The algorithm is summarized in Algorithm 3 and its accuracy guarantee is given in Theorem 7. Note that this is only a worst-case guarantee, and the practical accuracy is possibly much better: the number of samples in each cluster increases proportionally to the total mixing weights; the adaptive rate in Theorem 2 is also applicable when components within one cluster moderately overlap; we can postulate fewer components in one cluster based on information from other clusters.

Algorithm 3 Recover centers of a Gaussian mixture model in the unbounded case.

Input: $N = n + n'$ samples $X_1, X_2, \ldots, X_N$ drawn i.i.d. from a Gaussian mixture model, clustering parameter $L = \Omega(\sqrt{\log n})$, and mixing weights threshold τ.

Output: estimated centers $\hat{\mu}$.

Let s groups of samples $\mathcal{S}_1, \ldots, \mathcal{S}_s$ on intervals $I_1, \ldots, I_s$ be obtain from Algorithm 2 with n' test samples and n samples to cluster using length parameter L.

for $j = 1$ **to** s **do**

Let c_j be the center of I_j.

$\mathcal{S}_j' = \{x - c_j : x \in \mathcal{S}_j\}$.

Let $(\hat{w}, \hat{x})$ be mixing weights and centers from the denoised method of moments on $\mathcal{S}_j'$.

$\hat{\mathcal{C}}_j = \{\hat{x} + c_j : \hat{w} \geq \tau\}$.

end for

$\hat{\mu} = \cup_j \hat{\mathcal{C}}_j$.

Theorem 7. *Suppose in a k-component Gaussian location mixture model, the mixing weights are at least ϵ, the Gaussian scale is bounded, and the number of components $k \leq O(\frac{\log n}{\log \log n})$. Denote the set of centers by μ. Applying Algorithm 3 with $N = n + n'$ samples, where $n \geq n' \geq \Omega(\frac{\log k}{\epsilon})$, clustering parameter $L = \Theta(\sqrt{\log n + k \log \log n})$, and mixing weights threshold $\tau = \epsilon/(2k)$, with probability, say 0.9, the estimate on the set of centers $\hat{\mu}$ satisfies*

$$d_H(\hat{\mu}, \mu) \leq \begin{cases} O_k(L(\epsilon n)^{-\frac{1}{4k-2}}/\epsilon), & \sigma \text{ is known}, \\ O_k(L(\epsilon n)^{-\frac{1}{4k}}/\epsilon), & \sigma \text{ is unknown}, \end{cases}$$

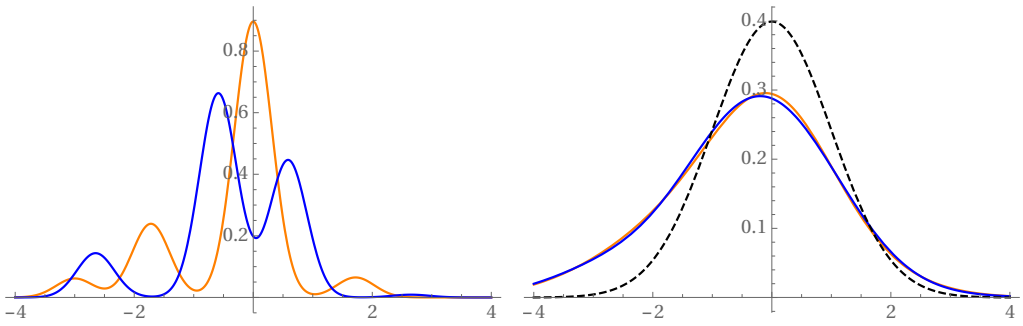
where the hidden constants in O_k are at most polynomial in k.

5 Lower bounds

This section introduces lower bounds on the estimation error in learning Gaussian location mixture models. The lower bounds match the accuracy guarantee in Theorem 1 within a polynomial factor in k, and thus our denoised method of moments is rate-optimal when k is a constant. The lower bounds are established through two instances of Gaussian location mixture models: they are statistically very close that it is impossibly to reliably distinguish the two models using a given number of samples, but they have different model parameters, including a difference between latent distributions in the Wasserstein distance, and also between the Gaussian variances in the case that

the scale is unknown. Then any estimate suffers a loss at least proportional to the parameters difference even in the presence of only these two models.

To show a vanishing statistical distance between two mixture models, one commonly used approach is *moment matching*: the mixture distributions are statistically close when the latent distributions match many moments, as demonstrated in Fig. 2. The connection is established



(a) Latent distributions (b) Mixture distributions

Figure 2: Densities of mixture distributions that match the first six moments. In (a), two different latent distributions coincide on their first six moments; In (b), after adding $Z \sim N(0,1)$ (the black dashed line) to the latent variables, the Gaussian mixtures become statistically much closer.

by various methods, for example, by orthogonal polynomials expansion [WV10, CL11], by Taylor series expansion [HP15, WY16], and by the relation to the best polynomial approximation [WY15]. Any Gaussian mixture model can be represented by $U + \sigma Z$, where U is discrete or a Gaussian mixture itself and Z is independently standard normal. Since common statistical distances are scale-invariant, including all f-divergences [AS66], we only consider the model $U + Z$ with U being ϵ-subgaussian:

$$\mathbb{E}[\exp(tU)] \le \exp(t^2\epsilon^2/2), \quad \forall\, t \in \mathbb{R},$$

where ϵ quantifies the signal-to-noise ratio. The χ^2-divergence between distributions of $U + Z$ and $U' + Z$ is $(O(\epsilon))^{2\ell+2}$ when U and U' match the first ℓ moments, as summarized in Lemma 9. The result is partially obtained by [WV10, CL11, HP15] in specific applications. A proof using Lemma 14 in Appendix C is given for completeness, and Remark 3 shows that the bound is not improvable.

Lemma 9. *If two ϵ-subgaussian random variables U, U' that match moments up to degree ℓ, where ϵ is less than some absolute constant, then*

$$\chi^2(U + Z \| U' + Z) \le (O(\epsilon))^{2\ell+2}.$$

Proof. Note that $\mathbb{E}[U'] = 0$, $\mathsf{var}[U'] \le \epsilon^2$, and $\mathbb{E}|U|^p, \mathbb{E}|U'|^p \le (O(\epsilon\sqrt{p}))^p$ by ϵ-subgaussianness. Applying Lemma 14 and the Stirling's approximation $n! > \sqrt{2\pi n}(n/e)^n$ yields that

$$\chi^2(U + Z \| U' + Z) \le e^{\epsilon^2/2} \sum_{j \ge \ell+1} \frac{(O(\epsilon))^{2j}}{\sqrt{2\pi j}} \le (O(\epsilon))^{2\ell+2}. \qquad \square$$

Remark 3 (Tightness of Lemma 9)**.** We show two latent distributions that match their first $2n-1$ moments while the χ^2-divergence is at least $(\Omega(\epsilon))^{4n}$, and therefore the result in Lemma 9 is not improvable. To this end, let the atoms and weights of the distribution of G_n be respectively the nodes and weights of the n-point quadrature rule as (10) under standard normal distribution

such that $\mathbb{E}[G_n^j] = \mathbb{E}[Z^j]$ for $j = 1, \ldots, 2n-1$, where Z is standard normal. Then G_n is 1-subgaussian (This can be seen from the moments dominance of G_n by standard normal. See a proof in Lemma 22) and thus ϵG_n is ϵ-subgaussian. Let Z' be independently standard normal, and then ϵG_n and $\epsilon Z'$ coincide on their first $2n-1$ moments. It is obtained in [WV10, (54)] that

$$\chi^2(\epsilon G_n + Z \| \epsilon Z' + Z) = \sum_{k \geq 2n} \frac{1}{k!} \left(\frac{\epsilon^2}{1+\epsilon^2} \right)^k |\mathbb{E}[H_k(G_n)]|^2,$$

where G_n, Z' are both independent of Z. Since $\mathbb{E}[H_{2n}(G_n)] = -n!$ (see Lemma 23), the χ^2-divergence is lower bounded by $\frac{(n!)^2}{(2n)!}(\frac{\epsilon^2}{1+\epsilon^2})^{2n} = (\Theta(\epsilon))^{4n}$ for $\epsilon = O(1)$.

To prove a lower bound, in view of Lemma 9, it boils down to finding two different latent distributions that match many moments. The optimal lower bounds follows from distributions that match as many moments as possible provided they are identifiable, and the largest degree is introduced in Section 3. The construction of two different distributions is also introduced in Remark 1 on the tightness of identifiability results. The lower bounds are correspondingly obtained in both cases where the variance is known or unknown in Proposition 6 and 7, and the exponents match the upper bounds in Section 4.1 and 4.2.

Known variance. Without loss of generality, we shall assume an unit variance. From Remark 1 there are two different k-atomic distributions, denoted by U and U', that match the first $2k-2$ moments. It follows from Lemma 9 that $(\Omega(\frac{1}{\epsilon}))^{4k-2}$ samples are necessary to reliably distinguish the mixture models $\epsilon U + Z$ and $\epsilon U' + Z$, and the accuracy in the Wasserstein distance is at most $\epsilon W_1(U, U')/2$. The best lower bound follows from maximizing the Wasserstein distance between two moment matching distributions, which is obtained by the following optimization problem:

$$\begin{aligned} \max \ & W_1(U, U') \\ \text{s.t. } & m_i(U) = m_i(U'), \quad i = 1, \ldots, 2k-2, \\ & |U|, |U'| \leq 1, \text{ both are } k\text{-atomic.} \end{aligned}$$

It turns out the maximal value of the above problem is $\Theta(1/k)$ (see a proof in Lemma 21). The lower bound is summarized in Proposition 6.

Proposition 6. *For any $\hat{U}$, there is some mixture model $U + \sigma Z$, where U is bounded and k-atomic, σ is bounded and is given, and Z is independently standard normal, such that using n samples from that mixture model $W_1(U, \hat{U}) \geq \Omega(\frac{1}{\sqrt{k}} n^{-\frac{1}{4k-2}})$ with constant probability.*

Remark 4. The proof can be easily extended to prove the optimality of (6). In the case of k_0 clusters of components, we construct two instances with identical $k_0 - 1$ single-component clusters, and the culprit is the difference within the last cluster with $k - k_0 + 1$ components that are difficult to distinguish. To this end, let V and V' be two $(k-k_0+1)$-atomic bounded random variables as in the proof of Proposition 6, and let the last cluster be either ϵV or $\epsilon V'$. Then the χ^2-divergence between two Gaussian mixture models is at most $(O(\epsilon))^{4(k-k_0)+2}$, which yields a minimax lower bound $\Omega_k(n^{-\frac{1}{4(k-k_0)+2}})$.

Unknown variance. The latent distribution is not necessarily to be k-atomic discrete distribution. The mixture model $U + Z$ is a k-component Gaussian location mixture as long as the latent distribution is. Let U be the k-point quadrature for the standard normal distribution, and let U'

be standard normal, and then they match their first $2k-1$ moments. The model of $\epsilon U + Z$ is a k-component Gaussian location mixture with an unit variance, and the model of $\epsilon U' + Z$ has one single component with variance $1+\epsilon^2$. Their difference on means are $\Theta_k(\epsilon)$ in the Wasserstein distance, and ϵ^2 on variance. As a result of Lemma 9, distinguishing those two mixture models requires at least $(\Omega(\frac{1}{\epsilon}))^{4k}$ samples. The lower bound in this case is summarized in Proposition 7.

Proposition 7. *For any $(\hat{U}, \hat{\sigma})$, there is some mixture model $U + \sigma Z$, where U is bounded and k-atomic, σ is bounded, and Z is independently standard normal, such that using n samples from that mixture model either $W_1(U, \hat{U}) \geq \Omega(\frac{1}{\sqrt{k}} n^{-\frac{1}{4k}})$ or $|\sigma^2 - \hat{\sigma}^2| \geq \Omega(n^{-\frac{1}{2k}})$ with constant probability.*

6 Numerical experiments

In this section, we evaluate our algorithm through empirical experiments, and compare the accuracy and running time with EM algorithm, both implemented in Python. Our moment projection step is implemented in semi-definite programming using CVXPY [DB16], and we implement the quadrature algorithm in [GW69], which recovers a discrete distribution from its moments. We also compare the accuracy of GMM obtained from the popular package gmm [Cha10] in R. The running time of GMM is ignored since the linear constrained optimization procedure used in gmm package is extremely slow for Gaussian mixture model: it takes days to finish in a mild scale as compared to minutes using our algorithm. We also ignore the comparison with MM since it failed in a constant fraction of experiments.

In our experiments with iterative numerical solvers, including EM and the optimization procedure in GMM, we need to set an initial guess and a stop criterion. For both algorithms, the initial guess is the best of five random guesses: We run the algorithm five times using different random guesses, with means drawn independently from a uniform distribution, and mixing weights drawn from a Dirichelet distribution. Among five estimates, we pick the one with the best objective function, which is the maximum likelihood for EM, and the minimum moments discrepancy for GMM. For the linearly constraint optimization solver in GMM, we use its default stop criterion; EM algorithm terminates when log-likelihood increases less than 10^{-3} or 5,000 iterations are reached.

Known variance. We run the experiments with five-component Gaussian mixture models, with unit variance in each component, means uniformly drawn from $[-1, 1]$, and mixing weights drawn from Dirichelet distribution with equal parameters. The sample size ranges from 500 to 5,000, and we plot the W_1 distance versus the sample size. We repeat 20 times under a fixed Gaussian mixture model, and plot the average error and the standard deviation. We also plot the running time versus the sample size. The random instance has the following parameters:

Weights:	0.123	0.552	0.010	0.080	0.235
Centers:	-0.236	-0.168	-0.987	0.299	0.150

The results are shown in Fig. 3. The accuracy is comparable among three methods. However, the running time of EM algorithm grows fast as sample size increases, and is about 15 times more than our DMM algorithm using 5,000 samples, partly because EM needs to access all samples in each iteration, instead of first summarizing data into a few moments. This is particularly restrictive in the presence of big data these days, and also restricts our experiments comparison on larger sample size. Though GMM applies the same approach as ours, the heuristic solver used in gmm package in R for this minimization problem is very slow and takes days to finish all experiments.

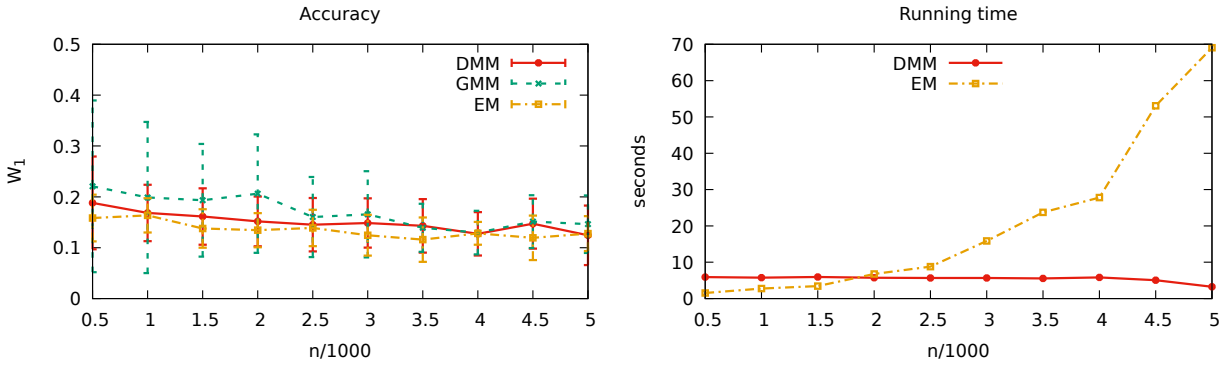


Figure 3: Comparison of different methods under a randomly generated five-component Gaussian mixture model.

The slow convergence of EM algorithm is more severe when there are more overlaps among components [RW84, KX03]. The accuracy also deteriorates when the log-likelihood is trapped in a flat area, since a lack a progress is no longer a good indication of convergence. Any fixed stop criterion is likely to terminate the EM algorithm early; a more stringent stop criterion improves the algorithm insubstantially but incurs much slower convergence. To see this, we run an experiments to learn a two-component Gaussian mixture model with unit variance in each component; however, the two components completely overlap, i.e., the sample generating model is $X = Z$. The setup of experiment is the same as before, and the results are shown in Fig. 4, where we add an additional line with label EM^+ and require the log-likelihood increases less than 10^{-4} instead of 10^{-3}. In

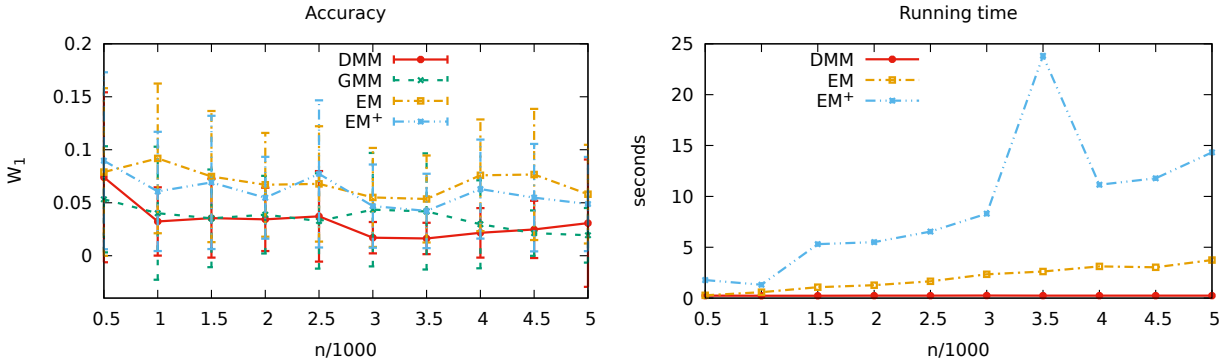


Figure 4: Comparison of different methods when components completely overlap.

this experiment, the performances of DMM and GMM are still similar, but we can clearly see the relative degradation of the EM algorithm. With a more stringent stop criterion in EM^+, the performance only slightly improves but the running time significantly worsens. More specifically, using 5,000 samples and comparing with the running time of DMM, EM is still about 15 times slower, but EM^+ is about 60 times slower. The gmm package finishes in an bearable time in this experiment as there are only two components, but it is still more than 200 times slower than DMM, and thus its running time is again ignored in the plot.

Unknown variance. We repeat the same setups as before, except that estimators no longer have access to the true variance in Gaussian components. The results are shown in Fig. 5. To estimate a five-component model, Lindsay's estimator involves the empirical moments up to degree 10, among which high order ones are inaccurate with limited number of samples. For instance, in the current experiment, the variance of 10^{th} moment is over one billion while we only sampled 5,000 samples. However, by the analysis is Section 4.2, the usual method of moments can always be exactly solved, so the bad performance is a consequence of the inaccuracy in high order moments. The gmm package achieves a better performance, and a possible reason is that high order moments only have small weights in the minimization problem, so the effect of inaccurate empirical moments is alleviated. When only two components need to be estimated and moments of degree up to four are involved, the performances become similar. The running time comparison is roughly the same as before and is ignored in Fig. 5 to save space, and we only mention some relative comparisons here: using the same 5,000 samples from the five-component random Gaussian mixture model, EM algorithm is 14 times slower than Lindsay's estimate, and the gmm package is more than 1,600 times slower; when two components completely overlap, EM algorithm is more than 100 times slower than Lindsay's estimate, and the gmm package is more than 2,500 times slower.

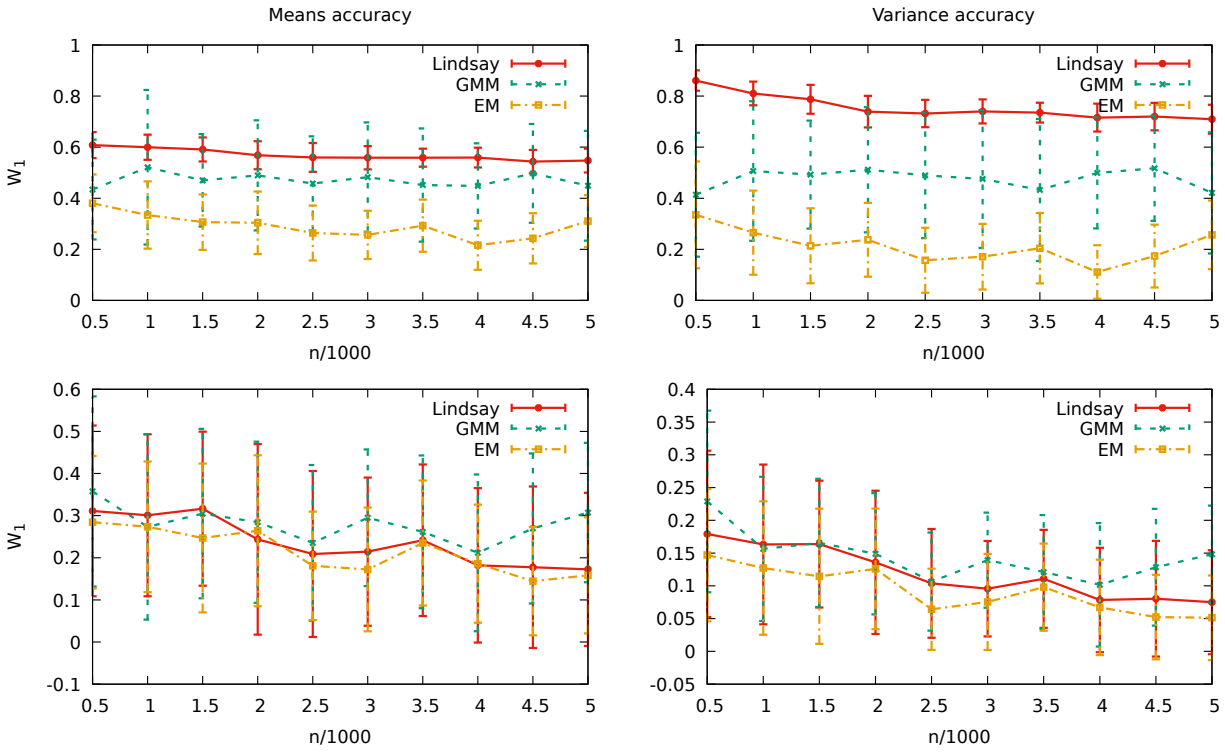


Figure 5: Comparison of different methods with unknown variance. The models in the first and second rows are the same as Fig. 3 and 4, respectively.

7 Extensions and discussions

7.1 Gaussian location-scale mixtures

This paper studies the optimal learning of the Gaussian location mixture model, a mixture of Gaussian with an arbitrary common variance. One immediate extension is the Gaussian location-scale mixture model with heteroscedastic components:

$$\sum_{i=1}^{k} w_i N(\mu_i, \sigma_i^2)$$

Parameter estimation for this model turns out to be significantly more difficult than the location mixture model, in particular

- Unbounded likelihood: It is well-known that the maximum likelihood estimator is ill-defined [KW56, p. 905]. For instance, consider $k = 2$, for any sample size n, we have

$$\sup_{p_1,p_2,\theta_1,\theta_2,\sigma} \prod_{i=1}^{n} \left[\frac{p_1}{\sigma_1} \varphi\left(\frac{X_i - \theta_1}{\sigma_1} \right) + \frac{p_2}{\sigma_2} \varphi\left(\frac{X_i - \theta_2}{\sigma_2} \right) \right] = \infty,$$

  achieved by, e.g., $\theta_1 = X_1, p_1 = 1/2$, $\sigma_2 = 1$, and $\sigma_1 \to 0$.

- In this model, identifiability is not yet completely settled. An unknown k-component Gaussian mixture model comprises $3k-1$ free parameters: k means, k variances, and k mixing weights summing up to one, so it is expected to be identified through its first $3k-1$ moments. It is recently resolved in [ARS16] that $3k-1$ moments can generally identify the Gaussian mixture distribution up to finitely many solutions, which is referred to as algebraic identifiability. In the case of two components, by a careful analysis of polynomial equations, six moments can uniquely identify the mixture distribution [Pea94], but this method is difficult to generalize in the case of more components. Identifying the mixture distribution uniquely, referred to as rational identifiability [ARS16], remains an open problem in general.

Besides the issue of identifiability, the optimal estimation accuracy under the Gaussian location-scale mixture model is resolved only in special cases. The sharp accuracy is only known in the case of two components to be $\Theta(n^{-1/12})$ for estimating means and $\Theta(n^{-1/6})$ for estimating variances [HP15], obtained by one robust variation based on Pearson's polynomial equations [Pea94]. In a k-component Gaussian mixture model, the accuracy is known to be $n^{-\Theta(1/k)}$ [MV10, KMV10], achieved by an exhaustive grid search on the parameter space. In addition to limited available accuracy results, [MV10, KMV10, HP15] all aim to recover parameters of all components (up to a global permutation), which necessarily requires many assumptions including lower bounds on mixing weights and differences between components. Those accuracy guarantees are inconvenient to apply in practice when no a priori knowledge is available on the unknown mixture model.

7.2 Multiple dimensions

The focus of this paper is a mixture of Gaussian on the real line, and a natural extension is a mixture of Gaussian in multiple dimensions. This is often studied by computer scientists under the framework of clustering, classification, or unsupervised learning, and usually requires nonoverlapping components [Das99, VW04]. One commonly used approach is dimensionality reduction: projecting data onto some lower dimensional subspace, clustering samples in that subspace, and

mapping back to the original space, and common choices of the subspace include random subspace and subspace from singular value decomposition. The approach using random subspace is analyzed in [Das99, AK01], and requires a pairwise separation polynomial in the dimensions; the subspace from singular value decomposition is analyzed in [VW04, KSV05, AM05, BV08], and requires a pairwise separation polynomial in the number of components.

When components are allowed to overlap significantly, the random projection approach is also adopted by [MV10,KMV10,HP15], where the learning problem in high dimensions is reduced to the real line, so the univariate learning algorithm is a foundational subroutine. We provide an algorithm using similar random projection method to learn a Gaussian mixture model in d dimensions in Algorithm 4 in the case of known variance, using the univariate algorithm in Section 4.1 as a subroutine, and obtain the estimation accuracy in Theorem 8. The unknown variance case can be obtained analogously, where we instead using the univariate algorithm in Section 4.2. However, the dependence on the dimension using the random projection method, which estimates each coordinate independently, is highly suboptimal.[3]

Algorithm 4 Learning a Gaussian mixture model in d dimension.

Input: n samples $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ drawn i.i.d. from a d-dimensional Gaussian mixture model, common covariance matrix Σ, and separation parameter τ.

Output: estimated Gaussian mixture model with weights and centers $(\hat{w}_j, \hat{\mu}_j)$.

Let $(b_1, \ldots, b_d)$ be a set of random orthonormal basis in $\mathbb{R}^d$.
Let $r = \frac{1}{\sqrt{d}} \sum_i b_i$.
Run the univariate algorithm in Section 4.1 with n projected samples $\langle X_1, r\rangle, \ldots, \langle X_n, r\rangle$ and common variance $r^\top \Sigma r$ to get weights and centers $\{(w_j, \mu_j)\}$.
Reordering the components such that $\mu_1 < \mu_2 < \cdots < \mu_k$.
Initialize k estimated components with weights $\hat{w}_j = w_j$ and centers $\hat{\mu}_j = 0$.
for $i = 1$ **to** d **do**
  Let $r' = r + \tau b_i$.
  Run the univariate algorithm in Section 4.1 with n projected samples $\langle X_1, r'\rangle, \ldots, \langle X_n, r'\rangle$ and common variance $r'^\top \Sigma r'$ to get centers $\{\mu'_j\}$ (weights are ignored).
  Reordering the components such that $\mu'_1 < \mu'_2 < \cdots < \mu'_k$.
  Let $\hat{\mu}_j = \hat{\mu}_j + b_i \frac{\mu'_j - \mu_j}{\tau}$ for $j = 1, \ldots, k$.
end for

Theorem 8. *Suppose in a k-component Gaussian mixture model in $\mathbb{R}^d$ with weights and centers (w_j, μ_j), the centers are bounded by M and pairwise separated by ϵ in ℓ_2-norm, and mixing weights are lower bounded by ϵ'. Given n i.i.d. samples drawn from that Gaussian mixture model with $n > (\Omega_k(\frac{M\sqrt{d}}{\delta\epsilon\epsilon'}))^{4k-2} \log \frac{d}{\delta}$, with probability at least $1 - 2\delta$, Algorithm 4 with $\tau = \frac{\delta\epsilon}{2Mk^2\sqrt{d}}$ yields $(\hat{w}_j, \hat{\mu}_j)$ with accuracy*

$$|w_j - (\pi\hat{w})_j| < O_k\left(\frac{M\sqrt{d}}{\delta\epsilon}\epsilon_n\right), \quad \|\mu_j - (\pi\hat{\mu})_j\|_2 < O_k\left(\frac{M^2 d}{\delta\epsilon\epsilon'}\epsilon_n\right),$$

for the same permutation π, where $\epsilon_n = \min\{(\frac{n}{\log(d/\delta)})^{-\frac{1}{4k-2}}, (\frac{\sqrt{d}}{\delta\epsilon})^{2k-2}\sqrt{\frac{\log(d/\delta)}{n}}\}$.

[3] Specifically, in d dimensions, estimating each coordinate independently suffers an ℓ_2 loss proportional to $\sqrt{d}$. However, it is possible to achieve $d^{1/4}$ using spectral methods in specific examples. See Lemma 29.

Proof. By the distribution of random projection in Lemma 28 and the union bound, with probability at least $1-\delta$, we have $|\langle \mu_i - \mu_j, r\rangle| > \frac{2\delta\epsilon}{k^2\sqrt{d}} \triangleq 2\epsilon_r$ for all pairs $i \neq j$, where ϵ_r is referred to as the separation of centers on this random direction. Without loss of generality, assume $\langle \mu_1, r\rangle < \cdots < \langle \mu_k, r\rangle$. By Theorem 1, with probability at least $1 - \frac{\delta}{d+1}$, the accuracy in W_1 distance is at most $O_k(M(\frac{n}{\log(d/\delta)})^{-\frac{1}{4k-2}})$, which is less than $0.1\epsilon_r\epsilon'$ when $n > (\Omega_k(\frac{M\sqrt{d}}{\delta\epsilon\epsilon'}))^{4k-2}\log\frac{d}{\delta}$. Then it follows from Theorem 2 that the W_1 distance is at most $O_k(M\epsilon_r^{2-2k}\sqrt{\frac{\log(k/\delta)}{n}})$, and thus is upper bounded by $O_k(M\epsilon_n)$. Denote the estimated weights on direction r by $\hat{w}_1, \ldots, \hat{w}_k$, and centers by $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$, and it follows from Lemma 1 that, after reordering components,

$$|\langle \mu_j, r\rangle - \tilde{\mu}_j| < O_k(M\epsilon_n/\epsilon'), \quad |w_j - \hat{w}_j| < O_k(M\epsilon_n/\epsilon_r). \tag{20}$$

On each direction $r_\ell = r + \tau b_\ell$, the centers are separated by at least $|\langle \mu_i - \mu_j, r_\ell\rangle| > 2\epsilon_r - 2M\tau > \epsilon_r$, so the same accuracy of the estimated centers in (20) also holds on each direction r_ℓ. Note that, the population mean of each component can be represented by $\mu_j = \sum_\ell b_\ell \langle b_\ell, \mu_j\rangle = \sum_\ell b_\ell \frac{\langle r_\ell, \mu_j\rangle - \langle r, \mu_j\rangle}{\tau}$. Therefore, $\|\hat{\mu}_j - \mu_j\|_2 \le \sqrt{d}\frac{O_k(M\epsilon_n/\epsilon')}{\tau} = \frac{\sqrt{d}M}{\epsilon_r}O_k(M\epsilon_n/\epsilon')$. □

It is interesting to directly extend the moment based method to high dimensions, which is challenging in both aspects of theory and algorithm. When component centers are in general position, the tensor structure in moments of spherical Gaussian mixture models is studied by [HK13] and the model is polynomially learnable using spectral decomposition method. To apply our method in high dimensions, the challenge is on the moments comparison results analogous to Proposition 1 and 2, the key step leading to the optimal rate. Both moments comparison results are proved by the primal formulation of the Wasserstein distance and its simple formula in (13), which is only a special case on the real line and no longer holds in multiple dimensions [Vil03]; Proposition 1 is alternatively proved through the dual approach (12) and the proof relies on the Newton's interpolation formula, which is again difficult to generalize to high dimensions. For an efficient learning algorithm, we rely on the semidefinite programming to denoise the vector of estimated moments. However, it remains an open problem how to conveniently describe the moment space in multiple dimensions [Las09], and it is unclear whether the moment space of discrete distributions is still convex, and how to find a quadrature rule from moments.

7.3 General finite mixtures

Though this paper focuses on learning Gaussian location mixture models, the moments comparison theorems in Section 3 are independent of properties of Gaussian. As long as moments of the latent distribution are estimated accurately, similar theory and algorithms can be obtained. Unbiased estimate of moments exists in many mixture models with widespread applications, including exponential mixtures [Jew82], Poisson mixtures [KX05], Gaussian scale mixtures [AM74], and notably the quadratic variance exponential family (QVEF) whose variance is at most a quadratic function of the mean, such as Gaussian, Poisson, gamma, binomial, and negative binomial [Mor82, (8.8)].

As a closely related topic of this paper, we discuss the Gaussian scale mixture model in details, which has been extensively studied in statistics literature [AM74] and is widely used in image and video processing [WS00, PSWS03]. In a Gaussian scale mixture, components are assumed to have the same mean, so samples from different components are significantly overlapped, and clustering based algorithms such as [Das99] no longer work in this problem. Nevertheless, our denoised method of moments in Section 4.1 is easily generalizable. Suppose the unknown Gaussian scale mixture is $X = \sqrt{V}Z$, where V is a k-atomic discrete latent distribution of the variances that are bounded, and

Z is independently standard normal. Note that the population moments are $\mathbb{E}[X^{2r}] = \mathbb{E}[V^r]\mathbb{E}[Z^{2r}]$, so an immediate unbiased estimate for the moments of V is $\hat{m}_r = \frac{1}{n}\sum_{i=1}^n X_i^{2r}/\mathbb{E}[Z^{2r}]$ with accuracy $O_r(1/\sqrt{n})$. Applying moments projection and then computing the quadrature rule analogously to the denoised method of moments in Section 4.1, we obtain an accuracy

$$W_1(V, \hat{V}) \le O_k(n^{-\frac{1}{4k-2}}),$$

with high probability. Moreover, the tightness of this result can be analogously obtained as in Section 5. To see this, let V and V' be two different k-atomic distributions on $[0, 1]$ with identical first $2k-2$ moments. Then the two mixture models $\sqrt{\epsilon V}Z$ and $\sqrt{\epsilon V'}Z$ coincide on their first $4k-3$ moments, and they are both $\sqrt{\epsilon}$-subgaussian. Applying Lemma 9 yields that any estimate with $(O_k(1/\sqrt{\epsilon}))^{8k-4}$ samples suffers a W_1 loss at least $\Omega_k(\epsilon)$ under either of the two models. By choosing $\epsilon = \Theta_k(n^{-\frac{1}{4k-2}})$, we conclude that the W_1 loss is at least $\Omega_k(n^{-\frac{1}{4k-2}})$ in the worst case.

8 Proofs

8.1 Proof of moments comparison theorems

Proposition 1 is proved under a slightly weaker assumption that X and Y both take values in a subset of $2k$ points in $[-1, 1]$, which immediately follows from Proposition 8. We provide three proofs, where the first two are based on the primal (coupling) formulation of W_1 distance (13), and the third on the dual formulation (12). In all proofs, the Wasserstein distance is analyzed by proxy of polynomials, whose expected values are linear combination of moments. Specifically,

- The first proof uses polynomials to interpolate step functions, whose expected values are the distribution function values. The closeness of moments imply the closeness of distribution functions and thus, by (13), a small Wasserstein distance. Similar idea applies to the proof of Proposition 2.
- The second proof finds a polynomial that preserves the sign of the difference between two distribution functions, and the Wasserstein distance given by the integral of the absolute difference (13) pertains to the integral of that polynomial.
- The third proof uses polynomials to approximate 1-Lipschitz functions, whose expected values are connected to the Wasserstein distance by the dual formulation (12). A small distance is promised by a small approximation error and close moments.

Proposition 8. *Suppose the union of supports of X and Y are ℓ points in $[-1, 1]$. Then*

$$\max_{i\in[\ell-1]} |m_i(X) - m_i(Y)| \ge (W_1(X, Y)/O(\ell))^{\ell-1}.$$

First proof of Proposition 8. Suppose $\max_{i\in[\ell-1]} |m_i(X) - m_i(Y)| = \delta$. Denote the union of supports of X and Y by $t_1 < t_2 < \cdots < t_\ell$. Then, with F_X and F_Y denoting the distribution functions of X and Y, respectively,

$$W_1(X, Y) = \sum_{r=1}^{\ell-1} |F_X(t_r) - F_Y(t_r)| \cdot |t_{r+1} - t_r|. \tag{21}$$

Construct a polynomial P_r of degree $\ell-1$ satisfying $P_r(t_1)=\cdots=P_r(t_r)=1$ and $P_r(t_{r+1})=\cdots=P_r(t_\ell)=0$, which determines the polynomial uniquely. Then, almost surely, $P_r(X)=\mathbf{1}_{\{X\le t_r\}}$ and $P_r(Y)=\mathbf{1}_{\{Y\le t_r\}}$. The polynomial P_r has a formula in Newton form in terms of divided differences

$$P_r(x)=\sum_{i=1}^{\ell}P_r[t_1,\dots,t_i]\prod_{j=1}^{i-1}(x-t_j)=\sum_{i=r+1}^{\ell}P_r[t_1,\dots,t_i]\prod_{j=1}^{i-1}(x-t_j),$$

where the last equality is due to $P_r[t_1,\dots,t_i]=0$ for $i\le r$. When each $|t_j|\le 1$, applying moments differences between X and Y, the expected values of the polynomial $\prod_{j=1}^{i-1}(x-t_j)$ between X and Y differ by at most $2^{i-1}\delta$. Then, applying the upper bound of the coefficients in Corollary 2,

$$\Delta_r\triangleq|\mathbb{E}[P_r(X)]-\mathbb{E}[P_r(Y)]|\le\sum_{i=r}^{\ell-1}\frac{\binom{i-1}{r-1}2^i\delta}{(t_{r+1}-t_r)^i}\le\sum_{i=r}^{\ell-1}\frac{2^{i-1}2^i\delta}{(t_{r+1}-t_r)^i}.\tag{22}$$

The right-hand side of (22) is further upper bounded by $\frac{\ell 4^{\ell-1}\delta}{(t_{r+1}-t_r)^{\ell-1}}$ since $|t_{r+1}-t_r|\le 2$. Then,

$$W_1(X,Y)=\sum_{r=1}^{\ell-1}(\Delta_r\wedge 1)\cdot|t_{r+1}-t_r|\le\sum_{r=1}^{\ell-1}\left(\frac{\ell 4^{\ell-1}\delta}{(t_{r+1}-t_r)^{\ell-1}}\wedge 1\right)\cdot|t_{r+1}-t_r|\le O(\ell)\delta^{\frac{1}{\ell-1}}.\tag{23}$$

The conclusion follows. □

Lemma 10. *Let $t_1\le t_2\le\dots$ be an ordered sequence (not necessarily distinct). Suppose $f(t_i)=1$ for $i\le r$, $f(t_i)=0$ for $i\ge r+1$, and $f[t_i,\dots,t_j]=0$ for both $i<j\le r$ and $j>i\ge r+1$. Then,*

$$f[t_i,\dots,t_j]=(-1)^{i-r}\sum_{L\in\mathcal{L}(i,j)}\prod_{(x,y)\in L}\frac{1}{t_x-t_y},\quad i\le r<r+1\le j,\tag{24}$$

where $\mathcal{L}(i,j)$ is the set of lattice paths[4] from $(r,r+1)$ to (i,j) using steps $(0,1)$ and $(-1,0)$.

Proof. Construct a upper triangular matrix with $a_{i,j}=f[t_i,\dots,t_j]$ when $i\le j$:

$$\begin{bmatrix}1&0&\cdots&0&a_{1,r+1}&\cdots&\\&1&\ddots&\vdots&\vdots&&\\&&1&0&a_{r-1,r+1}&\cdots&\\&&&1&a_{r,r+1}&\cdots&\\&&&&0&\cdots&0\\&&\mathbf{0}&&&\ddots&\vdots\\&&&&&&0\end{bmatrix}.$$

The value for $i\le r$ and $j\ge r+1$ can be obtained recursively by

$$a_{i,j}=\frac{a_{i,j-1}-a_{i+1,j}}{t_i-t_j}.\tag{25}$$

In the following, we prove (24) by induction using the recursive formula (25). We can directly compute the base cases: if $i=r$, we have $a_{i,j}=\prod_{v=r+1}^{j}\frac{1}{t_r-t_v}$; if $j=r+1$, we have $a_{i,j}=(-1)^{i-r}\prod_{v=i}^{r}\frac{1}{t_v-t_{r+1}}$. Suppose (24) holds for both $a_{i,j-1}$ and $a_{i+1,j}$. We evaluate $a_{i,j}$ using (25), which equals the desired result in (24). □

[4]For $a,b\in\mathbb{N}^2$, a lattice path from a to b using a set of steps S is a sequence $a=x_1,x_2,\dots,x_n=b$ with all increments $x_{j+1}-x_j\in S$.

Corollary 2. *Let f be the same as Lemma 10. Then $f[t_1, \ldots, t_i] = 0$ for $i \le r$, and*

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

Proof. Since the sequence $t_1 \le t_2 \le \ldots$ is ordered, then each of the $\binom{(r-1)+(i-(r+1))}{r-1}$ terms in the summation of (24) is at most $\frac{1}{(t_{r+1}-t_r)^{i-1}}$ in magnitude. □

Second proof of Proposition 8. Denote the union of supports of X and Y by $t_1 < t_2 < \cdots < t_\ell$. The distribution functions of X and Y, denoted by F_X and F_Y, possibly differ at $t_2, \ldots, t_{\ell-1}$. Then there exists a polynomial L of degree at most $\ell' \le \ell - 2$ satisfying $(F_X(x) - F_Y(x))L(x) \ge 0$ for all t, where the polynomial $L(x)$ is the product of $x - t_i$ over a subset of $\{t_2, \ldots, t_{\ell-1}\}$. Suppose $W_1(X, Y) = \delta$, from (21), then there exists r such that $|F_X(t_r) - F_Y(t_r)| \cdot |t_{r+1} - t_r| \ge \frac{\delta}{\ell-1}$.

Let $P(x) \triangleq \int_{t_1}^{x} L(t)\mathrm{d}t$, a polynomial of degree $\ell' + 1$. Since all $|t_j| \le 1$, the expected values of the polynomial P between X and Y differ by at most $2^{\ell'}$ times the maximal differences between the first $\ell' + 1$ moments of X and Y. Applying Fubini's theorem,

$$\mathbb{E}[P(X)] = \int_{t_1}^{t_\ell} P(x)\mathrm{d}F_X(x) = \int_{t_1}^{t_\ell} \int_{t_1}^{t_\ell} L(t)\mathbf{1}_{\{t \le x\}}\mathrm{d}t\mathrm{d}F_X(x) = \int_{t_1}^{t_\ell} L(t)(1 - F_X(t))\mathrm{d}t.$$

Since $(F_X(x) - F_Y(x))L(x)$ is always non-negative and $|F_X(t_r) - F_Y(t_r)| \cdot |t_{r+1} - t_r| \ge \frac{\delta}{\ell-1}$, the difference between the expected values of $P(X)$ and $P(Y)$ is

$$|\mathbb{E}[P(X)] - \mathbb{E}[P(Y)]| = \int_{t_1}^{t_\ell} |L(t)(F_X(t) - F_Y(t))|\mathrm{d}t \ge \frac{\delta}{\ell - 1} \int_{t_r}^{t_{r+1}} \frac{|L(t)|}{t_{r+1} - t_r}\mathrm{d}t.$$

When t is between t_r and t_{r+1}, the polynomial $L(t)$ is at least $(t_{r+1} - t)^a(t - t_r)^b$ in magnitude, for some $a \le r-1$ and $b \le k-r-1$ sum up to the degree of $L(x)$. The integral $\int_{t_r}^{t_{r+1}} (t_{r+1} - t)^a(t - t_r)^b\mathrm{d}t$ can be evaluated to be $B(a+1, b+1)(t_{r+1} - t_r)^{a+b+1}$, where $B(\cdot, \cdot)$ is the beta function. Then, $|\mathbb{E}[P(X)] - \mathbb{E}[P(Y)]| \ge \frac{\delta}{\ell-1}(t_{r+1} - t_r)^{a+b}B(a+1, b+1)$. Note that $|t_{r+1} - t_r| \ge \frac{\delta}{\ell-1}$, and thus $|\mathbb{E}[P(X)] - \mathbb{E}[P(Y)]| \ge (\frac{\delta}{O(\ell)})^{a+b+1}$. The conclusion follows since $a + b + 1 = \ell' + 1 \le \ell - 1$. □

Third proof of Proposition 8. Suppose $\max_{i \in [\ell-1]} |m_i(X) - m_i(Y)| = \delta$. Denote the union of supports of X and Y by $t_1 < t_2 < \cdots < t_\ell$. Fix any 1-Lipschitz function f. The expected values of $f(X)$ and $f(Y)$ only pertain to function values $f(t_1), \ldots, f(t_\ell)$, which can be interpolated by a polynomial of degree $\ell - 1$. However, the coefficients of the interpolating polynomial are possibly arbitrarily large.[5] To fix this, we slightly modify the value of f on those discrete points into another function denoted by $\tilde{f}$, and then interpolate $\tilde{f}$ by a polynomial P with some upper bounds on the coefficients. Then the difference between expected values of f is upper bounded by

$$|E[f(X)] - E[f(Y)]| \le 2 \max_{x \in \{t_1, \ldots, t_\ell\}} |\tilde{f}(x) - f(x)| + |\mathbb{E}[P(X)] - \mathbb{E}[P(Y)]|.$$

To this end, we define the values of $\tilde{f}$ recursively by $\tilde{f}(t_1) = f(t_1)$ and $\tilde{f}(t_i) = \tilde{f}(t_{i-1}) + (f(t_i) - f(t_{i-1}))\mathbf{1}_{\{t_i - t_{i-1} > \tau\}}$, where $\tau \le 2$ is a parameter we will optimize later. The definition of $\tilde{f}$ implies that $|\tilde{f}(x) - f(x)| \le \tau\ell$ for $x \in \{t_1, \ldots, t_\ell\}$. The polynomial P of degree $\ell - 1$ can be represented in Newton's formula: $P(x) = \sum_{i=1}^{\ell} P[t_1, \ldots, t_i] \prod_{j=1}^{i-1}(x - t_j)$. Since all $|t_j| \le 1$, applying moments differences between X and Y, the expected values of the polynomial $\prod_{j=1}^{i-1}(x - t_j)$ between X and Y

[5] For example, the polynomial to interpolate $f(-\epsilon) = f(\epsilon) = \epsilon, f(\epsilon) = 0$ is $P(x) = x^2/\epsilon$.

differ by at most $2^{i-1}\delta$. Since f is 1-Lipschitz, the first-order finite differences $P[t_i, t_{i+1}]$ are bounded by one in magnitude. Then, by induction and the definition of P, we have $|P[t_i, \dots, t_{i+j}]| \le (\frac{2}{\tau})^{j-1}$. Therefore,

$$|E[f(X)] - E[f(Y)]| \le 2\tau\ell + \sum_{i=2}^{\ell} \left(\frac{2}{\tau}\right)^{i-2} 2^{i-1}\delta \le 2\ell \left(\tau + \frac{4^{\ell-2}}{\tau^{\ell-2}}\delta\right).$$

The conclusion follows by letting $\tau = 4\delta^{\frac{1}{\ell-1}}$. □

The proof of Proposition 2 uses a similar idea as the first proof of Proposition 8 to approximate step functions. It is possible for an interpolating polynomial being almost surely equal to a step function under a discrete measure, but such equality is impossible under an arbitrary measure. A classical method to bound a distribution function by moments is to choose a polynomial that majorizes (or minorizes) a step function, and thus the expected value is a upper (or lower) bound of that distribution function, for example, in the proof of Chebyshev-Markov-Stieltjes inequality (see, e.g., [Akh65]). In our case, to bound the difference between two distribution functions with one being discrete, we further select the majorizing (or minorizing) polynomial to be almost surely equal to the step function under that discrete measure. See Fig. 6 for an illustration.

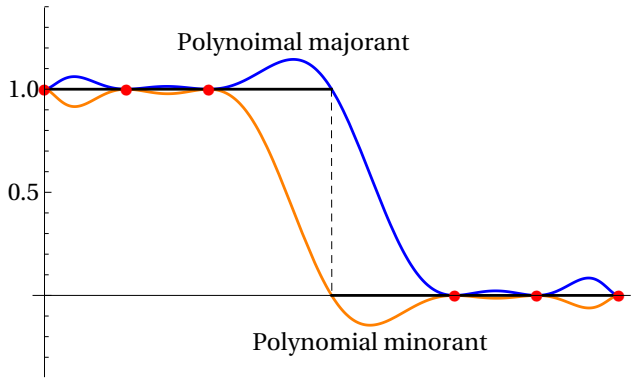


Figure 6: Polynomial majorant and minorant that coincide with the step function on 6 red points. The polynomials are of degree 12 and are obtained by Hermite interpolation in Section 2.2.

Proof of Proposition 2. Suppose $\max_{i\in[2k]} |m_i(X) - m_i(Y)| = \delta$, and X is supported on $x_1 < x_2 < \cdots < x_k$. For any t not equal to any of those x_i, construct polynomials P_t and Q_t of degree $2k$, satisfying the following equations which determine them uniquely:

$$\begin{aligned}
&P_t(x_i) = Q_t(x_i) = 1, && i \le m,\\
&P_t(t) = 1, Q_t(t) = 0,\\
&P_t(x_i) = Q_t(x_i) = 0, && i \ge m+1,\\
&P_t'(x_i) = Q_t'(x_i) = 0, && i = 1, \dots, k.
\end{aligned}$$

It is essential that $P_t(x) \ge \mathbf{1}_{\{x\le t\}}$ and $Q_t(x) \le \mathbf{1}_{\{x\le t\}}$ by Rolle's theorem (see a proof in [Akh65, p. 65]). By the Hermite interpolation formula, $P_t(x) = Q_t(x) + R_t(x)$, where $R_t(x) = \prod_i (\frac{x-x_i}{t-x_i})^2$. Then, with F_X and F_Y denoting the distribution functions of X and Y, respectively,

$$\begin{aligned}
&\mathbb{E}[Q_t(Y)] \le F_Y(t) \le \mathbb{E}[P_t(Y)] = \mathbb{E}[Q_t(Y)] + \mathbb{E}[R_t(Y)],\\
&\mathbb{E}[Q_t(X)] \le F_X(t) \le \mathbb{E}[P_t(X)] = \mathbb{E}[Q_t(X)].
\end{aligned}$$

Let $f(t) \triangleq |\mathbb{E}[Q_t(Y)] - \mathbb{E}[Q_t(X)]|$, and $g(t) \triangleq \mathbb{E}[R_t(Y)]$. Then the difference between distribution functions of X and Y is at most

$$|F_Y(t) - F_X(t)| \le (f(t) + g(t)) \wedge 1 \le f(t) \wedge 1 + g(t) \wedge 1. \tag{26}$$

When X takes values in $[-1, 1]$, the expected values of the polynomial Q_t between X and Y differ by at most 2^{2k}. Then $\mathbb{E}[\prod_i (Y - x_i)^2] \le 2^{2k}\delta$, and

$$\int (g(t) \wedge 1)\mathrm{d}t \le \int \left(\frac{2^{2k}\delta}{\prod_{i=1}^k (t - x_i)^2} \wedge 1 \right) \mathrm{d}t \le O(k)\delta^{\frac{1}{2k}}, \tag{27}$$

where the last inequality follows from Lemma 19. The polynomial Q_t (and also P_t) can be written in Newton form by generalized divided differences using Neville's algorithm. To this end, we expand $x_1 < \cdots < x_m < t < x_{m+1} < \cdots < x_k$ by replacing each x_i by two copies of itself, and denote the new sequence of length $2k+1$ by $t_1 = t_2 < t_3 = t_4 < \cdots < t_{2k} = t_{2k+1}$, where $t_{2m+1} = t$. Then,

$$Q_t(x) = \sum_{i=1}^{2k+1} Q_t[t_1, \ldots, t_i] \prod_{j=1}^{i-1} (x - t_j) = \sum_{i=2m+1}^{2k+1} Q_t[t_1, \ldots, t_i] \prod_{j=1}^{i-1} (x - t_j),$$

where the last equality is due to $Q_t[t_1, \ldots, t_i] = 0$ for $i \le 2m$. Applying the upper bound of the coefficients in Corollary 2 and the moments differences between X and Y,

$$f(t) \le \sum_{i=2m}^{2k} \frac{\binom{i-1}{2m-1} 2^i \delta}{(t - x_m)^i}, \quad x_m < t < x_{m+1}; \tag{28}$$

if $t < x_1$ then $Q_t(x) = 0$, and thus $f(t) = 0$. By analogous calculation to (23) and (27), we also have $\int (f(t) \wedge 1)\mathrm{d}t \le O(k)\delta^{\frac{1}{2k}}$. Since $W_1(X, Y) = \int |F_X(t) - F_Y(t)|\mathrm{d}t$, applying (26) yields that

$$W_1(X, Y) \le O(k)\delta^{\frac{1}{2k}}.$$

The conclusion follows. □

Proposition 4 is proved analogous to the first proof of Proposition 8, apart from a careful analysis of polynomial coefficients under separation assumptions. Proposition 5 is proved similarly by an extension of the proof of Proposition 2.

Proof of Proposition 4. In the the first proof of Proposition 8, the difference between distribution functions Δ_r in (22) is upper bounded by

$$\Delta_r \le \frac{\ell 4^{\ell - 1}\delta}{(t_{r+1} - t_r)^{\ell'} \gamma^{\ell - \ell' - 1}},$$

since all but ℓ' atoms are at least γ away from t_r. The conclusion follows analogously. □

Proof of Proposition 5. In the proof of Proposition 2, since any t is at least γ away from all but k' atoms, the integral in (27) can be upper bounded by

$$\int (g(t) \wedge 1)\mathrm{d}t \le O(k) \left(\frac{2^{2k}\delta}{\gamma^{2(k-k')}} \right)^{1/(2k')}.$$

Similarly, the moment difference in (28) is upper bounded by $\frac{2k4^{2k-1}\delta}{(t - x_m)^{2k'}\gamma^{2(k-k')}}$, and we have

$$\int (f(t) \wedge 1)\mathrm{d}t \le O(k) \left(\frac{k4^{2k-1}\delta}{\gamma^{2(k-k')}} \right)^{1/(2k')}.$$

The conclusion follows analogously. □

8.2 Proof and extension of Proposition 3

Without loss of generality, we shall assume $\sigma \geq \sigma'$. Then, in distribution, $X = U + \tau Z + \sigma' Z'$ and $X' = U' + \sigma' Z'$, where Z and Z' are independent standard normal, and $\tau = \sqrt{\sigma^2 - \sigma'^2}$. Since moments of X and X' are close, the moments of $U + \tau Z$ and U', which can be evaluated by the Hermite polynomial given in (15), are also close to each other by Lemma 18. The magnitude of τ, which quantifies the accuracy of σ^2, can be upper bounded by the accuracy of the first $2k$ moments, as shown in the following Lemma 11. Moreover, it follows from Proposition 2 that the distributions of $U + \tau Z$ and U' are close, in W_1 distance. Then immediately the distributions of U and U' are also close, since $W_1(U + \tau Z, U) = \Theta(\tau)$.

Lemma 11 (Accuracy on σ^2)**.** *Suppose U' takes k values in $[-1, 1]$ and Z is a standard normal random variable independent of U. Then*

$$\max_{i \in [2k]} |m_i(U + \tau Z) - m_i(U')| \geq (\tau / O(\sqrt{k}))^{2k}.$$

Proof. Suppose U' takes values in $\{x'_1, \ldots, x'_k\}$, and $\max_{i \in [2k]} |m_i(U + \tau Z) - m_i(U')| = \delta$. Let $P(x) = \prod_{i=1}^k (x - x'_i)^2 = \sum_{i=0}^{2k} a_i x^i$. Note that $P(U') = 0$ almost surely. When U' takes values in $[-1, 1]$, we have

$$\mathbb{E}[P(U + \tau Z)] = \mathbb{E}[P(U + \tau Z)] - \mathbb{E}[P(U')] \leq 2^{2k} \delta.$$

Since U is independent of Z, we have

$$\min_U \mathbb{E}[P(U + \tau Z)] = \min_x \mathbb{E}[P(x + \tau Z)] \geq \tau^{2k} \min_{y_1, \ldots, y_k} \mathbb{E}\left[\prod_i (Z + y_i)^2\right] = k! \tau^{2k},$$

where the last step used the next lemma. □

Lemma 12.

$$\min\{\mathbb{E}[p^2(Z)] : \deg(p) \leq k,\ p \text{ is monic}\} = k!$$

achieved by $p = H_k$.

Proof. Since p is monic, it can be written as $p = H_k + \sum_{j=0}^{k-1} \alpha_j H_j$. By orthogonality and the fact that $\mathbb{E}[H_j^2(Z)] = j!$, we have $\mathbb{E}[p^2(Z)] = k! + \sum_{j=0}^{k-1} \alpha_j^2 j!$ and the conclusion follows. □

Proof of Proposition 3. Without loss of generality, assume $\sigma \geq \sigma'$. Then X is in distribution equal to $U + \tau Z' + \sigma' Z$, where $\tau = \sqrt{\sigma^2 - \sigma'^2}$. By Lemma 18, $|m_r(U + \tau Z') - m_r(U')| \leq (O(\sqrt{r}))^r \delta$ for $r = 1, \ldots, 2k$. Let $\delta' = (O(\sqrt{k}))^{2k} \delta$. By Lemma 11, $\tau \leq O(\sqrt{k}) \delta'^{\frac{1}{2k}}$. By Proposition 2, $W_1(U + \tau Z, U') \leq O(k) \delta'^{\frac{1}{2k}}$. Since $W_1(U + \tau Z, U) \leq O(\tau)$, then, by triangle inequality, $W_1(U, U') \leq O(k) \delta'^{\frac{1}{2k}} + O(\tau) = O(k) \delta'^{\frac{1}{2k}}$. □

For an adaptive rate of Theorem 2 in the case of unknown variance, we need an extension of Proposition 3 when both Gaussian mixture models have k_0 clusters of centers, given by the following Proposition 9. The proof is also similar to Proposition 3, only that the moment comparison result in Proposition 2 is replaced by Proposition 5.

Proposition 9. *Under the conditions Proposition 3, if U and U' each has k_0 clusters of atoms and intercluster atoms are separated by at least γ, then,*

$$\sqrt{|\sigma^2 - \sigma'^2|},\ W_1(U, U') \leq O_k\left(\left(\frac{\delta}{\gamma^{2(k_0 - 1)}}\right)^{\frac{1}{2(k - k_0 + 1)}}\right).$$

Proof. Following similar proof of Proposition 3 above, let $\delta' = (O(\sqrt{k}))^{2k}\delta$. When U' has k_0 clusters of atoms, the largest one is of size at most $k - k_0 + 1$. For any $t \in \mathbb{R}$, it can be $\gamma/2$ close to at most $k - k_0 + 1$ atoms of U'. Then, it follows from Proposition 5 that

$$W_1(U + \tau Z', U') \le O(k) \left(\frac{k 4^{2k-1} \delta'}{\gamma^{2(k_0-1)}} \right)^{\frac{1}{2(k-k_0+1)}}.$$

The upper bound for τ follows from Lemma 13, and the upper bound for $W_1(U, U')$ follows from the triangle inequality. □

Lemma 13. *If U' is k-atomic, then*

$$W_1(U + \tau Z, U') \ge \Omega_k(\tau).$$

Proof. Note that W_1 distance is shift invariant, and $W_1(aX, aY) = |a| W_1(X, Y)$ for $a \in \mathbb{R}$. Then, for any $x \in \mathbb{R}$, we have $\inf W_1(x + \tau Z, Y) = \tau \inf W_1(Z, Y) \ge \Omega_k(\tau)$, where the infimum is over Y supported on at most k atoms. Therefore, for any coupling of $U + \tau Z$ and U', we have

$$\mathbb{E}|U + \tau Z - U'| = \mathbb{E}[\mathbb{E}[|U + \tau Z - U'||U]] \ge \Omega_k(\tau).$$ □

8.3 Proof of density estimation

Lemma 14 (Bound χ^2-divergence using moments difference)**.** *Suppose all moments of random variables U and U' exist, and U' is centered with variance σ^2. Then,*

$$\chi^2(U + Z \| U' + Z) \le e^{\frac{\sigma^2}{2}} \sum_{j \ge 1} \frac{(\Delta m_j)^2}{j!},$$

where $\Delta m_j = m_j(U) - m_j(U')$ denotes the j^{th} moment difference between U and U', Z is a standard normal random variable, and $U, U' \perp\!\!\!\perp Z$.

Proof. The proof applies orthogonal polynomials expansion that is also used in [WV10, CL11]. Denote the densities of two mixture distributions by $f(x) = \mathbb{E}[\phi(x - U)]$ and $g(x) = \mathbb{E}[\phi(x - U')]$, where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the density of the standard normal distribution. By the exponential generating function of the Hermite polynomials (see, e.g., [AS64, 22.9.17]), we have $\phi(x - t) = \phi(x) \sum_{k \ge 0} H_k(x) \frac{t^k}{k!}$. Then,

$$f(x) - g(x) = \phi(x) \sum_{j \ge 1} H_j(x) \frac{\Delta m_j}{j!}.$$

Since the function $x \mapsto e^x$ is convex, we obtain by Jensen's inequality that

$$g(x) = \phi(x) \mathbb{E}[\exp(U'x - U'^2/2)] \ge \phi(x) \exp(-\sigma^2/2).$$

Then, the χ^2-divergence is upper bounded by

$$\chi^2(U + Z \| U' + Z) = \int \frac{(f(x) - g(x))^2}{g(x)} \mathrm{d}x \le e^{\frac{\sigma^2}{2}} \int \phi(x) \left(\sum_{j \ge 1} H_j(x) \frac{\Delta m_j}{j!} \right)^2.$$

The conclusion follows by the orthogonality of Hermite polynomials. □

Lemma 15. *If U and U' each takes at most k values on $[-1,1]$, and the first $2k-1$ moments between U and U' differ be at most ϵ, then, for any $\ell \geq 2k$,*

$$|\mathbb{E}[U^\ell] - \mathbb{E}[U'^\ell]| \leq 3^\ell \epsilon.$$

Proof. Let $f(x) = x^\ell$ and denote the support of U and U' by $x_1 < \cdots < x_{k'}$ for $k' \leq 2k$. The function f can be interpolated on $x_1, \ldots, x_{k'}$ using a polynomial P of degree at most $2k-1$, where

$$P(x) = \sum_{i=1}^{k'} f[x_1, \ldots, x_i] \prod_{j=1}^{i-1}(x - x_j) = \sum_{i=1}^{k'} \frac{f^{(i-1)}(\xi_i)}{(i-1)!} \prod_{j=1}^{i-1}(x - x_j),$$

for some $\xi_i \in [x_1, \ldots, x_i]$ by Newton's interpolation formula. Therefore,

$$|\mathbb{E}[U^\ell] - \mathbb{E}[U'^\ell]| = |\mathbb{E}[P(U)] - \mathbb{E}[P(U')]| \leq \sum_{i=1}^{k'} \binom{\ell}{i-1} 2^{i-1}\epsilon \leq 3^\ell \epsilon. \qquad \square$$

Proof of Theorem 3. Using n independent samples the first $2k-1$ moments of U can be estimated within $\sqrt{O_k(\log(1/\delta))/n}$ with probability $1-\delta$. The conclusion follows from Lemma 14 and 15. $\square$

8.4 Proofs for Section 4.1

Proof of Lemma 4. Denote by $\tilde{m}'_r$ the average of $\gamma_r(X_i, \sigma)$ over n' samples. By Chebyshev inequality $|\tilde{m}'_r - m_r(U)| < O(\sqrt{\mathsf{var}[\gamma_r(X,\sigma)]/n'})$ with probability at least $3/4$. Let $\tilde{m}$ be the median over $O(\log \frac{1}{\delta})$ independent estimates, and then, by Hoeffding's inequality, $|\tilde{m}_r - m_r(U)| < O(\sqrt{\mathsf{var}[\gamma_r(X,\sigma)]/n'})$ with probability at least $1-\delta$. Therefore it boil down to an upper bound of $\mathsf{var}[\gamma_r(X,\sigma)]$. Recall the formula of $\gamma_r(x,\sigma)$ in (15). Since the standard deviation of a summation is at most the sum of individual standard deviations, $\square$

8.5 Proofs for Section 4.2

Proof of Lemma 5. The proof is similar to [Lin89, Theorem 5B]. Let $\hat{\mathbf{M}}_r(\sigma)$ denote the moment matrix associated with the empirical moments of $\gamma_i(X,\sigma)$ for $i \leq 2r$, and let $\hat{\sigma}_r$ denote the smallest non-negative root of $\det(\mathbf{M}_r(\sigma)) = 0$. Then $\hat{\sigma} = \hat{\sigma}_k$, and we can directly compute that $\hat{\sigma}_1 = s$. Since X is continuous, then almost surely, the empirical distribution has n points of support. By Theorem 4, the matrix $\hat{\mathbf{M}}_r(0)$ is positive definite and thus $\hat{\sigma}_r > 0$ for any $r < n$. For any $q < r$, if $\hat{\mathbf{M}}_r(\sigma)$ is positive definite, then $\hat{\mathbf{M}}_q(\sigma)$ as a leading principal submatrix is also positive definite. Since eigenvalues of $\hat{\mathbf{M}}_r(\sigma)$ are continuous functions of σ, then $\hat{\sigma}_r \leq \hat{\sigma}_q$. In particular, $\hat{\sigma}_k \leq \hat{\sigma}_1$. $\square$

Proof of Lemma 6. The proof of Lemma 5 concludes that $0 < \hat{\sigma} = \hat{\sigma}_k \leq \hat{\sigma}_{k-1} \leq \ldots \leq \hat{\sigma}_1 = s$, and for any $\sigma < \hat{\sigma}_j$, the matrix $\hat{\mathbf{M}}_j(\sigma)$ is positive definite. Since $\det(\hat{\mathbf{M}}_k(\hat{\sigma})) = 0$ by the definition of $\hat{\sigma}_k$, then, for some $r \in \{1, \ldots, k\}$, we have $\det(\hat{\mathbf{M}}_j(\hat{\sigma})) = 0$ for $j = r, \ldots, k$, and $\det(\hat{\mathbf{M}}_j(\hat{\sigma})) > 0$ for $j = 0, \ldots, r-1$. By Theorem 4, there is a r-atomic distribution whose j^{th} moment coincides with $\hat{\gamma}_j(\hat{\sigma})$ for $j \leq 2r$. It suffices to show that $r = k$ almost surely.

Since X is continuous, in the following we condition on the event that all samples $X_1, \ldots, X_n$ are distinct, which happens almost surely, without loss of generality. We first show that the empirical moments $(\hat{m}_1, \ldots, \hat{m}_n)$, where $\hat{m}_j = \frac{1}{n}\sum_i X_i^j$, have a joint density in $\mathbb{R}^n$. The Jabobian matrix of this transformation is

$$\frac{1}{n}\begin{bmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & n \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \\ \vdots & \ddots & \vdots \\ X_1^{n-1} & \cdots & X_n^{n-1} \end{bmatrix},$$

which is non-degenerate. Since those n samples $(X_1, \dots, X_n)$ have a joint density, then the empirical moments $(\hat{m}_1, \dots, \hat{m}_n)$ also have a joint density.

Suppose, for the sake of contradiction, that $r \le k-1$. Then $\det(\hat{\mathbf{M}}_{r-1}(\hat{\sigma})) > 0$ and $\det(\hat{\mathbf{M}}_r(\hat{\sigma})) = \det(\hat{\mathbf{M}}_{r+1}(\hat{\sigma})) = 0$. It follows from Lemma 26 that $\hat{\gamma}_{2r+1}(\hat{\sigma})$ is a deterministic function of $\hat{\gamma}_1(\hat{\sigma}), \dots, \hat{\gamma}_{2r}(\hat{\sigma})$. Since $\hat{\sigma}$ is also first non-negative root of $\hat{d}_r(\sigma) = 0$, it is uniquely determined by $(\hat{m}_1, \dots, \hat{m}_{2r})$. Therefore, $\hat{\gamma}_{2r+1}(\hat{\sigma})$, and thus $\hat{m}_{2r+1}$, are both deterministic functions of $(\hat{m}_1, \dots, \hat{m}_{2r})$, which happens with probability zero, since the sequence $(\hat{m}_1, \dots, \hat{m}_{2r+1})$ has a joint density. Consequently, $r \le k-1$ happens with probability zero. □

Proof of Lemma 7. We only show the upper tail bound $\mathbb{P}[Y \ge t]$. The lower tail bound of Y is equivalent to the upper tail bound of $-Y$. Suppose $\max_{i\in[2k]} |m_i(X) - m_i(Y)| = \delta$.

Suppose X takes at most k values, denoted by $x_1, x_2, \dots$. Let a polynomial of degree $2k$ with bounded coefficients be $P(x) = Q^2(x)$, where $Q(x) = \prod_i (x - x_i)^{d_i}$, and $d_i \ge 1$ and sum up to k. Note that $P(X) = 0$ almost surely. Then $\mathbb{E}[P(Y)] = \mathbb{E}[P(Y)] - \mathbb{E}[P(X)] \le 2^{2k}\delta$. By Markov inequality, for any $t \ge 2$,

$$\mathbb{P}[Y \ge t] \le \mathbb{P}[P(Y) \ge P(t)] \le \frac{\mathbb{E}[P(Y)]}{P(t)} \le \frac{\delta}{(\Omega(t))^{2k}}.$$

Suppose Y takes at most k values. If all those values are within $[-1, 1]$, then we are done. If at most $k-1$ values, denoted by $x_1, x_2, \dots$, are within $[-1, 1]$, let a polynomial of degree $2k$ with bounded coefficients be $P(x) = (x^2 - 1)Q^2(x)$, where $Q(x) = \prod_i (x - x_i)^{d_i}$, and $d_i \ge 1$ and sum up to $k-1$. Note that $\mathbb{E}[P(X)] \le 0$. Then $\mathbb{E}[P(Y)] \le \mathbb{E}[P(Y)] - \mathbb{E}[P(X)] \le 2^{2k}\delta$. Since $P(Y) \ge 0$ almost surely, the conclusion follows analogously by Markov inequality. □

Proof of Corollary 1. Suppose $\max_{i\in[2k]} |m_i(X) - m_i(\hat{X})| = \delta$. Let $\tau \triangleq \sqrt{|\sigma^2 - \hat{\sigma}^2|}$, and G be a random variable whose first $2k$ moments equal the corresponding moments of the standard normal distribution, with magnitude at most $O(\sqrt{k})$, for example, the quadrature of Z [Sze75]. Suppose $\sigma \ge \hat{\sigma}$. Then, by Lemma 18, $|m_r(U+\tau G) - m_r(\hat{U})| \le (O(\sqrt{r}))^r \delta$ for $r \le 2k$. Applying Lemma 7 with $\frac{U+\tau G}{O(\sqrt{k})}$ and $\frac{\hat{U}}{O(\sqrt{k})}$ yields the conclusion. Suppose $\sigma \le \hat{\sigma}$. Then $|m_r(U) - m_r(\hat{U} + \tau G)| \le (O(\sqrt{r}))^r \delta$ for $r \le 2k$. Since $\mathbb{P}[|\hat{U}| \ge t] \le \mathbb{P}[|\hat{U} + \tau G| \ge t - O(\sqrt{k})\tau]$, the conclusion follows again by Lemma 7. □

Proof of Theorem 5. Note that the r^{th} moment of $\hat{X} = \hat{U} + \hat{\sigma}Z$ equals the r^{th} empirical moment from n samples, for all $r \le 2k$. Since U and σ are both bounded, then by Lemma 27 and the union bound, with probability at least $1 - \delta$, we have $|m_r(X) - m_r(\hat{X})| \le \sqrt{\frac{(O(r))^r \log(k/\delta)}{n}}$ for all $r \le 2k$. By Lemma 5 the estimate $\hat{\sigma}$ is also bounded. Let $\epsilon \triangleq \sqrt{\frac{(O(k))^{2k} \log(k/\delta)}{n}}$ be an uniform accuracy of the estimates of first $2k$ moments. Define a random variable $\tilde{U} = \hat{U}\mathbf{1}_{\{|\hat{U}| \le O(\sqrt{k})\}}$, which is bounded by $O(\sqrt{k})$ and also takes at most k values. By the tail bound of the k-atomic random variable $\hat{U}$ in Corollary 1, $|m_j(\hat{U}) - m_j(\tilde{U})| \le (O(k))^k \epsilon$ for $j \le 2k$, and the transportation cost is $W_1(\hat{U}, \tilde{U}) \le (O(k))^k \epsilon$. Define $\tilde{X} = \tilde{U} + \hat{\sigma}Z$. Then, $|m_j(\hat{X}) - m_j(\tilde{X})| \le (O(k))^{2k}\epsilon$ and thus $|m_j(X) - m_j(\tilde{X})| \le (O(k))^{2k}\epsilon$, for $j \le 2k$. Applying Proposition 3 yields that $|\sigma^2 - \hat{\sigma}^2| \le O_k((n/\log\frac{1}{\delta})^{-\frac{1}{2k}})$ and $W_1(U, \tilde{U}) \le O_k((n/\log\frac{1}{\delta})^{-\frac{1}{4k}})$. The conclusion follows. □

8.6 Proofs for Section 4.4

Proof of Lemma 8. Let $M = \Omega(\sqrt{\log n})$ and $M' = \sqrt{2\log(1/\epsilon')}$ and then $L = M + M'$. Since the mixing weights are at least ϵ, then with probability at least $1 - ke^{-n'(\epsilon-\epsilon')}$, there exists at

least one sample $\tilde{X}_i \in [\mu_j \pm M']$ for each of the k components, so the interval $[\tilde{X}_i \pm L]$ covers the interval $[\mu_j \pm M]$. Note that, with probability at least $1 - n^{-\Omega(1)}$, each of the n samples to cluster is close to its latent center μ_j within $O(\sqrt{\log n})$, and thus belongs to the cluster that covers the interval $[\mu_j \pm M]$. Moreover, with probability at least $1 - n'^{-\Omega(1)}$, the Gaussian noises within the n' test samples are in magnitude at most $O(\sqrt{\log n'})$, and thus each interval I_r is within $[\mu_j \pm (L + O(\sqrt{\log n'}))]$ from some μ_j. Since there are at most k components, each cluster is within an interval of length $O(k(L + \sqrt{\log n'}))$. □

Proof of Theorem 7. We apply Lemma 8 with $n \geq n' \geq \Omega(\frac{\log k}{\epsilon})$, and $\epsilon' = \epsilon/2$. Let $M = \Theta(\sqrt{\log n + k \log\log n})$, and then $L = O(M)$ since $\frac{1}{\epsilon'} \leq O(\frac{n}{\log k})$. Then it follows the conclusion of Lemma 8 that, with probability at least $1 - n^{-\Omega(1)} - n'^{-\Omega(1)} - k^{-\Omega(1)}$, a sample in an interval is equivalent to its latent center also in the same interval, and the length of each interval is at most $O(kM)$. The intervals are independent of n samples for estimating centers, so they are treated as deterministic hereafter. Denote each interval by $I_j = [c_j \pm L_j]$. To apply the estimation guarantee of our denoised method of moments on a bounded interval, we need to show that the empirical moments from the subset of shifted samples $\mathcal{S}'_j = \{X_i - c_j : X_i \in I_j\}$ are close to the corresponding moments of the shifted Gaussian mixture model $U_j - c_j + \sigma Z$, where U_j consists of all components whose centers are in I_j and is distributed according to the conditional distribution of U given $U \in I_j$. To this end, we first compute the expected moments, conditioned on all Gaussian noises are no greater than M:

$$\mathbb{E}[(X - c_j)^r | X \in I_j, |Z| \leq M] = \mathbb{E}[(X - c_j)^r | U \in I_j, |Z| \leq M] = \mathbb{E}[(U'_j + \sigma Z)^r | |Z| \leq M],$$

where $U'_j = U_j - c_j$. Since $|U'_j| \leq L_j$, applying Lemma 16 yields that

$$|\mathbb{E}[(U'_j + \sigma Z)^r | |Z| \leq M] - \mathbb{E}[(U'_j + \sigma Z)^r]| \leq (L_j + O(M\sqrt{r}))^r \frac{r}{M} e^{-\frac{M^2}{2}}.$$

Note that $L_j \leq O(kM)$, so for $r \leq 2k$ the above difference is at most $\frac{(O(k))^{2k+1}}{n}$, and then the accuracy of empirical moments using $\mathcal{S}'_j$ is the same as the accuracy using i.i.d. samples from $U_j - c_j + \sigma Z$. Since each component has at least ϵ probability and $n \geq \Omega(\frac{\log k}{\epsilon})$, then with probability $1 - k^{\Omega(1)}$, each subset of samples $\mathcal{S}'_j$ is of size at least $\Omega(n\epsilon)$. The accuracy guarantees for the denoised method of moments in Theorem 1 in the bounded case yield that $W_1(\hat{U}'_j, U'_j) \leq O_k(L_j(n\epsilon)^{-\frac{1}{4k-2}}) \leq O_k(L(n\epsilon)^{-\frac{1}{4k-2}})$ when σ is known and is at most $O_k(L(n\epsilon)^{-\frac{1}{4k}})$ when σ is unknown. Therefore, after applying the mixing weights threshold $\tau = \epsilon/(2k)$, the estimated centers are close to the support of U_j in Hausdorff distance within $O(\frac{W_1(U_j, \hat{U}_j)}{\epsilon/(2k)})$ by Lemma 17. Consequently the same accuracy applies to the Hausdorff distance between unions of centers over all groups.

Lemma 16. *For $M \geq 1$,*

$$0 \leq \mathbb{E}[Z^r] - \mathbb{E}[Z^r | |Z| \leq M] \leq r(O(\sqrt{r}))^r \left(M^{r-1} e^{-\frac{M^2}{2}} \right).$$

Proof. For r odd, we have $\mathbb{E}[Z^r] - \mathbb{E}[Z^r | |Z| \leq M] = 0$. For r even, the left inequality is immediate since $x \mapsto x^r$ is increasing. For the right inequality,

$$\mathbb{E}[Z^r] - \mathbb{E}[Z^r | |Z| \leq M] = \mathbb{E}[Z^r] - \frac{\mathbb{E}[Z^r \mathbf{1}_{\{|Z| \leq M\}}]}{\mathbb{P}[|Z| \leq M]} \leq \frac{\mathbb{E}[Z^r] - \mathbb{E}[Z^r \mathbf{1}_{\{|Z| \leq M\}}]}{\mathbb{P}[|Z| \leq M]},$$

and the conclusion follows from Claim 2. □

Lemma 17. *Let μ be a discrete distribution whose atom has at least ϵ probability and is supported on S_μ. For any subset S of the support of another distribution ν,*

$$d_H(S_\mu, S) \le \frac{W_1(\mu,\nu)}{(\min_{y\in S}\nu\{y\}) \wedge (\epsilon - \nu(S^c))_+}.$$

Proof. This is a generalization of Lemma 2. Let X and Y be distributed according to any coupling of μ and ν. For any $y \in S$,

$$\mathbb{E}|X-Y| \ge \nu\{y\}\mathbb{E}[|X-Y||Y=y] \ge \min_{t\in S}\nu\{t\}d(y,\mathrm{supp}(\mu)).$$

For any $x \in \mathrm{supp}(\mu)$, we have $\mu\{x\} \ge \epsilon$. Since $\mathbb{P}[Y \notin S|X=x]\mathbb{P}[X=x] \le \nu(S^c)$, then $\mathbb{P}[Y \notin S|X=x] \le \nu(S^c)/\epsilon$, and thus $\mathbb{P}[Y \in S, X=x] \ge (\epsilon - \nu(S^c))_+$. Therefore,

$$\mathbb{E}|X-Y| \ge (\epsilon-\nu(S^c))_+\mathbb{E}[|X-Y||X=x, Y\in S] \ge (\epsilon-\nu(S^c))_+ d(x,S).$$

The proof is completed. □

□

8.7 Proofs for Section 5

Proof of Proposition 6. Let U and U' be two k-atomic centered random variables taking values in $[-1,1]$ with identical first $2k-2$ moments but Wasserstein distance at least $\Omega(1/k)$, whose existence is given by Lemma 21. Directly applying Lemma 9 yields that $\chi^2(\epsilon U + Z \| \epsilon U' + Z) \le (O(\epsilon))^{4k-2}$. This can improved to $(O(\frac{\epsilon}{\sqrt{k}}))^{4k-2}$ by a similar argument as Lemma 9 using $\mathbb{E}[|\epsilon U|^p], \mathbb{E}[|\epsilon U'|^p] \le (O(\epsilon))^p$ for all $p \ge 2k-1$ by boundedness of U and U'. Consequently, in distinguishing the two mixture models $\epsilon U + Z$ and $\epsilon U' + Z$, any test using $(O(\frac{\sqrt{k}}{\epsilon}))^{4k-2}$ samples makes an error with constant probability. Then any estimate $\hat{U}$ using $(O(\frac{\sqrt{k}}{\epsilon}))^{4k-2}$ samples makes an error of at least $\Omega(\frac{\epsilon}{k})$ in W_1 distance with constant probability under either mixture model $\epsilon U + Z$ or $\epsilon U' + Z$. Choosing $\epsilon = \Theta(\sqrt{k} n^{-\frac{1}{4k-2}})$ completes the proof. □

Proof of Proposition 7. Let U' be standard normal and U be the k-point quadrature under standard normal distribution, and then they have identical first $2k-1$ moments. Since U and U' are both 1-subgaussian (as explained in Remark 3), and $\epsilon U' + Z$ has the same distribution as $\sqrt{1+\epsilon^2}Z$, it follows from Lemma 9 that $\chi^2(\epsilon U + Z \| \sqrt{1+\epsilon^2}Z) \le (O(\epsilon))^{4k}$. Consequently, in distinguishing the two mixture models $\epsilon U + Z$ and $\sqrt{1+\epsilon^2}Z$, any test using $(O(\frac{1}{\epsilon}))^{4k}$ samples makes an error with constant probability. Note that the discrete latent distributions differ in W_1 distance by $\Omega(\frac{\epsilon}{\sqrt{k}})$ (see a proof in Lemma 24). Then any estimate $(\hat{U}, \hat{\sigma}^2)$ using $(O(\frac{1}{\epsilon}))^{4k}$ samples makes an error of at least $\Omega(\frac{\epsilon}{\sqrt{k}})$ in W_1 distance on $\hat{U}$ part or at least $\Omega(\epsilon^2)$ on $\hat{\sigma}^2$ part with constant probability, under either mixture model $\epsilon U + Z$ or $\sqrt{1+\epsilon^2}Z$. Choosing $\epsilon = \Theta(n^{-\frac{1}{4k}})$ completes the proof. □

A Standard form of the semidefinite programming (16)

Given a sequence $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_r)$ that is not necessarily a valid moment vector, we want to project it onto a legitimate moment vector by (16). By introducing an auxiliary variable $t \ge x^\top x$, the projection problem (16) is equivalent to

$$\begin{aligned} \min\ & t - 2\tilde{m}^\top x + \tilde{m}^\top \tilde{m}, \\ \text{s.t. } & t \ge x^\top x,\ x \text{ satisfies (9)}. \end{aligned}$$

This is a semidefinite programming since the constraint $t \geq x^\top x$ is equivalent to $\begin{bmatrix} t & x^\top \\ x & I \end{bmatrix} \succeq 0$ using Schur complement (see, e.g., [VB96]).

B Quadrature algorithm

Here we provide a basic algorithm to the find the quadrature rule. More efficient and stable algorithms are summarized in Section 2.1.

Algorithm 5 Quadrature rule

Input: a sequence of $2k-1$ moments $(m_1, \dots, m_{2k-1})$.

Output: nodes $x = (x_1, \dots, x_k)$ and weights $w = (w_1, \dots, w_k)$.

Let a polynomial P be

$$P(x) = \det \begin{bmatrix} 1 & m_1 & \cdots & m_k \\ \vdots & \vdots & \ddots & \vdots \\ m_{k-1} & m_k & \cdots & m_{2k-1} \\ 1 & x & \cdots & x^k \end{bmatrix}.$$

Let $(x_1, \dots, x_k)$ be the roots of the polynomial P.

Let w be

$$w = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_k^{k-1} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix}.$$

C Auxiliary lemmas

Lemma 18. *Let Z be a standard normal random variable independent of U and U', and $\sigma \in [0,1]$. Then,*

$$|m_r(U) - m_r(U')| \leq (O(\sqrt{r}))^r \max_{i \in [r]} |m_i(U + \sigma Z) - m_i(U' + \sigma Z)|.$$

Proof. Let $X = U + \sigma Z$ and $X' = U' + \sigma Z$, and $\max_{i \in [r]} |m_i(X) - m_i(X')| = \delta$. Note that $m_r(U) = \mathbb{E}[\gamma_r(X, \sigma)]$ and $m_r(U') = \mathbb{E}[\gamma_r(X', \sigma)]$, where γ_r is defined in (15). Then, applying the formula of Hermite polynomials,

$$|m_r(U) - m_r(U')| \leq \sum_{i=0}^{\lfloor r/2 \rfloor} \frac{r! \sigma^{2i}}{i!(r-2i)! 2^i} \delta = \delta \cdot \mathbb{E}(1 + \sigma Z)^r \leq \delta \cdot (O(\sqrt{r}))^r,$$

since $\mathbb{E}|Z|^r$ scales as $(O(\sqrt{r}))^r$. □

Lemma 19. *Let $r \geq 2$. Then,*

$$\int \left(\frac{\delta}{\prod_{i=1}^r |t - x_i|} \wedge 1 \right) \mathrm{d}t \leq 4r\delta^{\frac{1}{r}}.$$

Proof. Without loss of generality, let $x_1 \le x_2 \le \cdots \le x_r$. Note that

$$\int \left(\frac{\delta}{\prod_{i=1}^r |t - x_i|} \wedge 1 \right) \mathrm{d}t = \int_{-\infty}^{x_1} + \int_{x_1}^{\frac{x_1+x_2}{2}} + \int_{\frac{x_1+x_2}{2}}^{x_2} + \cdots + \int_{x_r}^{\infty}.$$

There are $2r$ terms in the summation and each term can be upper bounded by

$$\int_{x_i}^{\infty} \left(\frac{\delta}{|t - x_i|^r} \wedge 1 \right) \mathrm{d}t = \int_0^{\infty} \left(\frac{\delta}{t^r} \wedge 1 \right) \mathrm{d}t = \frac{r}{r-1} \delta^{\frac{1}{r}}.$$

The conclusion follows. □

Lemma 20. *Given any $2k$ distinct points $x_1 < x_2 < \cdots < x_{2k}$, there are two distributions that match the first $2k-2$ moments, with one distribution supported on $x_1, x_3, \ldots, x_{2k-1}$ and the other one supported on $x_2, x_4, \ldots, x_{2k}$.*

Proof. Consider the following linear equation

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{2k-2} & x_2^{2k-2} & \cdots & x_{2k}^{2k-2} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{2k} \end{pmatrix} = 0,$$

whose solutions are of the form $C \cdot w$, where we let the ℓ_1-norm of w be two. Since all weights sum up to zero, then positive weights in w sum up to 1 and negative weights sum up to -1. Let one distribution be support on x_i with weight w_i for w_i being positive, and the other one be supported on the remaining x_i's with the corresponding weights $|w_i|$. Then those two distributions match the first $2k-2$ moments.

It remains to show that weights in one non-zero solution have alternating signs. Note that all weights are non-zero: if one w_i is zero, then the solution must be all zero since the Vandermonde matrix is of full rank. One solution can be obtained by $w_{2k} = -1$ and

$$\begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_{2k-1} \\ \vdots & \ddots & \vdots \\ x_1^{2k-2} & \cdots & x_{2k-1}^{2k-2} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{2k-1} \end{pmatrix} = \begin{pmatrix} 1 \\ x_{2k} \\ \vdots \\ x_{2k}^{2k-2} \end{pmatrix}.$$

We have an explicit formula for $w_i = P_i(x_{2k})$ where P_i is an interpolating polynomial of degree $2k-2$ satisfying $P_i(x_j)$ being one for $j = i$ and being zero for all other $j \le 2k-1$. Specifically, $w_i = \frac{\prod_{j \ne i, j \le 2k-1}(x_{2k} - x_j)}{\prod_{j \ne i, j \le 2k-1}(x_i - x_j)}$. The proof is complete. □

Lemma 21.

$$\sup\{W_1(U, U') : m_i(U) = m_i(U') \ \forall\ i = 1, \ldots, \ell, |U|, |U'| \le 1\} = \Theta(1/(\ell+1)).$$

Furthermore, the supremum is $\frac{\pi - o(1)}{\ell+1}$ as $\ell \to \infty$, and is achieved by two distributions whose sizes of support differ by at most one and sum up to $\ell + 2$.

Proof. By the dual representation of W_1 distance in Section 2.3, the supremum is equal to

$$\sup_{f:1-\text{Lipschitz}} \sup\left\{\mathbb{E}[f(U)] - \mathbb{E}[f(U')] : m_i(U) = m_i(U') \ \forall\ i = 1, \dots, \ell, |U|, |U'| \le 1\right\}.$$

Since moment matching is the dual problem of the best polynomial approximation (see a proof in [WY16, Appendix E]), the optimal value is further equal to

$$\sup_{f:1-\text{Lipschitz}} \inf\left\{2 \sup_{|x|\le 1} |f(x) - P(x)| : P \text{ is a polynomial of degree } \ell\right\}.$$

The optimal value follows from the best uniform approximation error over 1-Lipschitz functions, a well-studied quantity in approximation theory (see, e.g., [Bus11]). The optimal distributions of the original problem can also be found from the dual problem: they are supported on the maximum points and the mimimum points of $P^* - f^*$, respectively, where f^* is the optimal 1-Lipschitz function and P^* is the optimal polynomial. The numbers of maximum points and minimum points follows from the Chebyshev alternating theorem (see, e.g., [Tim63]). □

Lemma 22. *Let G_n be the n-point quadrature under standard normal distribution such that $\mathbb{E}[G_n^j] = \mathbb{E}[Z^j]$ for $j = 1, \dots, 2n-1$, where Z is standard normal. For $j \ge 2n$, we have $\mathbb{E}[G_n^j] \le \mathbb{E}[Z^j]$ when j is even, and $\mathbb{E}[G_n^j] = \mathbb{E}[Z^j] = 0$ otherwise.*

Proof. When j is odd, by symmetry $\mathbb{E}[G_n^j] = \mathbb{E}[Z^j] = 0$. For $j \ge 2n$ being even, the conclusion immediately follows from the remainder (error) of quadrature rule, an extensively studied topic in numerical analysis [DR07]:

$$\int f(x)\mathrm{d}\nu(x) - \sum_{i=1}^n w_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n)!} \int \pi_n^2(x)\mathrm{d}\nu(x), \tag{29}$$

where w_i and x_i are weights and nodes of the n-point quadrature rule in (10) under ν, the point ξ depending on x is in the domain of ν, and $\pi_n(x) = \prod_i (x - x_i)$. The remainder of the quadrature rule can be analyzed from the perspective of Hermite polynomial interpolation, a key method in this paper, and is derived below for completeness. The evaluation of quadrature rule $\sum_{i=1}^n w_i f(x_i)$ as is equal to the integral of a polynomial P of degree $2n-1$, obtained from the Hermite interpolation:

$$P(x_i) = f(x_i), \quad P'(x_i) = f'(x_i), \qquad \forall\ i = 1, \dots, n.$$

Applying the error of Hermite interpolation [SB02] yields that

$$\int f(x)\mathrm{d}\nu(x) - \sum_{i=1}^n w_i f(x_i) = \int (f(x) - P(x))\mathrm{d}\nu(x) = \int \frac{f^{(2n)}(\xi(x))}{(2n)!}\pi_n^2(x)\mathrm{d}\nu(x),$$

and the remainder (29) follows immediately by the mean value theorem. □

Lemma 23. *Let G_n be the n-point quadrature under standard normal distribution such that $\mathbb{E}[G_n^j] = \mathbb{E}[Z^j]$ for $j = 1, \dots, 2n-1$, where Z is standard normal. Then $\mathbb{E}[H_j(G_n)] = 0$ for $j = 1, \dots, 2n-1$, and $\mathbb{E}[H_{2n}(G_n)] = -n!$.*

Proof. Expand $H_n^2(x)$, a polynomial of degree $2n$, using Hermite polynomials $H_n^2(x) = a_{2n}H_{2n}(x) + \cdots + a_0$, with $a_{2n} = 1$ according to the leading coefficient. By orthogonality $\mathbb{E}[H_j(Z)] = 0$ for all $j \ge 1$ and thus $\mathbb{E}[H_j(G_n)] = 0$ for $j = 1, \dots, 2n-1$. Then, $n! = \mathbb{E}[H_n^2(Z)] = a_0$. Recall from Section 2.1 that the quadrature G_n is supported on the roots of the n^{th} Hermite polynomial $H_n(x)$. Then, $0 = \mathbb{E}[H_n^2(G_n)] = \mathbb{E}[H_{2n}(G_n)] + a_0$, and the proof is completed. □

Lemma 24. *Let G_n be the n-point quadrature under standard normal distribution. Then $\mathbb{E}|G_n| \geq 1/\sqrt{4n+2}$ for $n \geq 2$.*

Proof. Recall that G_n is supported on zeros of the Hermite polynomial of degree n, which is at most $\sqrt{4n+2}$ in magnitude [Sze75]. The conclusion follows since $1 = \mathbb{E}[G_n^2] \leq \mathbb{E}|G_n|\sqrt{4n+2}$. □

Lemma 25 (Non-existence of unbiased estimators)**.** *Using finite samples $X_1, \ldots, X_m$ i.i.d. drawn from a two-component Gaussian mixture model $pN(s, \sigma^2)+(1-p)N(t, \sigma^2)$, with unknown parameters p, s, t, σ. For any $r \geq 2$, there is no unbiased estimator for the r^{th} moments of its latent distribution $ps^r + (1-p)t^r$.*

Proof. We first derive a few necessary conditions for an unbiased estimate, denoted by $g(x_1, \ldots, x_m)$, and then arrive at a contradiction. Expand the function under Hermite polynomials

$$g(x_1, \ldots, x_m) = \sum_{n_1, \ldots, n_m \geq 0} \alpha_{n_1, \ldots, n_m} \prod_i H_{n_i}(x_i),$$

and denote by $T_n(\mu, \sigma^2)$ the expected value of the Hermite polynomial $\mathbb{E}H_n(X)$ under Gaussian model $X \sim N(\mu, \sigma^2)$. Without loss of generality we may assume that the function g and the coefficients α are symmetric. Then, the expected value of the function g under $\sigma^2 = 1$ is

$$\mathbb{E}[g(X_1, \ldots, X_m)] = \sum_{n_1, \ldots, n_m \geq 0} \alpha_{n_1, \ldots, n_m} \prod_i (ps^{n_i} + (1-p)t^{n_i}), \tag{30}$$

which can be viewed as a polynomial in p, whereas the target is $ps^r + (1-p)t^r$, a linear function in p. Matching polynomial coefficients yields that

$$\sum_{n_1 + \cdots + n_m \geq 0} \alpha_{n_1, \ldots, n_m} t^{n_1 + \cdots + n_m} = t^r, \tag{31}$$

$$\sum_{n_1 + \cdots + n_m \geq 0} \alpha_{n_1, \ldots, n_m} (s^{n_1} - t^{n_1}) t^{n_2 + \cdots + n_m} \cdot m = s^r - t^r, , \tag{32}$$

$$\sum_{n_1 + \cdots + n_m \geq 0} \alpha_{n_1, \ldots, n_m} \prod_{i=1}^{j} (s^{n_i} - t^{n_i}) t^{n_{j+1} + \cdots + n_m} = 0, \ \forall\ j = 2, \ldots, m, \tag{33}$$

where we used the symmetry of the coefficients α. The equality (33) with $j = m$ yields that $\alpha_{n_1, \ldots, n_m} \neq 0$ only if at least one n_i is zero; then (33) with $j = m-1$ yields that $\alpha_{n_1, \ldots, n_m} \neq 0$ only if at least two n_i are zero; repeating this for $j = m, m-1, \ldots, 2$, we obtain that $\alpha_{n_1, \ldots, n_m}$ is nonzero only if at most one n_i is nonzero. Then the equality (32) implies that $\alpha_{n_1, \ldots, n_m}$ is nonzero only if exactly one $n_i = r$ and the coefficient is necessarily $\frac{1}{m}$. Therefore, it is necessary that the symmetric function is $g(x_1, \ldots, x_m) = \frac{1}{m} \sum_{i=1}^m H_r(x_i)$. However, this function is biased when $\sigma^2 \neq 1$. □

Lemma 26. *Given a sequence $\gamma_1, \gamma_2, \ldots$, let $\mathbf{H}_j$ denote the Hankel matrix of size $j+1$ using $1, \gamma_1, \ldots, \gamma_{2j}$. Suppose $\det(\mathbf{H}_{r-1}) \neq 0$, and $\det(\mathbf{H}_r) = \det(\mathbf{H}_{r+1}) = 0$. Then,*

$$\gamma_{2r+1} = (\gamma_{r+1}, \ldots, \gamma_{2r})(\mathbf{H}_{r-1})^{-1}(\gamma_r, \ldots, \gamma_{2r-1})^\top.$$

Proof. The matrices $\mathbf{H}_{r-1}$ and $\mathbf{H}_r$ are both of rank r by their determinants. We first show that the rank of $[\mathbf{H}_r, v]$, which is the first $r+1$ rows of $\mathbf{H}_{r+1}$ and is of dimension $(r+1) \times (r+2)$, is also r, where $v \triangleq (\gamma_{r+1}, \ldots, \gamma_{2r+1})^\top$. Suppose the rank is $r+1$. Then v cannot be in the image of $\mathbf{H}_r$. By symmetry of the Hankel matrix, the transpose of $[\mathbf{H}_r, v]$ is the first $r+1$ columns of $\mathbf{H}_{r+1}$.

Those $r+1$ columns are linearly independent when its rank is $r+1$. Since $\det(\mathbf{H}_{r+1}) = 0$, then the last column of $\mathbf{H}_{r+1}$ must be in the image of the first $r+1$ columns, which is a contradiction.

Since first r columns of $\mathbf{H}_{r+1}$ are linearly independent, and the first $r+1$ columns of $\mathbf{H}_{r+1}$ are of rank r. Then the $(r+1)^{\text{th}}$ column of $\mathbf{H}_{r+1}$ is in the image of the first r columns, and thus γ_{2r+1} is a linear combination of $\gamma_{r+1}, \ldots, \gamma_{2r}$. Since $\mathbf{H}_{r-1}$ is of full rank, the coefficients can be uniquely determined by $(\mathbf{H}_{r-1})^{-1}(\gamma_r, \ldots, \gamma_{2r-1})^\top$. □

Claim 1.

$$\mathbb{P}[Z > M] \le e^{-\frac{M^2}{2}}.$$

Proof. Applying Chernoff bound yields that

$$\mathbb{P}[Z > M] \le \exp(-\sup_t(tM - t^2/2)) = \exp(-M^2/2).$$ □

Claim 2. *For r even, and $M \ge 1$,*

$$\mathbb{E}[Z^r \mathbf{1}_{\{|Z|>M\}}] \le r(O(\sqrt{r}))^r \left(M^{r-1} e^{-\frac{M^2}{2}}\right).$$

Proof. Applying integral by parts yields that

$$\int_M^\infty x^r e^{-\frac{x^2}{2}} \mathrm{d}x = M^{r-1}e^{-\frac{M^2}{2}} + (r-1)M^{r-3}e^{-\frac{M^2}{2}} + (r-1)(r-3)M^{r-5}e^{-\frac{M^2}{2}} + \cdots + (r-1)!! \int_M^\infty e^{-\frac{x^2}{2}} \mathrm{d}x.$$

Applying Claim 1 and $(r-1)!! \le (O(\sqrt{r}))^r$, the conclusion follows. □

Lemma 27 (Accuracy of moment estimate)**.** *Given n samples from a Gaussian mixture model $X = U + \sigma Z$, where U and σ are both bounded, the r^{th} moments of X can be estimated within accuracy $\sqrt{\frac{(O(r))^r \log(1/\delta)}{n}}$ with probability at least $1-\delta$.*

Proof. Split n samples into $O(\log \frac{1}{\delta})$ groups of size $O(n/\log \frac{1}{\delta})$, and consider the median of r^{th} empirical moments. By Chebyshev inequality, in each group, with probability at least $3/4$, the r^{th} empirical moment is close to the population moments within $\sqrt{(O(r))^r \log(1/\delta)/n}$. By Hoeffding's inequality, the median has the same accuracy with probability at least $1-\delta$. □

Lemma 28 (Distribution of random projection)**.** *Given a unit vector a and random direction X,*

$$\mathbb{P}[|\langle a, X\rangle| < r] < r\sqrt{d}.$$

Proof. Let the area of the surface of d-dimensional unit ball be $S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$. By symmetry,

$$\mathbb{P}[|\langle a, X\rangle| < r] = \mathbb{P}[|X_1| < r] = \frac{\int_{-r}^r (\sqrt{1-x^2})^{d-2} S_{d-2}\sqrt{1-x^2}\mathrm{d}x}{S_{d-1}} = \frac{2S_{d-2}}{S_{d-1}} \int_0^r (1-x^2)^{\frac{d-3}{2}} \mathrm{d}x < r\sqrt{d},$$

where X_1 is the first coordinate of X. □

Lemma 29 (Accuracy of spectral methods)**.** *Given n samples from the Gaussian mixture model $X = U + Z$, where U takes values in $\{-\theta, \theta\}$ uniformly, $Z \sim N(0, I_d)$, and $\theta \in \mathbb{R}^d$ is bounded, denote the sample covariance matrix by $S = \frac{1}{n}\sum_i X_i X_i^\top$. Let $\hat{s}^2$ be the largest eigenvalue of $S - I_d$ if it is positive and zero otherwise, where $\hat{s}$ is non-negative, and $\hat{v}$ the corresponding normalized eigenvector, where we decree that $\theta^\top \hat{v} \ge 0$. Then, with $\hat{\theta} = \hat{s}\hat{v}$, when $n > d$, with high probability,*

$$\|\theta - \hat{\theta}\|_2 \le O(d/n)^{1/4}.$$

Proof. The given n samples can be represented in a matrix form $X = \theta\varepsilon^\top + Z \in \mathbb{R}^{d\times n}$, where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Rademacher random variables, and Z has independent standard normal entries. Note that $\varepsilon^\top\varepsilon = n$, so we have

$$S - I_d = \theta\theta^\top + B + C,$$

where $B = \frac{1}{n}ZZ^\top - I_d$ and $C = \frac{1}{n}(\theta\varepsilon^\top Z^\top + Z\varepsilon\theta^\top)$. With high probability, the spectral norm of B is at most $d/n + 2\sqrt{d/n}$ (see [DS01, Theorem II.13]), which is $O(\sqrt{d/n})$ when $n > d$, and the spectral norm of C is also $O(\sqrt{d/n})$. Then, the largest eigenvalue of $S - I_d$ differs from $\|\theta\|_2^2$ by at most $O(\sqrt{d/n})$ and thus $|\hat{s} - \|\theta\|_2| \le O(d/n)^{1/4}$. Since $\hat{v}$ maximizes $\|u^\top(S - I_d)u\|$ among all unit vectors $u \in \mathbb{R}^d$, including the direction of θ, then, applying spectral norms of B and C, we obtain that $\|\theta\|_2^2 \le (\theta^\top\hat{v})^2 + O(\sqrt{d/n})$, and consequently,

$$\|\theta - \|\theta\|_2\hat{v}\|_2^2 \le O(\sqrt{d/n}).$$

The conclusion follows from triangle inequality. □

Lemma 30. *The boundary of the space of the first $2k-1$ moments of all distributions on $\mathbb{R}$ corresponds to distributions with fewer than k atoms, while the interior corresponds to exactly k atoms.*

Proof. Given $m = (m_1, \ldots, m_{2k-1})$ that corresponds to a distribution of exactly k atoms, by [Lin89, Theorem 2A], the moment matrix $\mathbf{M}_{k-1}$ is positive definite. For any vector m' in a sufficiently small ball around m, the corresponding moment matrix $\mathbf{M}'_{k-1}$ is still positive definite. Consequently, the matrix $\mathbf{M}'_{k-1}$ is of full rank, and thus m' is a legitimate moment vector by [Las09, Theorem 3.4] (or [CF91, Theorem 3.1]). If m corresponds to a distribution with exactly $r < k$ atoms, by [Lin89, Theorem 2A], $\mathbf{M}_{r-1}$ is positive definite while $\mathbf{M}_r$ is rank deficient. Then, m is no longer in the moment space if m_{2r} is decreased. □

Acknowledgment

The authors are grateful to

References

[AK01] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[Akh65] Naum Ilich Akhiezer. *The classical moment problem: and some related questions in analysis*, volume 5. Oliver & Boyd, 1965.

[AM74] David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102, 1974.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[ARS16] Carlos Améndola, Kristian Ranestad, and Bernd Sturmfels. Algebraic identifiability of Gaussian mixtures. *International Mathematics Research Notices*, 2016.

[AS64] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1964.

[AS66] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[Ber80] Joseph Berkson. Minimum chi-square, not maximum likelihood! (with discussion). *The Annals of Statistics*, pages 457–487, 1980.

[BS10] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[Bus11] Jorge Bustamante. *Algebraic Approximation: A Guide to Past and Current Solutions*. Springer Science & Business Media, 2011.

[BV08] Spencer Charles Brubaker and Santosh Vempala. Isotropic PCA and affine-invariant clustering. In *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science, 2008.*, pages 551–560. IEEE, 2008.

[CF91] Raúl E Curto and Lawrence A Fialkow. Recursiveness, positivity, and truncated moment problems. *Houston Journal of Mathematics*, 17(4):603–635, 1991.

[Cha54] KC Chanda. A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, 41(1/2):56–61, 1954.

[Cha10] Pierre Chaussé. Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software*, 34(11):1–35, 2010.

[Che95] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233, 1995.

[Chr77] Elwin Bruno Christoffel. Sur une classe particulière de fonctions entières et de fractions continues. *Annali di Matematica Pura ed Applicata (1867-1897)*, 8(1):1–10, 1877.

[CL11] T.T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[Cra46] Harald Cramir. Mathematical methods of statistics. *Princeton U. Press, Princeton*, 1946.

[Das99] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *Foundations of computer science, 1999. 40th annual symposium on*, pages 634–644. IEEE, 1999.

[DB16] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[Dia87] Persi Diaconis. Application of the method of moments in probability and statistics. *Moments in mathematics*, (37):125–139, 1987.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[DR94] Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994.

[DR07] Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

[DS01] Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[Esc94] Michael D Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[EW95] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[Fer83] Thomas S Ferguson. Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24(1983):287–302, 1983.

[FM96] Ziding D Feng and Charles E McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 609–617, 1996.

[Gau68] Walter Gautschi. Construction of Gauss-Christoffel quadrature formulas. *Mathematics of Computation*, 22(102):251–270, 1968.

[Gau70] Walter Gautschi. On the construction of Gaussian quadrature rules from modified moments. *Mathematics of Computation*, 24(110):245–260, 1970.

[GVDV01] Subhashis Ghosal and Aad W Van Der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, pages 1233–1263, 2001.

[GW69] Gene H Golub and John H Welsch. Calculation of Gauss quadrature rules. *Mathematics of computation*, 23(106):221–230, 1969.

[GW00] C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Annals of Statistics*, 28(4):1105–1127, 2000.

[Hal05] Alastair R Hall. *Generalized method of moments*. Oxford University Press, 2005.

[Han82] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[HK15] Philippe Heinrich and Jonas Kahn. Optimal rates for finite mixture estimation. *arXiv:1507.04313*, 2015.

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

[Hub05] Peter J. Huber. *Robust statistics*. John Wiley & Sons, Inc., 2005.

[IJS01] Hemant Ishwaran, Lancelot F James, and Jiayang Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332, 2001.

[IZ02] Hemant Ishwaran and Mahmoud Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, pages 941–963, 2002.

[Jew82] Nicholas P Jewell. Mixtures of exponential distributions. *The annals of statistics*, pages 479–484, 1982.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

[Kos07] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[Kra32] Michel Krawtchouk. Sur le problème de moments. In *ICM Proceedings*, pages 127–128, 1932. Available at http://www.mathunion.org/ICM/ICM1932.2/Main/icm1932.2.0127.0128.ocr.pdf.

[KS53] Samuel Karlin and Lloyd S Shapley. *Geometry of moment spaces*. Number 12. American Mathematical Soc., 1953.

[KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

[KW56] Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

[KX03] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

[KX05] Dimitris Karlis and Evdokia Xekalaki. Mixed Poisson distributions. *International Statistical Review*, 73(1):35–58, 2005.

[Lai78] Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[Las09] Jean Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2009.

[LB00] Jonathan Q Li and Andrew R Barron. Mixture density estimation. In *Advances in neural information processing systems*, pages 279–285, 2000.

[Lin81] Bruce G Lindsay. Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work*, pages 95–109. Springer, 1981.

[Lin89] Bruce G Lindsay. Moment matrices: applications in mixtures. *The Annals of Statistics*, pages 722–740, 1989.

[Lin95] Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[Mor82] Carl N Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80, 1982.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[Ngu13] XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[PL01] Ramani S Pilla and Bruce G Lindsay. Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88(2):535–550, 2001.

[PSWS03] Javier Portilla, Vasily Strela, Martin J Wainwright, and Eero P Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image processing*, 12(11):1338–1351, 2003.

[Red81] Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, pages 225–228, 1981.

[RG97] Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[Rut62] Heinz Rutishauser. On a modification of the QD-algorithm with Graeffe-type convergence. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 13(5):493–496, 1962.

[RW84] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[RW97] Kathryn Roeder and Larry Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

[SB02] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, New York, NY, 3rd edition, 2002.

[Sch17] Konrad Schmüdgen. *The moment problem*. Springer, 2017.

[SD71] RA Sack and AF Donovan. An algorithm for Gaussian quadrature given modified moments. *Numerische Mathematik*, 18(5):465–478, 1971.

[SMA00] Wilfried Seidel, Karl Mosler, and Manfred Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3):481–487, 2000.

[ST43] James Alexander Shohat and Jacob David Tamarkin. *The problem of moments*. Number 1. American Mathematical Soc., 1943.

[Str65] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, pages 423–439, 1965.

[Sze75] G. Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, RI, 4th edition, 1975.

[Tim63] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. Pergamon Press, 1963.

[Usp37] James Victor Uspensky. Introduction to mathematical probability. 1937.

[VB96] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

[VDG96] Sara Van De Geer. Rates of convergence for the maximum likelihood estimator in mixture models. *Journal of Nonparametric Statistics*, 6(4):293–310, 1996.

[VdV00] Aad W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, Cambridge, United Kingdom, 2000.

[Vil03] C. Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, 2003.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Springer Verlag, Berlin, 2008.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[WS00] Martin J Wainwright and Eero P Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in neural information processing systems*, pages 855–861, 2000.

[WV10] Yihong Wu and Sergio Verdú. The impact of constellation cardinality on Gaussian channel capacity. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 620–628. IEEE, 2010.

[WY15] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv:1504.01227*, 2015.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[XJ96] Lei Xu and Michael I Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.