On the best approximation by finite Gaussian mixtures

Yun Ma
Tsinghua University
mayun21@mails.tsinghua.edu.cn

Yihong Wu
Yale University
yihong.wu@yale.edu

Pengkun Yang
Tsinghua University
yangpengkun@tsinghua.edu.cn

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206756

***Abstract*—We consider the problem of approximating a general Gaussian location mixture by finite mixtures. The minimum order of finite mixtures that achieve a prescribed accuracy (measured by various f-divergences) are determined within constant factors for the family of compactly supported or subgaussian mixing distributions. While the upper bound is achieved using the technique of local moment matching, the lower bound is established by relating the best approximation error to the low-rank approximation of certain trigonometric moment matrices and weighted moment matrices, followed by a refined spectral analysis of the minimum eigenvalue of these matrices. In the case of Gaussian mixing distributions, this result corrects a previous lower bound in [1].**

***Index Terms*—Gaussian mixture, density approximation, complexity measure, convergence rate, non-asymptotic analysis, moment matrix, orthogonal polynomials.**

I. Introduction

Let ϕ denote the standard normal density. For a probability distribution P on the real line, denote by f_P the marginal density of the Gaussian convolution $P * \phi$, that is

$$f_P(x) = \int \phi(x-\theta) dP(\theta). \tag{1}$$

We refer to P and f_P as the *mixing distribution* and the *mixture*, respectively. Given a general mixture f_P, the problem of interest is how to best approximate it by a finite mixture f_{P_m}, where the support size of P_m is at most m (i.e., m-atomic).

Let $d(f,g)$ denote a loss function that measures the approximation error of g by f. Concrete examples include L_p distances or f-divergences [2], the latter of which, including the total variation $\mathsf{TV}(f,g)$, squared Hellinger distance $H^2(f,g)$, the Kullback-Leibler divergence $\mathsf{KL}(f\|g)$, and the χ^2-divergence $\chi^2(f\|g)$, are the focus of the present paper. The best approximation error of f_P by an m-component mixtures is

$$\mathcal{E}^\star(m,P,d) \triangleq \inf_{P_m \in \mathcal{P}_m} d(f_{P_m}, f_P) \tag{2}$$

where $\mathcal{P}_m$ denotes the collection of all m-atomic distributions. Considering the worst instance of this pointwise quantity, we define

$$\mathcal{E}^\star(m,\mathcal{P},d) \triangleq \sup_{P\in\mathcal{P}} \mathcal{E}^\star(m,P,d) \tag{3}$$

as the worst-case approximation error over a family $\mathcal{P}$ of mixing distributions by m-component mixtures. It is well-known that the optimization problem (2) is *nonconvex* (in the location parameters) and is generally hard to solve. This shares the essential difficulty of approximation by neural nets with one hidden layer [3].

In information theory, the Gaussian convolution structure arises in the context of Gaussian channels [4], where the input and output distributions correspond to P and f_P respectively. The channel capacity determines the maximal rate at which information can be reliably transmitted, which, under the second moment constraint, is achieved by a Gaussian input distribution. In practice, the input may be constrained to be finitely valued due to modulation. To address this issue, [1] studied the Gaussian channel capacity under input cardinality constraints, in particular, the rate of convergence to the Gaussian channel capacity when the cardinality grows. It turns out that this capacity gap is precisely characterized by $\mathcal{E}^\star$ under the KL divergence – see (11).

The problem of approximation by finite mixtures also naturally arises in nonparametric statistics and empirical process theory. Classical results show that the complexity of a class of distributions, as manifested by their metric entropy, plays a crucial role in determining the rate of convergence of nonparametric density estimation [5], [6]. If the distribution family is parametric, its entropy is often determined by the dimension of the parameter space. However, nonparametric families are infinite-dimensional and determining its entropy entails more delicate analysis including finite-dimensional approximation. To describe the most economical approximation by finite mixtures, let us define

$$m^\star(\epsilon,P,d) = \min\{m\in\mathbb{N} : \exists P_m \in \mathcal{P}_m, d(f_{P_m}, f_P) \le \epsilon\}, \tag{4}$$

i.e., the smallest order of a finite mixture that approximates a given mixture f_P within a prescribed accuracy ϵ. For uniform approximation over $\mathcal{P}$, define

$$m^\star(\epsilon,\mathcal{P},d) = \sup_{P\in\mathcal{P}} m^\star(\epsilon,P,d), \tag{5}$$

which offers a meaningful *complexity measure* for the class $\{f_P : P \in \mathcal{P}\}$ and is closely related to more classical complexity notions such as the metric entropy. In fact, most of the existing constructive bounds on the metric or bracketing entropy for general Gaussian mixtures are obtained by first finding a discrete approximation then quantizing the weights and atoms, and the resulting upper bounds increase with $m^\star$ [7]–[10]. Hence, tightened upper bounds for metric entropy immediately follow.

Clearly, determining $m^\star$ and that of $\mathcal{E}^\star$ are equivalent by the dual formula

$$m^\star(\epsilon, \mathcal{P}, d) = \inf\{m : \mathcal{E}^\star(m, \mathcal{P}, d) \le \epsilon\}. \tag{6}$$

Next, we state our main theorems in terms of $m^\star$.

A. Main Results

Our main results give tight non-asymptotic characterizations of $m^\star$ for the family of compactly supported or subgaussian mixing distributions. For the former, consider

$$\mathcal{P}_M^{\mathsf{Bdd}} \triangleq \{P : P[-M, M] = 1\}. \tag{7}$$

Theorem 1: Suppose $M = O\Big((\epsilon\sqrt{\log \epsilon^{-1}})^{-1}\Big)$. Then, for $d \in \{\mathsf{TV}, H, \mathsf{KL}, \chi^2\}$,

$$m^\star(\epsilon, \mathcal{P}_M^{\mathsf{Bdd}}, d) = \Theta\left[\frac{\log \frac{1}{\epsilon}}{\log\Big(1 + \frac{1}{M}\sqrt{\log \frac{1}{\epsilon}}\Big)} \vee 1\right].^{1} \tag{8}$$

We provide some interpretations of Theorem 1. By definition, $m^\star$ increases with M and decreases with ϵ. In fact, (8) captures an "elbow-effect" depending on the relationship between M and ϵ. If the support of mixing distributions is not too wide, i.e., $M = O\Big(\sqrt{\log \frac{1}{\epsilon}}\Big)$, $m^\star$ has a slower growth with respect to ϵ as $\frac{\log \epsilon^{-1}}{\log\log \epsilon^{-1}}$; when $M = \Omega\Big(\sqrt{\log \frac{1}{\epsilon}}\Big)$ and $M = O\Big[(\epsilon\sqrt{\log \epsilon^{-1}})^{-1}\Big]$, the finite mixture needs to cover a wider range, and $m^\star$ grows as $M\sqrt{\log \frac{1}{\epsilon}}$.

Next we consider the family of σ^2-subgaussian distributions

$$\mathcal{P}_\sigma^{\mathsf{SubG}} \triangleq \{P : P[|X| > t] \le 2e^{-\frac{t^2}{2\sigma^2}}, \forall t > 0\}. \tag{9}$$

Theorem 2: Suppose $c_1 \le \sigma \le \epsilon^{-c_2}$ for some constants $c_1 > 0$ and $0 < c_2 < 1$. Then for $d \in \{\mathsf{TV}, H, \mathsf{KL}, \chi^2\}$,

$$m^\star(\epsilon, \mathcal{P}_\sigma^{\mathsf{SubG}}, d) = \Theta\left[\sigma \log \frac{1}{\epsilon} \vee 1\right]. \tag{10}$$

One way to reconcile Theorems 1 and 2 is to notice that each σ^2-subgaussian distribution is effectively supported (except for a total mass that is polynomially small in ϵ) on $[-C\sigma\sqrt{\log \frac{1}{\epsilon}}, C\sigma\sqrt{\log \frac{1}{\epsilon}}]$ for some large constant C, so that the complexity of $\mathcal{P}_\sigma^{\mathsf{SubG}}$ coincides with that of $\mathcal{P}_M^{\mathsf{Bdd}}$ with $M = \Theta(\sigma\sqrt{\log \epsilon^{-1}})$. While the upper bound essentially pursues this idea, our lower bound applies a different analysis.

We now briefly discuss the proof strategies for the main results. We prove the upper bound under the (stronger) χ^2-divergence and the lower bound under the (weaker) TV distance. For the upper bound, we extend the local moment matching argument in previous work [9], which constructs a discrete approximation by matching the moments for the mixing distribution conditioned on each subinterval in a partition of the effective support of the mixture. This approach can be further extended to distribution families with general tail conditions. The matching lower bound is the major contribution of this paper, which is shown by relating the best approximation error to the *low-rank approximation* of (trigonometric) *moment matrices* followed by a refined spectral analysis. The application of orthogonal polynomials also plays a crucial role in this analysis. In fact, the matching lower bound for Theorems 1 and 2 is proved for the uniform and the Gaussian mixing distribution, respectively.

B. Comparison with Previous Results

Below we give a brief overview of previous results. The upper bound for the compact support case is discussed in [7]–[9]. Among them the best result [9, Lemma 1] gives an upper bound of $m^\star(\epsilon, \mathcal{P}_M^{\mathsf{Bdd}}, \mathsf{TV}) = O(M\sqrt{\log \epsilon^{-1}} \vee \log \epsilon^{-1})$. Theorem 1 strengthens this result by bounding the χ^2-divergence and establishing a faster rate. For the subgaussian case, [11, Lemma 7] gives a simple $\log \epsilon^{-1}$ upper bound for 1-subgaussian family, while our Theorem 2 further discusses the effect of the subgaussian parameter σ. The specific approximation problem when P is $N(0, \sigma^2)$ is studied in [1], [12] in the context of the finite-constellation capacity. In fact,

$$\mathcal{E}^\star(m, N(0, \sigma^2), \mathsf{KL}) = C - C_m \tag{11}$$

where $C = \max_{P_X : \mathbb{E}[X^2] \le \sigma^2} I(X; X + Z) = \frac{1}{2}\log(1 + \sigma^2)$ is the Gaussian channel capacity with input X and additive noise $Z \sim N(0, 1)$, and $C_m = \max_{P_X \in \mathcal{P}_m : \mathbb{E}[X^2] \le \sigma^2} I(X; X + Z)$ is the capacity under input cardinality constraint. While quantized Gaussian only achieves an error that is polynomial in m, an exponential upper bound $\mathcal{E}^\star(m, N(0, \sigma^2), \mathsf{KL}) = O\Big(\sigma^2(\frac{\sigma^2}{1+\sigma^2})^{2m}\Big)$ is shown in [1, Theorem 8] using the Gauss quadrature. Theorem 2 generalizes this result to subgaussian family with an improved exponent for large σ.

Compared with these constructive upper bounds, the lower bound is far less understood. For Gaussian distribution, [1, Eq. (66)] claims that $\mathcal{E}^\star(m, N(0, \sigma^2), \mathsf{KL}) \ge (\frac{\sigma^2}{2+\sigma^2} + o(1))^{2m}$; however, the sketched proof turns out to be flawed, which results in a wrong dependency of the exponent on large σ. This is now corrected in Theorem 2 (see also Propositions 2 and 4), which shows that the exponential convergence is indeed tight. The exact optimal exponent, however, remains open.

C. Related Work

The problem of approximation by location mixtures is first addressed by the celebrated Tauberian theorem of Wiener [13], which gives a general characterization of whether the translation family of a given function is dense in $L^1(\mathbb{R}^d)$ or $L^2(\mathbb{R}^d)$ in terms of its Fourier transform. Convergence rates have been studied over the past few decades, with a wide range of applications in approximation theory, machine learning, and information theory [1], [12], [14]–[16]. For example, Barron [14] obtained dimension-free convergence rate for the location and scale m-mixture class of sigmoidal functions, a fundamental result in the theory of neural networks. For Gaussian models, Wu and Verdú [1] linked this problem to the Gaussian channel capacity under input cardinality constraint (cf. (11)).

[1] For any positive sequences a_n and b_n, write $a_n = O(b_n)$ if $a_n \le Cb_n$ for some absolute constant $C > 0$, $a_n = \Omega(b_n)$ when $b_n = O(a_n)$, $a_n = o(b_n)$ when $\lim a_n/b_n = 0$, and $a_n = \Theta(b_n)$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$ hold. For any $x, y \in \mathbb{R}$, $x \vee y \triangleq \max\{x, y\}$.

More recently, [17]–[19] showed the consistent approximation over various families with general location-scale mixtures. The problem is also related to some recent work [20]–[22] on the convergence rate of the empirical distribution to the underlying distribution, both smoothed by Gaussian convolution, under the so-called smoothed Wasserstein distance or f-divergences.

In the statistics literature, understanding the complexity of a distribution class plays an important role in nonparametric density and functional estimation [6]. Information-theoretic risk bounds are obtained on the basis of metric entropy for a variety of loss functions (see [23, Chapter 32] for an overview). In addition, metric entropy of the Gaussian mixture family is crucial for analyzing the statistical analysis of the sieve and nonparametric maximum likelihood estimator (MLE) in mixture models as well as posterior concentration [7]–[10], [24]–[27]. These results all rely on metric entropy of the Gaussian mixture class (with respect to truncated L_∞-norm) obtained via approximation by finite mixtures.

D. Notations

We denote $[k] = \{1, \ldots, k\}$ for $k \in \mathbb{N}$. We use the Kronecker's delta notation $\delta_{jk} = \mathbf{1}_{\{j=k\}}$. We use bold symbols to represent vectors and matrices. For vector $\mathbf{x}$, denote $\mathbf{x}^\top$ and $\mathbf{x}^\star$ as transpose and Hermitian transpose respectively, and $\mathrm{Diag}(\mathbf{x})$ as the corresponding diagonal matrix. Denote $\|\cdot\|$ as the Euclidean norm for vector or spectral norm for matrix, and let $\|\cdot\|_F$ be the matrix Frobenius norm. Write $\lambda_{\min}(\mathbf{A})$ as the smallest eigenvalue of matrix $\mathbf{A}$.

In this paper, we present the main ideas and the proof sketches. The details are omitted due to space limit.

II. Preliminaries on (Trigonometric) Moment Matrices

The theory of moments is fundamental in many areas such as probability, statistics, and approximation theory [28]. Given a distribution P and $X \sim P$, denote its k-th moment by $m_k = m_k(P) = m_k(X) = \mathbb{E}_P[X^k]$. The moment matrix associated with P of order $n+1$ is the following Hankel matrix:

$$\mathbf{M}_n = \begin{bmatrix} m_0 & m_1 & \cdots & m_n \\ m_1 & m_2 & \cdots & m_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ m_n & m_{n+1} & \cdots & m_{2n} \end{bmatrix}_{(n+1)\times(n+1)}. \quad (12)$$

Denote the vector of monomials as $\mathbf{X}_n = (1, X, \ldots, X^n)^\top$, The moment matrix of P can be equivalently represented as $\mathbf{M}_n = \mathbb{E}_P[\mathbf{X}_n \mathbf{X}_n^\top]$. Consequently, if P is discrete with no more than m atoms, the moment matrix of any order is of rank at most m, and P can be uniquely determined by its first $2m-1$ moments [28].

The above formulation can also be adapted to the trigonometric moments. For $k \in \mathbb{Z}$, denote $t_k = t_k(P) = t_k(X) = \mathbb{E}_P[e^{ikX}]$ as the k-th order Fourier coefficients (or characteristic functions) of P. Define the Toeplitz matrix

$$\mathbf{T}_n = \begin{bmatrix} t_0 & t_1 & \cdots & t_n \\ t_{-1} & t_0 & \cdots & t_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ t_{-n} & t_{-(n-1)} & \cdots & t_0 \end{bmatrix}_{(n+1)\times(n+1)} \quad (13)$$

as the trigonometric moment matrix associated with P of order $n+1$. $\mathbf{T}_n$ is equivalently the ordinary moment matrix of $Z = e^{iX}$ in the sense that $\mathbf{T}_n = \mathbb{E}_P[\mathbf{Z}_n \mathbf{Z}_n^\star]$ for $\mathbf{Z}_n = (1, Z, \ldots, Z^n)^\top$.

Our proof of the converse results relies on classical theory of moment matrices. For Hankel moment matrices, the seminal work [29] studied the asymptotic behavior of small eigenvalues for Gaussian and exponential weights. Systematical treatments for general distribution classes are given by a series of work [30]–[33]. [34] proposes a generalized result for certain forms of weighted Hankel matrices. [35] gives characterization for eigenvalues of Toeplitz forms, which applies in particular to the trigonometric moment matrices.

The moment matrices are Hermitian and positive definite provided that the corresponding distribution has infinite support [29]. In fact, the smallest eigenvalue of the moment matrix plays an important role in the derivation of the lower bound. [36] analyzes eigenvalues of Gaussian Toeplitz matrices, but the lower bound is suboptimal. [31] introduces a framework of bounding the smallest eigenvalue, extending the method of [29]. Specifically, $\mathbf{M}_n$ is related to the orthogonal polynomials on the real line, and $\mathbf{T}_n$ is related to the orthogonal polynomials on the unit circle.

For this reason, we briefly introduce the general theory of those orthogonal polynomials. Given a distribution P, the associated set of orthogonal polynomials on the real line satisfies: 1) $p_k(x)$ is a polynomial of degree k; 2) $\int p_j(x) p_k(x) dP(x) = a_k \delta_{jk}$ for some $a_k \geq 0$. For the orthogonal polynomials defined on the unit circle, the latter orthogonality condition is replaced by $\int p_j(e^{i\theta}) \overline{p_k(e^{i\theta})} dP(\theta) = a_k \delta_{jk}$. Say $\{p_k\}$ are orthonormal if they are orthogonal and normalized such that $a_k \equiv 1$. We refer the readers to [35] for a comprehensive review of orthogonal polynomials and [37] for orthogonal polynomials on the unit circle.

III. Achievability via moment matching

In this section, we give upper bounds of finite mixture approximation of the distribution family $\mathcal{P}$. For any $P \in \mathcal{P}$, we need to construct an m-atomic distribution P_m achieving a small approximation error measured by $d(f_{P_m}, f_P)$. Previous results have shown that comparing moments is useful in determining the approximation accuracy. For example, [1, Theorem 8] considers the m-point Gaussian quadrature with a scale parameter σ that matches the first $2m-1$ moments of $N(0, \sigma^2)$. For the compactly supported family, [9, Lemma 1] provides a moment matching approximation.

In our non-asymptotic analysis framework, when the relative scale of the parameter of the distribution family compared

with m changes, different treatments are needed. Consider the compactly supported family as an example. The previous result [9] becomes suboptimal unless M grows with m at a certain rate, as shown in (14). To achieve a tight upper bound, we bound the χ^2-divergence by moment differences and construct moment matching approximations either globally or locally depending on the relationship between parameters. The Gauss quadrature serves as a classical approach to the global discrete approximation in the sense of matching moments. In the case of local approximations, we partition the support set into subintervals and match the moments of each conditional distribution. Similar constructions have appeared in [9] by applying Caratheódory's theorem to the conditional moments.

Then we extend our non-asymptotic analysis for the compactly supported family to the subgaussian family via a truncation argument. This approach is also applicable for other distribution families with general tail bounds (e.g., subexponential tails and bounded moment conditions), by applying (17) with the corresponding tail probability bounds.

Proposition 1: There exists a universal constant $C > 0$ such that for $m \in \mathbb{N}$ and $M > 0$:

$$\mathcal{E}^\star(m, \mathcal{P}_M^{\mathsf{Bdd}}, \chi^2) \le \begin{cases} \exp\left(-m \log \frac{m}{M^2}\right), & m \ge CM^2; \\ \exp\left(-\frac{\log C}{2C} \frac{m^2}{M^2}\right), & \sqrt{C}M \le m \le CM^2. \end{cases} \tag{14}$$

Proposition 2: There exist universal constants $C, C' > 0$ such that for $m \in \mathbb{N}$ and $0 < \sigma \le C'm$:

$$\mathcal{E}^\star(m, \mathcal{P}_\sigma^{\mathsf{SubG}}, \chi^2) \le \exp\left(-Cm \log\left(1 + \frac{1}{\sigma}\right)\right). \tag{15}$$

Note that the upper bound of $m^\star$ in Theorems 1 and 2 directly follows from Propositions 1 and 2 in view of (6). Additionally, as will be shown in Proposition 5, we have that $\mathcal{E}^\star(m, \mathcal{P}_M^{\mathsf{Bdd}}, \mathsf{TV})$ (resp. $\mathcal{E}^\star(m, \mathcal{P}_\sigma^{\mathsf{SubG}}, \mathsf{TV})$) is at least some constant when $M = \Omega(m)$ (resp. $\sigma = \Omega(m)$). This result coincides with Propositions 1 and 2 where the upper bounds degenerate to constants for very large M or σ. It follows that no consistent approximation exists in these regimes.

The following lemma stated in [38, Lemma 9] is useful in the derivation of χ^2-divergence upper bound, which bounds the χ^2-divergence by comparing moments.

Lemma 1: Suppose all moments of P and Q exist, and Q is centered with variance σ^2. Then

$$\chi^2(f_P \| f_Q) \le e^{\frac{\sigma^2}{2}} \sum_{j \ge 1} \frac{(m_j(P) - m_j(Q))^2}{j!}. \tag{16}$$

Next we sketch the proof of Propositions 1 and 2.

Proof Sketch of Propositions 1 and 2: We first show the upper bound for compactly supported distributions, then prove the result for subgaussian family via a truncation argument.

- *Global moment matching:* Fix a distribution $P \in \mathcal{P}_M^{\mathsf{Bdd}}$. For any $m \in \mathbb{N}$, the Gauss quadrature rule implies the existence of some $P_m \in \mathcal{P}_m$ that matches the first $2m-1$ moments of P. If $m \ge CM^2$, applying Lemma 1 to the centered distributions yields an $\exp\left(-\Omega\left(m \log \frac{m}{M^2}\right)\right)$ upper bound.
- *Local moment matching:* If $\sqrt{C}M \le m \le CM^2$, we partition the interval $[-M, M]$ into k subintervals $I_j = [-M + (j-1)\frac{2M}{k}, -M + j\frac{2M}{k}]$ for $j \in [k]$, where $k = \min\{n \in \mathbb{N} : \lfloor m/n \rfloor \ge C(M/n)^2\}$. Note that k is finite when $m \ge \sqrt{C}M$.
  Specifically, we apply the results from global moment matching to conditional distributions. Denote $P_{(j)}$ as the conditional version of P on I_j, that is, for any Borel set A, $P_{(j)}(A) = \frac{P(I_j \cap A)}{P(I_j)}$. Then $P = \sum_{j=1}^k P(I_j) P_{(j)}$. According to the $m \ge CM^2$ case, for each $P_{(j)}$, there exists some $\tilde{P}_j$ supported on at most $\lfloor m/k \rfloor$ atoms such that $\chi^2(f_{\tilde{P}_j} \| f_{P_{(j)}}) \le \exp\left(-\Omega\left(\frac{m^2}{M^2}\right)\right)$. Let $P_m = \sum_{j=1}^k P(I_j) \tilde{P}_j$, then by Jensen's inequality we have $\chi^2(f_{P_m} \| f_P) \le \exp\left(-\Omega\left(\frac{m^2}{M^2}\right)\right)$.
- *Truncation under subgaussian conditions:* Fix $P \in \mathcal{P}_\sigma^{\mathsf{SubG}}$. Let $t \in \mathbb{R}_+$ be a parameter to be determined. Define P_t and P_t^c as the conditional version of P on $I_t \triangleq [-t, t]$ and I_t^c, respectively. For any distribution Q, calculation shows that

$$\chi^2(f_Q \| f_P) \le \frac{2}{P(I_t)} \left(\chi^2(f_Q \| f_{P_t}) + P(I_t^c)\right), \tag{17}$$

  which converts the approximation problem to the subgaussian P into that of its conditional version P_t, which is compactly supported, with an additional term of subgaussian tail. Let Q be the approximation of P_t achieving the convergence rate in (14) for $M = t$. Then the two terms on the right-hand side of (17) are bounded by (14) and (9) respectively, as functions of m, t, and σ. Finally, the result (15) follows from choosing a suitable $t = t^\star(m, \sigma)$ to balance the terms, which also varies depending on the regimes in (14).

■

IV. Converse via spectra of moment matrices

In this section, we give lower bounds for the best approximation error by finite mixture approximation. To do so, we propose a general framework applicable to various distribution families and obtain the first valid lower bound that applies to the approximation error of any finite mixture. Note that for each $P \in \mathcal{P}$, $m^\star(\epsilon, P, d)$ lower bounds $m^\star(\epsilon, \mathcal{P}, d)$. We choose $\mathrm{Unif}[-M, M]$, the uniform distribution on $[-M, M]$, and $N(0, \sigma^2)$ as the representative in the family of compactly supported and σ^2-subgaussian mixing distributions, respectively. The connection between these two choices is that the truncated version of $N(0, \sigma^2)$ in $[-M, M]$ with a sufficiently large σ is essentially $\mathrm{Unif}[-M, M]$. Also, they are one of the hardest cases to approximate in their respective class. The following results provide matching lower bounds to Propositions 1 and 2, and also imply the desired lower bounds for $m^\star$.

Proposition 3: Suppose that $m \geq 2\sigma$. There exists some universal constant c such that

$$\mathcal{E}^\star(m, N(0,\sigma^2), \mathsf{TV}) \geq c \exp\left(-\left(4+\frac{\pi^2}{24}\right)\frac{m}{\sigma} - 3\log m\right). \quad (18)$$

Proposition 4: There exists a universal constant C_1 such that the following holds. When $m \geq C'M^2$ for some sufficiently large constant C',

$$\mathcal{E}^\star(m, \mathrm{Unif}[-M,M], \mathsf{TV}) \geq \exp\left(-C_1 m \log\left(\frac{m}{M^2}\right)\right); \quad (19)$$

and for any $M > 0$, $m \in \mathbb{N}$,

$$\mathcal{E}^\star(m, \mathrm{Unif}[-M,M], \mathsf{TV}) \geq \exp\left(-C_1\frac{m^2}{M^2} - \log 4m\right). \quad (20)$$

Before sketching the proof, we briefly introduce the main ideas in our analysis. First, by applying the variational representation with the set of trigonometric functions, the lower bound reduces to comparing two trigonometric moment matrices. Since the trigonometric moment matrix of an m-atomic distribution is of rank at most m, we can think of the problem from the perspective of low-rank approximation towards that of $\mathrm{Unif}[-M,M]$, which always has full rank. By the Eckart-Young-Mirsky theorem, it is sufficient to bound the smallest eigenvalue of the full-rank matrix. To handle this problem, we further extend the spectral analysis procedure adopted in [29] and [31] to the scaled trigonometric moment matrices encountered in the derivation. The results are finally obtained by evaluating the spectral norm of certain matrix consisting of the coefficients of the corresponding orthogonal polynomials.

Proof Sketch of Proposition 3: Denote $X \sim P$, $Z \sim N(0,1)$, $Y = X + Z \sim f_P$ and $X_m \sim P_m$ for $P_m \in \mathcal{P}_m$.

- *Lower bound via trigonometric moments:* Let $P = N(0,\sigma^2) \in \mathcal{P}_\sigma^{\mathsf{SubG}}$. Consider the variational representation for total variation distance $\mathsf{TV}(P,Q) = \frac{1}{2}\sup_{\|f\|_\infty \leq 1} |\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)|$. Set the test function as $f_\omega(x) = \exp(i\omega x)$ and choose $\omega \in \delta\mathbb{Z}$, where δ is a frequency parameter to be chosen. Using the Gaussian characteristic function, we obtain the lower bound

$$\mathsf{TV}(f_{P_m}, f_P) \geq \sup_{k\in\mathbb{Z}, |k|\leq m} \frac{|t_k(\delta X_m) - t_k(\delta X)|}{2\exp(k^2\delta^2/2)} \quad (21)$$

$$\geq \frac{\|\mathbf{T}_m(\delta X_m) - \mathbf{T}_m(\delta\sigma Z)\|_F}{2(m+1)\exp(m^2\delta^2/2)} \quad (22)$$

$$\geq \frac{\lambda_{\min}(\mathbf{T}_m(\delta\sigma Z))}{2(m+1)\exp(m^2\delta^2/2)} \quad (23)$$

  where (23) follows from the Eckart-Young-Mirsky theorem and the fact that $\mathrm{rank}(\mathbf{T}_m(\delta X_m)) \leq m$.
- *Spectral Analysis for trigonometric moment matrix:* For $\mathbf{T} \triangleq \mathbf{T}_m(\delta\sigma Z)$, the Rayleigh quotient indicates that

$$\lambda_{\min}(\mathbf{T}) = \min_{\mathbf{x}\in\mathbb{R}^{m+1}\setminus\{\mathbf{0}\}} \frac{\mathbf{x}^\top\mathbf{T}\mathbf{x}}{\|\mathbf{x}\|^2}. \quad (24)$$

  Let $\pi_m(w) = \sum_{j=0}^m x_j w^j$. Also expand π_m with orthogonal polynomials $\{\varphi_k\}$ on the unit circle as $\pi_m(w) = \sum_{k=0}^m c_k\varphi_k(w)$. Specifically, $\{\varphi_k(x)\}$ is the Rogers-Szegö polynomials [37, Eq. (1.6.51)]. Denote $\mathbf{x} = (x_0,\ldots,x_m)^\top$ and $\mathbf{c} = (c_0,\ldots,c_m)^\top$ as the two series of coefficients respectively. It follows from the definition that $\mathbf{x} = \mathbf{R}_m^\top\mathbf{c}$, where $\mathbf{R}_m$ is an $(m+1)\times(m+1)$ matrix that encodes the coefficients of $\{\varphi_k(x)\}_{k=0}^m$. Moreover, the orthogonality of $\{\varphi_k\}$ implies that

$$\|\mathbf{c}\|_2^2 = \mathbb{E}\left[|\pi_m(e^{i\delta\sigma Z})|^2\right] = \mathbf{x}^\top\mathbf{T}\mathbf{x}. \quad (25)$$

  As a result, (24) becomes

$$\lambda_{\min}(\mathbf{T}) = \min_{\mathbf{c}\in\mathbb{R}^{m+1}} \frac{\|\mathbf{c}\|^2}{\|\mathbf{R}_m\mathbf{c}\|^2} \geq \frac{1}{\|\mathbf{R}\|_F^2}. \quad (26)$$

  We then upper bound $\|\mathbf{R}\|_F^2$ by a detailed calculation using the explicit form of $\{\varphi_k(x)\}$. (18) is finally obtained by a suitable choice of δ to balance the terms in (23).

■

The proof of Proposition 4 also follows similar strategies. Specifically, (20) simply follows from the expansion under the standard monomial basis; (19) requires extra efforts as we are unaware of the general explicit formula of the associated orthogonal polynomials on the unit circle. To derive (19), we turn to the variational lower bound of the χ^2-divergence and determine the smallest eigenvalue of a certain *weighted moment matrix* via a similar spectral analysis procedure. Then, we convert the χ^2-divergence lower bound to a TV lower bound by analyzing the Gaussian tails. Consequently, (19) and (20) together yield a tight TV lower bound for Proposition 4. However, for the case of Proposition 3, this indirect approach only yields suboptimal dependency on σ.

Finally, we consider the regime when σ diverges. In this case, we expect more components are needed to approximate the mixture $N(0,1+\sigma^2)$ and it is natural to conjecture that the approximation is impossible unless m is at least proportional to σ. This however is not captured by Proposition 3 due to the extra polynomial term in m. The next result makes this intuition precise in a strong sense for both the Gaussian and the uniform mixing distributions.

Proposition 5: For some universal constant C, we have that

$$\mathcal{E}^\star(m, N(0,\sigma^2), \mathsf{TV}) \geq 1 - C\frac{m}{\sigma}\sqrt{\log\frac{\sigma}{m}};$$

$$\mathcal{E}^\star(m, \mathrm{Unif}[-M,M], \mathsf{TV}) \geq 1 - C\frac{m}{M}\sqrt{\log\frac{M}{m}}.$$

The impossibility of non-trivial approximation for small m can be explained as follows. If the mixing distribution P is smooth enough with a large M (or σ), then f_P can be flat over a large region. In comparison, when m is too small, f_{P_m} inevitably has many modes and cannot approximate f_P well.

ACKNOWLEDGMENTS

Y. Wu is supported in part by the NSF Grant CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. P. Yang is supported in part by the NSFC Grant 12101353 and Tsinghua University Initiative Scientific Research Program.

REFERENCES

[1] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 620–628.

[2] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, pp. 146–158, 1975.

[3] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[4] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley & Sons, 2006.

[5] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, no. 5, pp. 1564 – 1599, 1999.

[6] A. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes: with applications to statistics*. New York: Springer Series in Statistics. Springer-Verlag, 1996.

[7] S. Ghosal and A. van der Vaart, "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, vol. 35, no. 2, pp. 697 – 723, 2007.

[8] S. Ghosal and A. W. van der Vaart, "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities," *The Annals of Statistics*, vol. 29, no. 5, pp. 1233 – 1263, 2001.

[9] C.-H. Zhang, "Generalized maximum likelihood estimation of normal mixture densities," *Statistica Sinica*, vol. 19, no. 3, pp. 1297–1318, 2009.

[10] S. Saha and A. Guntuboyina, "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising," *The Annals of Statistics*, vol. 48, no. 2, pp. 738 – 762, 2020.

[11] Y. Polyanskiy and Y. Wu, "Self-regularizing property of nonparametric maximum likelihood estimator in mixture models," *arXiv:2008.08244*, 2020.

[12] Y. Wu and S. Verdú, "Functional properties of MMSE," *2010 IEEE International Symposium on Information Theory*, pp. 1453–1457, 2010.

[13] N. Wiener, "Tauberian theorems," *Annals of mathematics*, vol. 33, no. 1, pp. 1–100, 1932.

[14] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[15] P. J. S. Ferreira, "Neural networks and approximation by superposition of Gaussians," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 1997, pp. 3197–3200.

[16] J. A. Cuesta-Albertos, C. M. Bea, and J. M. R. Rodríguez, "Shape of a distribution through the l_2-Wasserstein distance," *Distributions with Given Marginals and Statistical Modelling*, pp. 51–61, 2002.

[17] H. D. Nguyen and G. McLachlan, "On approximations via convolution-defined mixture models," *Communications in Statistics - Theory and Methods*, vol. 48, no. 16, pp. 3945–3955, 2019.

[18] T. T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan, "Approximation by finite mixtures of continuous density functions that vanish at infinity," *Cogent Mathematics & Statistics*, vol. 7, no. 1, p. 1750861, 2020.

[19] T. Nguyen, F. Chamroukhi, H. D. Nguyen, and G. J. McLachlan, "Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces," *Communications in Statistics - Theory and Methods*, pp. 1–12, May 2022.

[20] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, 2020.

[21] H.-B. Chen and J. Niles-Weed, "Asymptotics of smoothed Wasserstein distances," *Potential Analysis*, vol. 56, no. 4, pp. 571–595, 2022.

[22] A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin, "Rate of convergence of the smoothed empirical Wasserstein distance," *arXiv:2205.02128*, 2022.

[23] Y. Polyanskiy and Y. Wu, *Information theory: from coding to statistical learning*. Cambridge University Press, 2023, draft: http://www.stat.yale.edu/~yw562/teaching/itbook-export.pdf.

[24] X. Shen and W. H. Wong, "Convergence rate of sieve estimates," *The Annals of Statistics*, vol. 22, no. 2, pp. 580 – 615, 1994.

[25] W. H. Wong and X. Shen, "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs," *The Annals of Statistics*, vol. 23, no. 2, pp. 339 – 362, 1995.

[26] C. R. Genovese and L. Wasserman, "Rates of convergence for the Gaussian mixture sieve," *The Annals of Statistics*, vol. 28, no. 4, pp. 1105 – 1127, 2000.

[27] J. A. Soloff, A. Guntuboyina, and B. Sen, "Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood," 2021.

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

[29] G. Szegö, "On some Hermitian forms associated with two given curves of the complex plane," *Transactions of the American Mathematical Society*, vol. 40, pp. 450–461, 1936.

[30] H. Widom and H. Wilf, "Small eigenvalues of large Hankel matrices," *Proceedings of the American Mathematical Society*, vol. 17, no. 2, pp. 338–344, 1966.

[31] Y. Chen and N. Lawrence, "Small eigenvalues of large Hankel matrices," *Journal of Physics A*, vol. 32, pp. 7305–7315, 1999.

[32] C. Berg, Y. Chen, and M. E. H. Ismail, "Small eigenvalues of large Hankel matrices: The indeterminate case," *Mathematica Scandinavica*, vol. 91, pp. 67–81, 1999.

[33] Y. Chen and D. Lubinsky, "Smallest eigenvalues of Hankel matrices for exponential weights," *Journal of Mathematical Analysis and Applications*, vol. 293, pp. 476–495, 05 2004.

[34] F. Štampach and P. Šťovíček, "Spectral representation of some weighted Hankel matrices and orthogonal polynomials from the Askey scheme," *Journal of Mathematical Analysis and Applications*, vol. 472, no. 1, pp. 483–509, 2019.

[35] G. Szegö, *Orthogonal polynomials*, 4th ed. Providence, RI: American Mathematical Society, 1975.

[36] J. Pasupathy and R. Damodar, "The Gaussian Toeplitz matrix," *Linear Algebra and its Applications*, vol. 171, pp. 133–147, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0024379592902559

[37] B. Simon, *Orthogonal polynomials on the unit circle. Part 1: Classical theory*. AMS Colloquium Publications, 01 2005, vol. 54.

[38] Y. Wu and P. Yang, "Optimal estimation of Gaussian mixtures via denoised method of moments," *The Annals of Statistics*, vol. 48, no. 4, pp. 1981 – 2007, 2020.