Self-regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models

Yury Polyanskiy and Yihong Wu*

August 31, 2020

Abstract

Introduced by Kiefer and Wolfowitz [KW56], the nonparametric maximum likelihood estimator (NPMLE) is a widely used methodology for learning mixture models and empirical Bayes estimation. Sidestepping the non-convexity in mixture likelihood, the NPMLE estimates the mixing distribution by maximizing the total likelihood over the space of probability measures, which can be viewed as an extreme form of overparameterization.

In this paper we discover a surprising property of the NPMLE solution. Consider, for example, a Gaussian mixture model on the real line with a subgaussian mixing distribution. Leveraging complex-analytic techniques, we show that with high probability the NPMLE based on a sample of size n has $O(\log n)$ atoms (mass points), significantly improving the deterministic upper bound of n due to Lindsay [Lin83a]. Notably, any such Gaussian mixture is statistically indistinguishable from a finite one with $O(\log n)$ components (and this is tight for certain mixtures). Thus, absent any explicit form of model selection, NPMLE automatically chooses the right model complexity, a property we term *self-regularization*. Extensions to other exponential families are given. As a statistical application, we show that this structural property can be harnessed to bootstrap existing Hellinger risk bound of the (parametric) MLE for finite Gaussian mixtures to the NPMLE for general Gaussian mixtures, recovering a result of Zhang [Zha09].

Contents

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu. Y.W. is with the Department of Statistics and Data Science, Yale University, New Haven, CT, email: yihong.wu@yale.edu.

1 Introduction

Nonparametric maximum likelihood estimator (NPMLE) is a useful methodology for various statistical problems such as density estimation, regression, censoring model, deconvolution, and mixture models (see the monographs [GW92, GJ14]). Oftentimes optimizing over a massive (infinite-dimensional) parameter space can lead to undesirable properties, such as non-existence[1] and roughness, and runs the risk of overfitting. These shortcomings can be remedied by the method of sieves [Gre81] or explicit regularization [GG71, Sil82] at the expense of losing the main advantages of the NPMLE – the full adaptivity (tuning parameters-free) and the computational tractability. However, for certain problems including shape constraints (such as monotonicity [Gre56, Bir89] and log-concavity [DR09, CSS10, DW16, KS16]) and mixture models [Lin95, Zha09, SG20], a striking observation is that *unpenalized* NPMLE achieves superior performance and has become the method of choice for both theoretical investigation and practical computation. While basic structural properties of NPMLE has been well understood, these results are frequently too conservative to explain its superior statistical performance. This paper studies the *typical* structure of NPMLE for mixture models as well as its statistical consequences.

Consider a parametric family of densities $\{p_\theta : \theta \in \Theta\}$ with respect to some dominating measure μ on $\mathbb{R}$, where the parameter space Θ is assumed to be a subset of $\mathbb{R}$. Given a mixing distribution (prior) π on Θ, we denote the induced mixture density as:

$$p_\pi(x) \triangleq \int_\Theta p_\theta(x)\pi(d\theta). \tag{1}$$

Introduced by Kiefer and Wolfowitz [KW56] (see also an earlier abstract by Robbins [Rob50]), the NPMLE for the mixing distribution is defined as a maximizer of the mixture likelihood given n data points $x_1, \ldots, x_n$:

$$\widehat{\pi}_{\mathrm{NPMLE}} \in \arg\max_{\pi \in \mathcal{M}(\Theta)} \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i), \tag{2}$$

where $\mathcal{M}(\Theta)$ denotes the collection of all probability measures on Θ. We refer the readers to the monograph of Lindsay [Lin95] for a systematic treatment on the NPMLE. Although the convex optimization problem (2) is infinite-dimensional, over the years various computationally efficient algorithms have been obtained; see [Lin95, Chapter 6] and more recent developments in [JZ09, KM14]. The NPMLE provides a highly useful primitive for empirical Bayes and compound estimation problem, in which one first apply the NPMLE to learn a prior then execute the corresponding Bayes estimator of the learned prior. This strategy can be used as a universal means for denoising and achieves the state-of-the-art empirical Bayes performance [JZ09].

We summarize a few known structural properties of the NPMLE. The first existence and uniqueness result was obtained by Simar [Sim76] for the Poisson mixture, followed by Jewell [Jew82] for mixtures of exponential distributions. It was shown that the (unique) solution $\widehat{\pi}_{\mathrm{NPMLE}}$ to the optimization problem (2) is a discrete distribution, whose number of atoms (mass points) is at most the number of distinct values of the observations and consequently at most the sample size n.[2] These results have been significantly extended in [Lai78, Lin83a, Lin83b, GW92, LR93] which show that the NPMLE solution is unique and n-atomic for all exponential families with densities with respect to the Lebesgue measure. Although the bound $|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le n$ is the best possible, which seems to suggest the estimator exhibits significant overfitting (since an n-component mixture requires

[1] For example, it is easy to see that NPMLE for the class of unimodal densities does not exist.

[2] The existence of such an atomic maximizer is a direct consequence of Carathéodary theorem [Egg58, Chapter 2, Theorem 18]; the uniqueness takes effort to show.

$2n-1$ parameters to describe), in practice the support size is much smaller than n. Understanding this phenomenon is the main motivation behind this work.

To anchor the discussion, let us focus on the Gaussian location model, where $p_\theta(x) = \varphi(x-\theta)$ is the density of $N(\theta,1)$ and $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ is the standard normal density, so that p_π is the convolution $\pi * \varphi$. It is well known that for finite Gaussian mixtures, the likelihood is non-concave in the location parameters; furthermore, spurious local maxima can exist even with infinite sample size [JZB+16] which pose difficulty for heuristic methods such as the EM algorithm. To sidestep the non-convexity, the approach of NPMLE can be viewed as an extreme form of *overparameterization*, which postulates a potentially infinite Gaussian mixture so as to convexify the optimization problem. Since overparameterized models are prone to overfitting, it is of significant interest to understand the typical model size fitted by the NPMLE. To this end, the worst-case bound $|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le n$ is not useful. In fact, this bound can be tight, e.g., when the n observations are extremely spaced out [Lin95, p. 116]. This, however, is *not* a typical configuration of the sample if it consists of independent observations. Indeed, in practice it has been observed that NPMLE tends to fit a Gaussian mixture with much fewer components than n. This not only explains the absence of overfitting, but is a highly desirable property for interpretability of the NPMLE solution, and is clearly not explained by the worst-case bound. Based on numerical evidence, Koenker and Gu [KG19] suggested that the number of atoms of the NPMLE is typically $O(\sqrt{n})$. As our main result shows next, it is in fact $O(\log n)$.

Theorem 1 (Gaussian mixture model)**.** *Let p_θ be the density of $N(\theta,1)$. Let $x_{\min} = \min_{i\in[n]} x_i$ and $x_{\max} = \max_{i\in[n]} x_i$. Then there exists a universal constant C_0, such that*

$$|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le C_0(x_{\max} - x_{\min})^2. \tag{3}$$

Consequently, suppose $x_1,\dots,x_n$ are drawn independently from $\pi * N(0,1)$ for some s-subgaussian mixing distribution π, i.e., $\int \pi(d\theta)e^{t\theta} \le e^{st^2/2}$ for all $t\in\mathbb{R}$. Then for any $\tau > 0$, there exists some constant $C = C(s,\tau)$ such that with probability at least $1-n^{-\tau}$,

$$|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le C\log n. \tag{4}$$

A few remarks are in order:

Remark 1 (Tightness of Theorem 1)**.** The $O(\log n)$ upper bound in Theorem 1 is tight in the following sense:

- First, it is necessary to select a model of size $\Omega(\log n)$ in order to be compatible with existing statistical guarantees on the NPMLE. Indeed, suppose the true density p_π is $N(0,\sigma^2)$ for some $\sigma > 1$ (i.e. the mixing distribution is another Gaussian). It is known that the Hellinger distance between $N(0,\sigma^2)$ and any k-Gaussian mixture (k-GM) with unit variance is at least $\exp(-O(k))$ [WV10]. Therefore, if $|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le c\log n$ for some small constant c, the bias would be too big, violating the Hellinger risk bound $\mathbb{E}[H^2(p_\pi, p_{\widehat{\pi}_{\mathrm{NPMLE}}})] = O(\frac{\log^2 n}{n})$ on the NPMLE [Zha09] (see Section 4).

- On the other hand, there is no statistical value to fit a model of size bigger than $\Omega(\log n)$. Indeed, it is easy to show by moment matching (see, e.g., [WY20, Lemma 8]) that for any subgaussian π, $p_\pi = \pi * N(0,1)$ can be approximated by a k-GM within total variation (TV) distance $\exp(-\Omega(k))$. Therefore, there exists a k-GM density $p_{\pi'}$ with $k = C\log n$, such that $\mathrm{TV}(p_\pi, p_{\pi'}) = o(1/n)$. As such, one can couple the original sample $X_1,\dots,X_n$ drawn from p_π with the sample $X_1',\dots,X_n'$ drawn from $p_{\pi'}$, so that with probability $1-o(1)$, $X_i = X_i'$ for

all $i = 1, \dots, n$. From this simulation perspective, $p_{\pi'}$ is an equally plausible "ground truth" that explains the data, and hence, statistically speaking, there is no reason to fit a mixture model with more than $C \log n$ components.

From the above two aspects, one can view $\Theta(\log n)$ as the "effective dimension" of the Gaussian mixture model with subgaussian mixing distributions (i.e. each doubling of the sample size *unlocks* a new parameter of the model class). Thus it is a remarkable fact that NPMLE picks up the right model size without explicit model selection penalty. For this reason, we refer to the phenomenon described in Theorem 1 as *self-regularization*. In order to further quantify self-regularization and determine what the correct model size is, we formalize a framework called the statistical degree in Section 5.1.

Remark 2 (Poisson mixture)**.** Using only classical results, one can get a glimpse of the self-regularization property of the NPMLE by considering the Poisson model. Suppose $x_1, \dots, x_n$ are drawn independently from a Poisson mixture for some subexponential mixing distribution on the mean parameter. Since the observations are non-negative integers, the number of distinct values of in the sample is at most $x_{\max} + 1$, which is $O(\log n)$ with probability $1 - o(1)$ by a union bound. Thus the NPMLE for the Poisson mixture is $O(\log n)$-atomic with high probability, which is again the optimal model size. Clearly, this argument does not generalize to continuous distributions such as the Gaussian mixture model in which all observations are distinct with probability one. Nevertheless, the *range* of the data still grows logarithmically and Theorem 1 shows that the number of atoms in the NPMLE can be bounded by the squared range.

Remark 3 (Model selection and penalized MLE)**.** Define the likelihood value of the best k-GM fit as

$$L_{\mathrm{opt}}(k) \triangleq \max_{\pi \in \mathcal{M}_k} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i). \tag{5}$$

Note that this is a non-convex optimization problem, since the likelihood is not concave in the location parameters. As $k \to \infty$, $L_{\mathrm{opt}}(k)$ approaches the objective value of the (convex optimization) NPMLE (2). Theorem 1 shows that with high probability with respect to the randomness of the sample, the curve $k \mapsto L_{\mathrm{opt}}(k)$ *flattens* when k surpasses $C \log n$ for some constant C. This has the following immediate bearing on model selection. Various criteria (such as AIC or BIC [Ler92, Ker00]) have been proposed for the mixture model: given a penalty function $\mathrm{pen}(k)$ that strictly increases in k, select a model size by solving

$$\max_{k=1,\dots,K} \{L_{\mathrm{opt}}(k) - \mathrm{pen}(k)\}$$

where K is a pre-defined maximal model size. Theorem 1 shows that for Gaussian mixtures, regardless of the actual model size, there is no need to choose K bigger than $C \log n$, which also suggests $K = C \log n$ a universal choice. It is shown in [Ler92, Ker00] that BIC (with $\mathrm{pen}(k) = \frac{k}{2} \log n$) is consistent in estimating the order of the mixture model. Complementing this result, Theorem 1 shows that regardless of the choice of penalty, any penalized MLE will not choose a model size bigger than $C \log n$ with high probability.

Remark 4 (Comparison with shape-constrained estimation)**.** The structure of the NPMLE is much less well understood for mixture models than for shape-constrained estimation. For example, for monotone density, the NPMLE (known as the Grenander estimator [Gre56]) of a decreasing density on $[0, 1]$ with n observations is known to be piecewise constant with at most n pieces. Denote by k_n by its number of pieces. Under appropriate conditions it is shown that in general $k_n = O(n^{1/3})$

with high probability [Gro11, Lemma 3.1]. In the special case where the data are drawn from the uniform distribution, k_n is asymptotically $N(\log n, \log n)$ [GL93, Theorem 2]. These results are made possible thanks to an explicit characterization of the NPMLE in terms of empirical processes, a luxury we do not have in mixture models.

On the other hand, there is a clear analogy for the structural behavior of NPMLE for monotone density and mixture models: In the former, if the data are drawn from uniform distribution (one-piecewise constant), the NPMLE will fit a piecewise constant density with $O(\log n)$ pieces; in the latter, if the data are drawn from a single Gaussian, the NPMLE will fit a Gaussian mixture with $O(\log n)$ components. From this perspective one could say there is some mild overfitting in NPMLE; nevertheless, it is a modest (and fair) price to pay for being completely automatic and computationally attractive.

Theorem 1 is further extended in Theorem 3 to general exponential families, which shows that there is some degree of universality to the $O(\log n)$ upper bound. As we will see in Section 2, bounding the number of atoms in NPMLE boils down to counting critical points of functions of the form $F(\theta) = \sum_{i=1}^n w_i p_\theta(x_i)$, where w_i's are nonnegative weights. We accomplish this task using methods from complex analysis. Roughly speaking, the strategy is as follows: First, we localize the roots of F' in a compact interval, say $[-r, r]$. Then, we bound the number of zeros of F' in the complex disk of radius r, in terms of its maximal modulus on the complex disks. This leads to a deterministic upper bound, as a function of the sample, on the number of atoms of the NPMLE. Finally, we analyze the high-probability behavior of this upper bound when the sample consists of iid observations. We note that in the special case of Gaussian model, counting the number of critical points of F has been studied, independently, in the context of a seemingly unrelated information-theoretic problem [DYPS20]; see also Section **??**.

Note that statistical guarantees on NPMLE, typically in terms of Hellinger risk of density estimation, have been obtained in [GvdV01, GvdV07, Zha09, SG20]. These results follow the usual route of analyzing MLE using entropy numbers and only uses the zeroth order optimality condition, and therefore cannot produce any *structural* information on the optimizer such as the number of atoms. (For example, such analysis applies equally to $\widehat{\pi}_{\mathrm{NPMLE}}$ convolved with an arbitrarily small Gaussian, which now has infinitely many atoms.) A structural result, such as Theorem 1, can only be obtained by "opening up the optimization blackbox", by examining the exact optimality conditions, as we indeed do below. In turn, a pleasant consequence of the self-regularizing property is a simpler proof of the statistical guarantee of the NPMLE in [Zha09], by bootstrapping existing that of the (parametric) MLE for finite Gaussian mixtures [MM11] to general mixtures.

The remainder of the paper is organized as follows: Section 2 recalls the first-order optimality condition for the NPMLE. Following [Lin83b], Section 3 studies the NPMLE for mixtures of exponential family and bounds its number of atoms as well as analyzing its typical behavior. Section 4 derives Hellinger risk bounds for the NPMLE in the Gaussian mixture model. Section 5 concludes the paper by discussing the concept of "self-regularization", its ramifications and open problems.

Throughout the paper, we use standard asymptotic notations: For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ if $a_n \geq cb_n$ holds for all n and some absolute constant $c > 0$, $a_n \lesssim b_n$ if $a_n \gtrsim b_n$, and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold; the notations $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ are similarly defined. We write $a_n = o(b_n)$ or $b_n = \omega(b_n)$ or $a_n \ll b_n$ or $b_n \gg a_n$ if $a_n/b_n \to 0$ as $n \to \infty$.

2 Optimality condition

In this section we review the first-order optimality condition (both necessary and sufficient) for characterizing the NPMLE. Denote the objective function in (2) by

$$\ell(\pi) = \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i).$$

Let $\widehat{\pi} = \widehat{\pi}_{\mathrm{NPMLE}}$. Since $\ell(\widehat{\pi}) \geq \ell((1-\epsilon)\widehat{\pi} + \epsilon\delta_\theta)$ for any $\epsilon \in [0,1]$ and any $\theta \in \mathbb{R}$, we arrive at the first-order optimality condition $\frac{d}{d\epsilon}\ell((1-\epsilon)\widehat{\pi} + \epsilon\delta_\theta)\big|_{\epsilon=0} \leq 0$, namely,[3]

$$D_{\widehat{\pi}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \frac{p_\theta(x_i)}{p_{\widehat{\pi}}(x_i)} \leq 1, \quad \forall \theta \in \mathbb{R}. \tag{6}$$

Furthermore, averaging the LHS of (6) over $\widehat{\pi}$ and using the definition of the mixture density in (1), we have

$$\int \widehat{\pi}(d\theta) D_{\widehat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\int \widehat{\pi}(d\theta) p_\theta(x_i)}{p_{\widehat{\pi}}(x_i)} = 1.$$

We conclude that

$$\mathrm{supp}(\widehat{\pi}) \subset \{\text{Global maximizers of } D_{\widehat{\pi}}\}. \tag{7}$$

In particular, the number of atoms of $\widehat{\pi}$ is at most the number of critical points of $D_{\widehat{\pi}}$.

Example 1 (Poisson mixture)**.** As a concrete example, let us consider the Poisson model, where $p_\theta(x) = \frac{\theta^x}{x!}e^{-\theta}$ and $x \in \mathbb{Z}_+$. Thus

$$\frac{d}{d\theta} D_{\widehat{\pi}}(\theta) = e^{-\theta}\left(\sum_{i=1}^{n} w_i (x_i \theta^{x_i - 1} - \theta^{x_i})\right),$$

for some nonnegative weights $\{w_i\}$. Note that the quantity inside the parenthesis is a polynomial of θ of degree at most $x_{\max}$. Therefore, the number of critical points of $D_{\widehat{\pi}}(\theta)$ and hence the number of atoms of $\widehat{\pi}_{\mathrm{NPMLE}}$ are at most $x_{\max}$. This result is first observed[4] in [Sim76], which slightly improves the bound $x_{\max} + 1$ in Remark 2. For other models, the first-order condition typically does not reduce to a polynomial equation.

3 Exponential families

Following [Lin83a, Lin83b], we consider the following exponential family. Let p_0 be a base density (with respect to some dominating measure μ) on $\mathbb{R}$, whose moment generating function (MGF) and cumulant generating function is defined as

$$L(\theta) = \mathbb{E}_{X \sim p_0}[e^{\theta X}], \quad \kappa(\theta) = \log L(\theta), \tag{8}$$

[3] The condition (6) is also sufficient for the global optimality of $\widehat{\pi}$. Indeed, for any π, by the concavity of ℓ, Jensen's inequality implies $\ell(\widehat{\pi}) - \ell(\pi) \geq \frac{1}{\epsilon}[\ell(\widehat{\pi}) - \ell((1-\epsilon)\widehat{\pi} + \epsilon\pi)]$ all $\epsilon \in (0,1)$. Sending $\epsilon \to 0$ yields $\ell(\widehat{\pi}) - \ell(\pi) \geq -\frac{d}{d\epsilon}\ell((1-\epsilon)\widehat{\pi} + \epsilon\delta_\theta)\big|_{\epsilon=0} = 1 - \int \pi(d\theta) D_{\widehat{\pi}}(\theta) \geq 0$.

[4] The derivation here differs slightly with the original argument of [Sim76] which treats the system of $\{1, x, \ldots, x^k, e^x\}$.

and is assumed to be finite for all $\theta \in (\underline{\theta}, \overline{\theta})$, where $\underline{\theta}, \overline{\theta} \in [-\infty, \infty]$. Define the following exponential family of densities with natural parameter θ:

$$p_\theta(x) = \exp(\theta x - \kappa(\theta)) p_0(x).$$

Notable examples include:

- Gaussian location model $N(\theta, s)$: $p_0 = N(0,1)$, $L(\theta) = e^{\frac{\theta^2}{2s}}$ and $\kappa(\theta) = \frac{\theta^2}{2s}$.
- Poisson model $\mathrm{Poi}(e^\theta)$: $p_0 = \mathrm{Poi}(1)$, $L(\theta) = \exp\left(e^\theta - 1\right)$ and $\kappa(\theta) = e^\theta - 1$.

We need the following facts on the MGF:

Lemma 2. *1. $L(\theta) > 0$ for all $\theta \in \mathbb{R}$.*

2. κ is strictly convex and hence

$$\mu(\theta) \triangleq \kappa'(\theta) = \frac{L'(\theta)}{L(\theta)}$$

is strictly increasing in θ. Furthermore, if the distribution p_0 is fully supported on $\mathbb{R}$, then $\mu(\pm\infty) = \pm\infty$.

3. L has an analytic extension on the strip $\{z \in \mathbb{C} : \underline{\theta} < \Re(z) < \overline{\theta}\}$. Furthermore, for each disk $D(z_0, r)$ contained in this strip, with $z_0 = x_0 + iy_0$,

$$\sup_{z \in D(z_0, r)} |L(z)| \le \max\{L(x_0 - r), L(x_0 + r)\} \tag{9}$$

and

$$\sup_{z \in D(z_0, r)} |L'(z)| \le \inf_{\epsilon > 0} \frac{1}{\epsilon} \max\{L(x_0 - r - \epsilon), L(x_0 + r + \epsilon)\}. \tag{10}$$

Next we focus on continuous exponential families for which the NPMLE solution is known to be unique [LR93]. The following is a deterministic bound on the number of atoms of the NPMLE.

Theorem 3. *Fix $x_{\min} \le \min_{i \in [n]} x_i$ and $x_{\max} \ge \max_{i \in [n]} x_i$. Define $\theta_{\min} = \mu^{-1}(x_{\min})$, $\theta_{\max} = \mu^{-1}(x_{\max})$. Let $r = \frac{\theta_{\max} - \theta_{\min}}{2}$, $a = \frac{x_{\max} - x_{\min}}{2}$, and $x_0 = \frac{x_{\max} + x_{\min}}{2}$. Assume that $x_{\min} \le \mu(0) \le x_{\max}$. For each $\delta > 0$ such that $\delta < \frac{1}{5} \min\{\theta_{\min} - \underline{\theta}, \overline{\theta} - \theta_{\max}\}$,*

$$|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \le \frac{N_1}{\log \frac{2r + 2\delta}{2r + \delta}}$$

where

$$\begin{aligned}
N_1 &= 2(a + |\mu(0)| + |x_0|)(|\theta|_{\max} + 2\delta) + \kappa_{\max} + \log \frac{|x|_{\max} + \frac{1}{\delta}}{\tau} \\
\tau &= \max\{\mu(\theta_{\max} + \delta) - x_{\max}, x_{\min} - \mu(\theta_{\min} - \delta)\} \\
|\theta|_{\max} &= \max\{\theta_{\max}, -\theta_{\min}\} \\
|x|_{\max} &= \max\{x_{\max}, -x_{\min}\} \\
\kappa_{\max} &= \kappa(\theta_{\min} - 3\delta) \vee \kappa(\theta_{\max} + 3\delta)\,.
\end{aligned}$$

Remark 5. Roughly speaking, by choosing $\delta \asymp r$, Theorem 3 shows that $|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \lesssim |\theta|_{\max} a + \kappa_{\max}$.

Proof. Starting from (7), we bound the number of critical points of the following function

$$F(\theta) \triangleq \sum_{i=1}^{n} w_i \frac{p_\theta}{p_0}(x_i) = \sum_{i=1}^{n} w_i \exp(\theta x_i - \kappa(\theta)), \tag{11}$$

where $\sum_{i=1}^{n} w_i = 1$ and $w_i = c\frac{p_0(x_i)}{p_{\tilde{\pi}}(x_i)}$ and c is the normalization constant. Then

$$F'(\theta) = \sum_{i=1}^{n} w_i \exp(\theta x_i - \kappa(\theta))[x_i - \mu(\theta)],$$

Since $\mu = \kappa' = \frac{L'}{L}$ and $L(\theta)$ has no real roots (Lemma 2), we conclude that the critical points of $F(\theta)$ are the real roots of the following function:

$$G(\theta) \triangleq \sum_{i=1}^{n} w_i \exp(\theta x_i)[x_i L(\theta) - L'(\theta)]. \tag{12}$$

We first notice that all real roots of G should be on $[\theta_{\min}, \theta_{\max}]$. Indeed, by the strict monotonicity of μ, we have $G(\theta) > 0$ for $\theta > \theta_{\max}$ and $G(\theta) < 0$ for $\theta < \theta_{\min}$.

Next, let us extend definition (12) to a complex argument z and modify the function by introducing:

$$g(z) = G(z + \theta_0)e^{-(z+\theta_0)x_0} = \mathbb{E}[e^{(z+\theta_0)(Y-x_0)}(YL(z + \theta_0) - L'(z + \theta_0))], \qquad z \in \mathbb{C}$$

where $\theta_0 = \frac{\theta_{\min}+\theta_{\max}}{2}$ and $x_0 = \frac{x_{\min}+x_{\max}}{2}$, and $\mathbb{P}[Y = x_i] = w_i$. Note that the number of zeros of g in $z \in [-r, r]$ is the same as the total number of real zeros of G. We will overbound this quantity by counting all zeros of g in a disk of radius r on $\mathbb{C}$. To that end, we define $M_g(\rho) \triangleq \sup\{|g(z)| : |z| \le \rho\}$. We next fix $\delta_4 > \delta_3 > \delta_2 > 0$ such that $\theta_{\max} + \delta_4 < \overline{\theta}$ and $\theta_{\min} - \delta_4 > \underline{\theta}$. Set $r_2 = r + \delta_2$ and $r_1 = r + \delta_3$. On one hand, since $\mu(\theta_{\max}) = x_{\max}$ and $\mu(\theta_{\min}) = x_{\min}$, we have

$$M_g(r_2) \ge |g(r_2)| = |G(\theta_{\max}+\delta_2)|e^{-x_0(r_2+\theta_0)} \ge e^{-a(|\theta|_{\max}+\delta_2)-x_0(\theta_{\max}+\delta_2)}(\mu(\theta_{\max}+\delta_2)-x_{\max})\cdot L(\theta_{\max}+\delta_2)$$

and similarly

$$M_g(r_2) \ge |g(-r_2)| = |G(\theta_{\min}-\delta_2)|e^{-x_0(-r_2+\theta_0)} \ge e^{-a(|\theta|_{\max}+\delta_2)-x_0(\theta_{\min}-\delta_2)}(x_{\min}-\mu(\theta_{\min}-\delta_2))\cdot L(\theta_{\min}-\delta_2)$$

By the convexity of κ we have $\kappa(\theta_{\max} + \delta_2) \ge (\theta_{\max} + \delta_2)\mu(0)$ and $\kappa(\theta_{\min} - \delta_2) \ge (\theta_{\min} - \delta_2)\mu(0)$. Defining $\tau = \max(\mu(\theta_{\max} + \delta_2) - x_{\max}, x_{\min} - \mu(\theta_{\min} - \delta_2))$, we thus obtain

$$M_g(r_2) \ge e^{-(a+|\mu(0)|)(|\theta|_{\max}+\delta_2)-|x_0||\theta|_{\max}}\tau. \tag{13}$$

On the other hand, by Lemma 2 we have

$$\sup_{|z|\le r_1} |L'(\theta_0 + z)| \le \frac{1}{\delta_4 - \delta_3} \sup_{|z|\le r_1+\delta_4} |L(\theta_0 + z)| = \frac{1}{\delta_4 - \delta_3} L_4, \quad L_4 \triangleq L(\theta_{\min} - \delta_4) \vee L(\theta_{\max} + \delta_4)$$

Since $\sup_{|z|\le r_1} |L(z)| \le L(\theta_{\min} - \delta_3) \vee L(\theta_{\max} + \delta_3) \le L_4$ we conclude

$$M_g(r_1) \le e^{a(|\theta|_{\max}+\delta_3)} \left(|x|_{\max} + \frac{1}{\delta_4 - \delta_3}\right) L_4 \tag{14}$$

Now setting $\delta_4 = 3\delta, \delta_3 = 2\delta, \delta_2 = \delta$ we get

$$\log \frac{M_g(r_1)}{M_g(r_2)} \le N_1 \,.$$

The result then follows by the following lemma after also simplifying

$$\frac{r_1^2 + r_2 r}{r_1(r_2 + r)} \ge \frac{r_1 + r}{r_2 + r} = \frac{2r + 2\delta}{2r + \delta} \,.$$

□

Lemma 4. *Let f be a non-zero holomorphic function on a disk of radius r_1. Let $n_f(r) \triangleq |\{z : |z| \le r, f(z) = 0\}|$ and $M_f(r) \triangleq \sup_{|z|<r} |f(z)|$. For any $r < r_2 < r_1$ we have*

$$n_f(r) \le \frac{1}{\log \frac{r_1^2 + r_2 r}{r_1(r_2 + r)}} \log \frac{M_f(r_1)}{M_f(r_2)} \,.$$

This bound is achieved by $f(z) = \left(\frac{r - z}{1 - rz}\right)^n$.

Proof. Without loss of generality we assume $r_1 = 1$. If $M_f(1) = \infty$ then there is nothing to prove. Otherwise, the bound is equivalent to showing

$$M_f(r_2) \le M_f(1) C(r, r_2)^{-n_f(r)} \,, \qquad C(r, r_2) \triangleq \frac{1 + r_2 r}{r_2 + r} > 1 \tag{15}$$

which means that every zero inside rD reduces the magnitude of f on the boundary of $r_2 D$ by a factor C. To show this, let us denote by $\{a_i\}$ the list of $n = n_f(r)$ zeros of f inside rD (with multiplicity). Thus we can write

$$f(z) = g(z) \prod_{i=1}^{n} B_{a_i}(z) \,, \tag{16}$$

where $B_a(z) \triangleq \frac{|a|}{a} \frac{a - z}{1 - \bar{a} z}$ is the Blaschke factor, and $g(z)$ is holomorphic on D (and has no zeros in the closed disk of radius r, but this is not going to be used below). Let us show that for any $|z| \le r_2$ and $|a| < r_2$ we have

$$|B_a(z)| \le \frac{|a| + r_2}{1 + |a| r_2} \,. \tag{17}$$

Indeed, by the maximum principle it is sufficient to consider $z = r_2 e^{i\phi}, \phi \in [0, 2\pi)$ and by rotating the disk, we can also assume $a > 0$. Then

$$|B_a(r_2 e^{i\phi})|^2 = \frac{(a - r_2 \cos\phi)^2 + r_2^2 \sin^2 \phi}{(1 - a r_2 \cos\phi)^2 + a^2 r_2^2 \sin^2 \phi} = \frac{a^2 + r_2^2 - 2 a r_2 \cos\phi}{1 + a^2 r_2^2 - 2 a r_2 \cos\phi} \,. \tag{18}$$

Since $a^2 + r_2^2 < 1 + a^2 r_2^2$ we find that (18) is maximized at $\phi = \pi$, thus proving (17). Furthermore, from (18) applied with $r_2 = 1$ we also note that $|B_a(z)| = 1$ whenever $|z| = 1$, which via (16) implies $M_g(1) = M_f(1)$.

Finally, from (16)-(17) and the fact that $|g(z)| \le M_f(1)$ for all $z \in D$ we conclude that for any $|z| \le r_2$ we have

$$|f(z)| \le M_f(1) \prod_{i=1}^{n} \frac{|a_i| + r_2}{1 + |a_i| r_2} \,.$$

This concludes the proof of (15) after noticing that each factor is upper bounded by $\frac{1}{C(r, r_2)}$. □

As an application of Theorem 3, we now prove Theorem 1 for Gaussian location mixtures.

Proof. Choose $x_{\max} = \max_{i\in[n]} x_i$ and $x_{\min} = \min_{i\in[n]} x_i$. Recall that p_θ denote the density of $N(\theta, 1)$. In this model, we have $\underline{\theta} = -\infty$, $\overline{\theta} = \infty$, $\kappa(\theta) = \theta^2/2$ and $\mu(\theta) = \theta$. Thus $\theta_{\min} = x_{\min}$, $\theta_{\max} = x_{\max}$, $r = a = \frac{1}{2}(x_{\max} - x_{\min})$, $\tau = \delta$. Conveniently, note that for *location family*, we have the following translation invariance: Let $T_x(\pi)$ denote the pushforward of π under the translation $\cdot + x$. Then $\widehat{\pi}_{\mathrm{NPMLE}}(x_1 + x, \dots, x_n + x) = T_x(\widehat{\pi}_{\mathrm{NPMLE}}(x_1, \dots, x_n))$. Therefore, without loss of generality, we can assume $x_{\min} = -r \le 0 \le x_{\max} = r$, so that $x_0 = 0$ and $|x|_{\max} = r$.

Choosing $\delta = r$ yields (3). Finally, the high-probability statement follows from $\mathbb{P}[|x_i| \ge \tau] \le \exp(-c\tau^2)$ for some constant c and a union bound. □

The examples of Gaussian and Poisson models (Theorem 1 and Example 1) seem to suggest that NPMLE is always $O(\log n)$-atomic with high probability. Indeed, there is some degree of universality to this bound, as the following result shows. The extra condition we impose essentially says that the tail probability $\mathbb{P}_0\{|X| \ge a\}$ behaves as $\exp(-a^c)$ for some $c > 1$. For notational convenience, we will assume that the base measure p_0 is symmetric around zero.

Theorem 5. *Fix $2 < K_0 \le K_1$ and $\theta_0, b, \beta > 0$. Then there exist n_0, C depending on (K_0, K_1, β, b) with the following property. Consider any density p_0 symmetric around zero whose log-MGF satisfies*

$$K_0\kappa(\theta) \le \kappa(2\theta) \le K_1\kappa(\theta) \qquad \forall|\theta| > \theta_0\,. \tag{19}$$

Let $x_1, \dots, x_n \overset{i.i.d.}{\sim} p_\pi$ for some mixing distribution π supported on the interval $[-b, b]$. Then for all $n \ge n_0$, with probability $1 - 2n^{-\beta}$, $\widehat{\pi}_{\mathrm{NPMLE}}$ has at most $C \log n$ atoms.

Remark 6. Theorem 5 shows that the Gaussian tail is not essential for the $O(\log n)$ result to hold. In fact, consider any smooth density p_0 such that $-\log p_0(x) \asymp |x|^\alpha$ for $\alpha > 1$. Then by saddle-point approximation we have $\kappa(\theta) \asymp \theta^{\alpha/(\alpha-1)}$ as $\theta \to \infty$, which satisfies (19).

On the other hand, compactly supported families are excluded since for those distributions $\kappa(\theta)$ is asymptotically linear (with a slope given by the essential supremum of p_0) as $\theta \to \infty$. Furthermore, exponential tails are also excluded. This is directly related to the open problem with mixtures of exponential distributions which will be discussed in Section 5.2.

Proof of Theorem 5. We start by establishing properties of $\kappa(\cdot)$ and $\mu(\cdot)$ implied by conditions of the theorem. Under the symmetry assumption of p_0, $\kappa(\theta)$ is an even convex function with $\kappa(\theta) \ge \kappa(0) = 0$ and $\mu(0) = 0$. From the convexity of κ we get for any $\theta > 0$

$$\kappa(\theta) \le \kappa(\theta/2) + \frac{\theta}{2}\mu(\theta)$$
$$\kappa(2\theta) \ge \kappa(\theta) + \theta\mu(\theta)$$

And thus, for $\theta > \theta_0$ we get

$$\mu(\theta)\theta \ge C_0\kappa(\theta), \quad C_0 = 2\frac{K_0 - 1}{K_0} > 1 \tag{20}$$
$$\mu(\theta)\theta \le C_1\kappa(\theta), \quad C_1 = K_1 - 1 > 0 \tag{21}$$

Clearly, also, $\mu(\theta) \to \infty$ as $\theta \to \infty$ and hence p_0 is supported on the whole of $\mathbb{R}$. Thus we have $\underline{\theta} = -\infty$ and $\overline{\theta} = \infty$.

Define the rate function for $a > 0$:

$$E(a) \triangleq \sup_{\theta>0} a\theta - \kappa(\theta),$$

which is achieved at $\theta = \rho \triangleq \mu^{-1}(a)$, so that $E(a) = a\rho - \kappa(\rho)$. From (20)-(21) (noting $C_0 > 1$) we conclude that as $a \to \infty$ (and hence $\rho \to \infty$) we get:

$$E(a) \asymp a\rho \asymp \kappa(\rho) \tag{22}$$

We need to establish one more consequence of (19). Namely, there exists $\theta_0' > 0$ and $m_0 \in \mathbb{N}$ such that for any $\theta_1 > \theta_0'$ there exists $\theta^* \in [\theta_1, 2^{m_0-1}\theta_1]$ such that

$$\mu(2\theta^*) - \mu(\theta^*) > 1\,. \tag{23}$$

To show this, we select $m_0 > \log_2 \frac{4(K_0-1)}{K_0-2}$ and $\theta_0' \geq \theta_0$ so large that $\mu(\theta_0') \geq \frac{4m_0}{K_0-2}$. Now suppose (for the sake of contradiction) that for all $\theta \in [\theta_1, 2^{m_0-1}\theta_1]$ we have

$$\mu(2\theta) - \mu(\theta) \leq 1\,.$$

Denoting $\theta_2 \triangleq 2^{m_0}\theta_1$, applying the above inequality repeatedly yields $\mu(\theta_2) \leq \mu(\theta_1) + m_0$. Consequently, from the convexity of κ we have

$$\kappa(\theta_2) \leq \kappa(\theta_1) + (2^{m_0} - 1)\theta_1(\mu(\theta_1) + m_0)\,.$$

On the other hand,

$$\kappa(\theta_2/2) \geq \kappa(\theta_1) + (2^{m_0-1} - \theta_1)\mu(\theta_1)\,.$$

Taking the ratio of these, we get from (19):

$$\kappa(\theta_1) + (2^{m_0} - 1)\theta_1(\mu(\theta_1) + m_0) \geq K_0\left(\kappa(\theta_1) + (2^{m_0-1} - 1)\theta_1\mu(\theta_1)\right)\,.$$

Rearranging terms we arrive at

$$2^{m_0}\theta_1\left(\left(\frac{K_0}{2} - 1\right)\mu(\theta_1) - m_0\right) \leq (\theta_1\mu(\theta_1) - \kappa(\theta_1))(K_0 - 1) - \theta_1 m_0\,.$$

Dropping all negative terms on the right, and noticing that by the choice of θ_0' we have $(\frac{K_0}{2} - 1)\mu(\theta_1) - m_0 \geq \frac{1}{2}(\frac{K_0}{2} - 1)\mu(\theta_1)$, we conclude

$$2^{m_0}\frac{1}{2}\left(\frac{K_0}{2} - 1\right)\theta_1\mu(\theta_1) \leq (K_0 - 1)\theta_1\mu(\theta_1)\,.$$

By the choice of m_0, however, this is impossible. Hence, there must exist θ^* satisfying (23).

Having established (22) and (23) we proceed to the proof of the theorem. Let $X \sim p_\pi$. By the Chernoff bound, for any $\theta > 0$, $\mathbb{P}[X \geq a] \leq e^{-\theta a}\mathbb{E}[e^{\theta X}]$. Here

$$\mathbb{E}[e^{\theta X}] = \int \pi(d\theta') \int dx p_0(x) e^{(\theta'+\theta)x - \kappa(\theta')} = \int \pi(d\theta') \frac{L(\theta' + \theta)}{L(\theta')} \leq L(\theta + b),$$

where the last inequality follows from the fact that L is an even function lower bounded by $L(0) = 1$, and $L(\theta' + \theta) \leq L(\theta + b)$ for any $\theta > 0$ and $\theta' \in [-b, b]$. Optimizing over θ we set $\theta = \rho - b$ and obtain $\mathbb{P}[X \geq a] \leq e^{ab - E(a)}$, provided that $\rho > b$.

Since we aim to apply Theorem 3, we need to choose $x_{\max}$ and $x_{\min}$. We set them as follows. First we set θ_1 so that $a_1 = \mu(\theta_1)$ verifies $E(a_1) - a_1 b = (1 + \beta)\log n$. Note that as $n \to \infty$ we have $a_1, \theta_1 \to \infty$. In the sequel, we assume that n is so large that $\theta_1 > \theta_0'$ and $\theta_1 > b$. Notice that from (22) we have $E(a_1) \gg a_1$ and hence

$$E(a_1) \asymp \theta_1 a_1 \asymp \kappa(\theta_1) \asymp \log n\,. \tag{24}$$

Next, having selected θ_1 we use (23) to select $\theta_{\max} = \theta^*$. We set $x_{\max} = \mu(\theta_{\max})$ and $x_{\min} = -x_{\max}, \theta_{\min} = -\theta_{\max}$, so that $x_0 = (x_{\min}+x_{\max})/2 = 0$. Then we have $\mathbb{P}[X \geq x_{\max}] \leq \mathbb{P}[X \geq a_1] \leq n^{-(1+\beta)}$. Similarly, $\mathbb{P}[X \leq -x_{\min}] \leq n^{-(1+\beta)}$. By the union bound this implies that with probability at least $1 - 2n^{-\beta}$, we have $x_{\min} \leq \min_i x_i \leq \max_i x_i \leq x_{\max}$. Now we apply Theorem 3 with $\delta = 2\theta_{\max}$, obtaining

$$|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \lesssim \theta_{\max} x_{\max} + \kappa(4\theta_{\max}) + \log \frac{x_{\max} + \frac{1}{2\theta_{\max}}}{\tau}, \tag{25}$$

where $\tau = \mu(3\theta_{\max}) - \mu(\theta_{\max}) \geq \mu(2\theta_{\max}) - \mu(\theta_{\max}) > 1$ by (23). Consequently, the last term in (25) is dominated by the first.

Finally, we note that $\theta_{\max} \in [\theta_1, 2^{m_0-1}\theta_1]$ and thus $\theta_{\max} \asymp \theta_1$. From (19) we have $\kappa(\theta_{\max}) \asymp \kappa(\theta_1) \asymp \log n$. Similarly, from (22) we get $\theta_{\max} x_{\max} \asymp \kappa(\theta_{\max}) \asymp \log n$. In all, the right-hand side of (25) is $\asymp \log n$ as claimed. □

4 Statistical consequences on NPMLE

In this section we show how the self-regularization property of the NPMLE allows one to "bootstrap" existing results on MLE in finite Gaussian models to infinite mixtures.

The following statistical guarantee on NPMLE is due to Zhang [Zha09], improving over previous result of [GvdV01, GvdV07].

Theorem 6. *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_\pi \triangleq \pi * \varphi$ and let $\widehat{\pi} = \widehat{\pi}_{\mathrm{NPMLE}}(X_1, \ldots, X_n)$ be given in (2). Then*

$$\sup_{\pi \in \mathcal{M}_{\mathsf{SG}}(1)} \mathbb{E}_\pi[H^2(p_{\widehat{\pi}}, p_\pi)] \lesssim \frac{\log^2 n}{n}, \tag{26}$$

where $\mathcal{M}_{\mathsf{SG}}(s)$ denote the collection of all s-subgaussian distributions on $\mathbb{R}$.

Next we show that using the self-regularization of the NPMLE, Theorem 6 can be deduced from existing guarantees on MLE in finite mixture models. We need a couple of auxiliary results, whose proofs are deferred to the end of this section. The following result is on approximating a general Gaussian mixture by finite mixtures:

Lemma 7. *Let π be 1-subgaussian. For any $a > 0$ and any $k \in \mathbb{N}$, there exists a k-atomic π' supported on $[-a, a]$, such that*

$$\mathrm{TV}(p_\pi, p_{\pi'}) \leq 2e^{-a^2/2} + 2e^{a^2/4} \left(\frac{ea^2}{2k}\right)^k$$

Next we recall the statistical guarantee on the parametric MLE in finite Gaussian mixtures. By standard results on MLE (cf. e.g. [vdG00]), this can be deduced from the bracketing entropy for this class, which has been thoroughly investigated in the literature [GvdV01, GW00, Zha09, MM11]. The following result is a corollary of the entropy bound of Maugis and Michel in [MM11].

Lemma 8. *Let $a \geq 1$ and $k \in \mathbb{N}$. Let $\widehat{\pi}_{k,a} = \widehat{\pi}_{k,a}(Y_1, \ldots, Y_n)$ is the (parametric) MLE defined in (28), where $Y_i \overset{i.i.d.}{\sim} p_\pi$. There exists a universal constant C such that*

$$\sup_{\pi \in \mathcal{M}_{k,a}} \mathbb{E}_\pi[H^2(p_{\widehat{\pi}_{k,a}}, p_\pi)] \leq \frac{Ck}{n} \log \frac{na^2}{k}, \tag{27}$$

where $\mathcal{M}_{k,a}$ denotes the collection of all k-atomic distributions on $[-a, a]$.

Proof of Theorem 6. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\pi = \pi * N(0,1)$ for some 1-subgaussian π. Define the event $E_0 \triangleq \{|X_{\max}| \le a_0\}$, where $a_0 = \sqrt{C_0 \log n}$ for some large absolute constant C_0. Then E_0 has probability at least $1 - n^{-2}$. By Theorem 1, on the event E_0, $\widehat{\pi}$ is supported on $[-a_0, a_0]$ and $|\mathrm{supp}(\widehat{\pi})| \le C_1 a_0^2 = C_1 C_0 \log n \triangleq k_0$. Then for any $k \ge k_0$ and $a \ge a_0$, on the event E_0, we have

$$\widehat{\pi} = \widehat{\pi}_{k,a}(X_1, \dots, X_n) \triangleq \operatorname*{argmax}_{\pi \in \mathcal{M}_{k,a}} \sum_{i=1}^n \log p_\pi(X_i). \tag{28}$$

(In case that (28) has multiple maximizers, $\widehat{\pi}$ is chosen to be any one of them.)

Pick $a = \sqrt{C_1 \log n}$ and $k = C_2 \log n$ such that $a \ge a_0$, $k \ge k_0$, and $a^2/k \le 1/10$. Applying Lemma 7 with this choice, we obtain a k-atomic distribution π' supported on $[-a, a]$ such that $\mathrm{TV}(p_\pi, p_{\pi'}) \le n^{-3}$. Let $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} p_\pi = \pi * N(0,1)$ for some 1-subgaussian π. Then $\mathrm{TV}(\mathrm{Law}(X_1, \dots, X_n), \mathrm{Law}(Y_1, \dots, Y_n)) \le n^{-2}$. Then there exists a coupling such that $X_i = Y_i$ for $i = 1, \dots, n$ with probability at least $1 - n^{-2}$. Let E_2 denote this event.

On the event of $E_1 \cap E_2$, we have

$$\widehat{\pi} = \widehat{\pi}_{\mathrm{NPMLE}}(X_1, \dots, X_n) = \widehat{\pi}_{k,a}(X_1, \dots, X_n) = \widehat{\pi}_{k,a}(Y_1, \dots, Y_n).$$

Now we are in a position to pass the statistical guarantee on the parametric MLE in finite Gaussian mixtures to the NPMLE. Applying Lemma 8 with $k \asymp \log n$ and $a \asymp \sqrt{\log n}$, we have

$$\mathbb{E}[H^2(p_{\widehat{\pi}_{k,a}}, p_{\pi'})] \lesssim \frac{\log^2 n}{n}. \tag{29}$$

Finally, using the fact that $H^2 \le \mathrm{TV}$ and the triangle inequality for Hellinger, we have

$$\mathbb{E}[H^2(p_{\widehat{\pi}}, p_\pi) \mathbf{1}_{\{E_0 \cap E_1\}}] \le 2\mathbb{E}[H^2(p_{\widehat{\pi}_{k,a}(Y_1,\dots,Y_n)}, p_{\pi'})] + 2H^2(p_\pi, p_{\pi'}) \le C_3 \frac{\log^2 n}{n}.$$

The proof is completed since $H^2 \le 2$ and $E_0 \cap E_1$ has probability at least $1 - 2n^{-2}$. □

Remark 7. The following minimax lower bound is shown in [Kim14]:

$$\inf_{\widehat{p}} \sup_{\pi \in \mathcal{M}_{\mathsf{SG}}(1)} \mathbb{E}_\pi[H^2(\widehat{p}, p_\pi)] \gtrsim \frac{\log n}{n}, \tag{30}$$

which differs from the upper bound in Theorem 6 by $\log n$. As frequently observed in the density estimation literature, such a logarithmic factor can be attributed to the fact that the analysis of the MLE is based on the global entropy bound. Thus obtaining a local version of the entropy bound in [MM11] can potentially close this gap and establish the sharp optimality of the NPMLE in achieving the lower bound in (30).

Proof of Lemma 7. Without loss of generality, assume that π has zero mean. Let $\widetilde{\pi}$ denote the conditional version of π on $[-a, a]$. By the data processing inequality of total variation,

$$\mathrm{TV}(\pi * N(0,1), \widetilde{\pi} * N(0,1)) \le \mathrm{TV}(\pi, \widetilde{\pi}) = \pi([-a,a]^c) \le 2e^{-a^2/2}$$

where the last inequality follows from π being 1-subgaussian. Next, let π' denote the k-point Gauss quadrature of $\widetilde{\pi}$, such that π' and $\widetilde{\pi}$ have identical first $2k-1$ moments, and π' is also supported on $[-a, a]$. Then by moment-matching approximation (see [WY20, Lemma 8]), we have

$$\chi^2(\pi' * N(0,1) \| \widetilde{\pi} * N(0,1)) \le 4e^{a^2/2} \left(\frac{ea^2}{2k}\right)^{2k}.$$

Using the fact that $2\mathrm{TV}^2 \le \chi^2$ and the triangle inequality, the previous two displays yield the desired bound. □

Proof of Lemma 8. Let $N_{[]}(\epsilon)$ denote the bracketing number of the class of k-GM densities $\mathcal{P}_{k,a} \triangleq \{p_\pi : \pi \in \mathcal{M}_{k,a}\}$ with respect to the Hellinger distance. Applying Eq. (B.8) in [MM11, Proposition B.4] (with $\alpha = Q = 1$, $D(k,\alpha) = 3k$, $\lambda_m = \lambda_M = 1$, so that $\mathcal{I} \asymp K \log a$), we have

$$\log N_{[]}(\epsilon) \lesssim k \log \frac{a}{\epsilon}, \tag{31}$$

Next we can apply standard results on the density estimation guarantee (in Hellinger distance) for the MLE (see e.g. [vdG00, Theorem 7.4]). Define $J(\epsilon) \triangleq \int_{\epsilon^2}^{\epsilon} \sqrt{\log N_{[]}(u)} du$. By (31), we have $J(\epsilon) \lesssim \epsilon \sqrt{k \log \frac{a}{\epsilon}}$. Thus $\mathbb{E}[H^2(p_{\widehat{\pi}_{k,a}}, p_\pi)] \lesssim \epsilon_n^2$, where $\sqrt{n}\epsilon_n^2 = J(\epsilon_n)$ so that $\epsilon_n \asymp \sqrt{\frac{k}{n} \log \frac{na^2}{k}}$. □

5 Discussions

5.1 Statistical degree

In this subsection we discuss the concept of self-regularization. Loosely speaking, an unregularized estimator can be said to achieve some form of self-regularization if it returns a density with $o(n)$ components, which improves over the worst-case upper bound of n. Expanding on the reasoning in Remark 1, below we introduce a formal framework and provide a perspective on what may be the correct model size.

Consider a sequence of nested statistical models $M_1 \subset M_2 \subset \cdots M \subset \mathcal{M}(\mathcal{X})$, where k is a parameter that encodes the "model complexity" of M_k. For example, in linear models, M_k denotes those with k-sparse regression coefficients; in shape-constrained setting, M_k can be the set of k-piecewise constant or log-affine densities; in our setting of mixture models, M_k is the set of all k-GM densities.

Given a sample of size n, we define the statistical degree K_n as

$$K_n \triangleq \inf \left\{ k : d_{\max}(M, M_k) \le \frac{1}{3\sqrt{n}} \right\}, \tag{32}$$

where $d_{\max}(A, B) \triangleq \sup_{P \in A} \inf_{Q \in B} H(P, Q)$ denotes the best approximation error (in the Hellinger distance) of the model class A by members of B. By definition, K_n is the largest k so that any density in M can be made statistically indistinguishable (on the basis of n observations) from some density in M_k; in other words, given a sample of size n drawn independently from any $f \in M$, one can simulate it with probability at least $1 - c$ for some constant c using one drawn from some $f_k \in M_k$. From this simulation perspective, there is no statistical reason to fit a model of complexity bigger than K_n; on the other hand, it does not compromise the statistical performance (in terms of the Hellinger rate) to restrict to models of complexity at most K_n. Thus, we view achieving the statistical degree K_n as a criterion of self-regularization. As shown in Remark 1, for the class M of Gaussian mixtures with subgaussian mixing distributions, we have $K_n = \Theta(\log n)$, which coincides with the typical model size fitted by the NPMLE.

Next, we discuss a simple example where the self-regularization of the unpenalized NPMLE can be established directly.

Example 2. Consider observations taking non-negative integer values in $\mathcal{X} = \mathbb{Z}_+$. For each $k \ge 1$, let M_k denote the set of distributions supported on $\{0, \dots, k\}$, and let M the class of 1-subgaussian distributions on $\mathbb{Z}_+$. It is clear that the statistical degree in this case is $K_n = \Theta(\sqrt{\log n})$. Indeed, the upper bound follows from truncation and the uniform subgaussian tail, and the lower bound follows from considering an explicit distribution such as $P(j) \propto e^{-j^2}$.

Given $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} P \in M$, the NPMLE for P (without enforcing the subgaussianity) is simply the empirical distribution $\widehat{P}$, where

$$\widehat{P}(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i = j\}}, \quad j \in \mathbb{Z}_+. \tag{33}$$

By a union bound, there exists a constant C, such that with probability $1 - o(1)$, $\widehat{P}(j) = 0$ for all $j \geq k = C\sqrt{\log n}$. In other words, with high probability we automatically have $\widehat{P} \in M_k$ for some k that agrees with the statistical degree.

Note that the self-regularizing property in Example 2 is a simple consequence of the explicit expression of the NPMLE in (33). In contrast, for mixture models in Theorems 1 and 3 we need to resort to the optimality condition and complex-analytic techniques, due to the lack of close-form expression of NPMLE in mixture models. Another major difference is that for mixture models M_k is non-convex and hence optimizing the likelihood over M_k can be expensive. Quite spectacularly, the full relaxation over all measures somehow automatically solves the nonconvex optimization (and for the right k).

5.2 Self-regularization for mixtures of exponentials

Although we have not identified an example of a mixture model where the number of atoms of NPMLE is $\omega(\log n)$, the program of analyzing the NPMLE in Theorem 3 and Theorem 5 does have its limitations. As a leading example, let us consider mixtures of exponential distributions, which is among the earliest results on the structure of NPMLE [Jew82] (see also [GW92, Sec. 2.1]). Since the tail is exponential, this model is outside the scope of Theorem 5.

Example 3 (Exponential mixture)**.** Consider the exponential distribution $\mathrm{Exp}(\theta)$ with density $p_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{\{x>0\}}$ and $\theta > 0$. In this case, the NPMLE is defined as

$$\widehat{\pi}_{\mathrm{NPMLE}} = \arg \max_{\pi \in \mathcal{M}(\mathbb{R}_+)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i), \quad p_\pi(x) = \int \theta e^{-\theta x} \pi(d\theta). \tag{34}$$

Upon normalization, the gradient (6) is proportional to the function

$$F(\theta) = \sum_{i=1}^{n} w_i \theta e^{-\theta x_i}, \tag{35}$$

where $\sum_{i=1}^n w_i = 1$ and $w_i \geq 0$. Thus the atoms of the NPMLE are roots of $F'(\theta) = \sum_{i=1}^n w_i e^{-\theta x_i}(1 - \theta x_i)$, which are localized in the interval $[a, b]$ with $a = 1/x_{\max}$ and $b = 1/x_{\min}$. Following the proof of Theorem 3, to bound the number of roots of F', we can apply Lemma 4 to $f(\theta) = F'(\theta - \frac{a+b}{2})$ and $r = \frac{b-a}{2}$. Choose $r_2 = \frac{a+b}{2}$ and $r_1 = 2r = b - a$. Since $f(r_2) = F'(0) = 1$, we have $M_f(r_2) \geq 1$. Moreover, it is clear that $M_f(r_1) \leq \exp(Cbx_{\max})$ for some constant C. Thus an application of Lemma 4 shows that

$$|\mathrm{supp}(\widehat{\pi}_{\mathrm{NPMLE}})| \lesssim \frac{x_{\max}}{x_{\min}}. \tag{36}$$

However, in the stochastic setting the above bound is too loose to be useful. Indeed, suppose $x_1, \dots, x_n$ are drawn independently from a single exponential distribution, say, $\mathrm{Exp}(1)$. Then with high probability, we have $x_{\min} = \Theta_P(\frac{1}{n})$ and $x_{\max} = \Theta_P(\log n)$. Thus (36) yields $|\widehat{\pi}_{\mathrm{NPMLE}}| = O(n \log n)$, which is even worse than the deterministic bound of $|\widehat{\pi}_{\mathrm{NPMLE}}| \leq n$. Clearly, the culprit

of this looseness stems from the fact that the data-generating distribution is supported on $\mathbb{R}_+$ which has a boundary at zero. Since the smallest observation will be on the order of $\frac{1}{n}$, *a priori* one can only localize the atoms of the NPMLE in an interval of width $\Theta(n)$, which is much worse than $\Theta(\sqrt{\log n})$ in the Gaussian model. Similar problems also arise in other distributions whose support has boundary points, such as Gamma or Beta families.

Open question: Given $X_i \overset{\text{i.i.d.}}{\sim} p_\pi$ where $\operatorname{supp}(\pi) \subset [1,2]$, prove that with probability $1-o(1)$, we have

$$|\operatorname{supp}(\widehat{\pi}_{\text{NPMLE}})| = \Theta(\log n) \tag{37}$$

The crucial $O(\log n)$ upper bound would follow from the following analytic *conjecture:* For any distribution π on $[-a,a]$ the convolution $(\pi * h)(x) \triangleq \int h(x-y)\pi(dy)$ has at most $O(a)$ critical points, where $h(x) = e^{-e^x+x}$ is the density of a Gompertz distribution.

5.3 Compactly supported NPMLE

So far we have focused on unconstrained NPMLE, where the likelihood is maximized over all mixing distributions. In case where one has extra knowledge such as compact support, moment constraint, or sparsity, these information can be incorporated into the optimization problem as linear constraints leading to potentially improved statistical performance. This begs the question: to what extent does constraint help the self-regularization of the NPMLE. Specifically,

1. If the unconstrained solution fails to self-regularize, does adding constraints make it so?
2. If the unconstrained solution is already self-regularizing, does adding constraints make it more so?

We briefly discuss these two aspects below.

For the first problem, let us continue Example 3 on exponential mixtures, where we pointed out that the program in Theorem 3 does not resolve the self-regularization of unconstrained NPMLE. Nevertheless, it is easy to show that adding a support constraint to NPMLE does resolve conjecture (37). Indeed, suppose that the parameter θ is bounded from above by some constant θ_0, in which case one can consider the following support-constrained version of (34):

$$\widehat{\pi}'_{\text{NPMLE}} = \arg\max_{\pi \in \mathcal{M}([0,\theta_0])} \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i). \tag{38}$$

Thanks to the constraint, we only need to count the number of critical points of (35) in the interval $[0,\theta_0]$. Applying the same argument in Example 3 now with $r = \theta_0/2 = O(1)$ yields $|\widehat{\pi}'_{\text{NPMLE}}| \lesssim x_{\max} = O_P(\log n)$. Note that using moment matching and Taylor expansion we can show that the statistical degree for exponential mixtures with parameters bounded away from zero and infinity (say, $\operatorname{supp}(\pi) \subset [1,2]$) is $O(\log n)$. We *conjecture* that the statistical degree K_n in this case is $\Theta(\log n)$ and if so, the argument above shows that $\widehat{\pi}'_{\text{NPMLE}}$ does self-regularize.

For the second problem, let us revisit the Gaussian location mixture. Suppose the mixing distribution is supported on a compact interval, say, $[-1,1]$. Theorem 1 shows that the unconstrained NPMLE is $O(\log n)$-atomic with high probability. However, when the mixing distribution is compactly supported, the moment-matching argument in [WY20, Lemma 8] shows that the statistical degree in fact reduces to $O(\frac{\log n}{\log\log n})$. Again, we *conjecture* that in this case $K_n \asymp \frac{\log n}{\log\log n}$. Then a natural question is whether NPMLE with support constraint $\widehat{\pi}'_{\text{NPMLE}}$ defined as in (38)

with maximization over $\{\pi : \mathrm{supp}(\pi) \in [-1,1]\}$ achieves a better self-regularization of $O(\frac{\log n}{\log\log n})$ atoms.[5]

The main bottleneck of proving this is the following. Note that similar to the proof of Theorem 3 we can reduce to the problem of counting the critical point of (11), which for Gaussian model simplifies to

$$F(\theta) = \sum_{i=1}^{n} w_i \varphi(\theta - x_i), \quad w_i \propto \frac{1}{(\widehat{\pi}'_{\mathrm{NPMLE}} * \varphi)(x_i)}. \tag{39}$$

However this time we are not interested in bounding the number of all critical points of F, but only those in $[-1,1]$. Thus, we can set $r = 1$, $r_1 \asymp \sqrt{\log n}$ in the application of Lemma 4. The issue is with setting r_2. If we could show that F must have at least one point z_0 inside a disk of radius $O(1)$ such that

$$|F'(z_0)| > n^{-C} \tag{40}$$

for some C (with high probability), then invoking Lemma 4 with $r_2 = O(1)$ would conclude that F' has at most $O(\frac{\log n}{\log\log n})$ roots inside the unit disk. It is tempting to conjecture further that z_0 satisfying (40) exists for arbitrary choice of $\{w_i, x_i\}_{i=1}^n$, s.t. $|x_i| \lesssim \sqrt{\log n}$. Alas, this stronger conjecture does not hold as [PW20, Section 2] constructs $\{w_i, x_i\}_{i=1}^{O(\log n)}$ such that $|F'(z)| \le n^{-C\log\log n}$ for all $|z| = O(1)$. Therefore unlike the proof of Theorem 1, here we cannot ignore the stochastic origin of x_i and that w_i is inversely proportional to the fitted likelihood at x_i (see (39)). Since $\widehat{\pi}'_{\mathrm{NPMLE}}$ itself is random, proving this property of $G(z)$ seems to require a delicate analysis of "small-ball" probabilities of the empirical process. This is left for future work.

5.4 Maxima of Gaussian mixtures

In the special case of the Gaussian location mixture, Theorem 3 translates to the following statement on the Gaussian convolution: For any distribution π supported on the interval $[-a,a]$, the convolution $\pi * \varphi$ has at most $O(a^2)$ critical points. This result has been shown independently in the recent work [DYPS20, Theorem 6] by similar techniques using a corollary of Jensen's formula from [Tij71]. Can this bound be improved? The answer is negative and we next give a simple construction of a Gaussian mixture with $\Omega(a^2)$ local maxima.[6]

Lemma 9. *Let h be a continuous probability density on $\mathbb{R}$ with characteristic function $\widehat{h}$ and CDF H. Suppose we have ω_0 and $a > 0$ such that*

$$|\widehat{h}(\omega_0)| > 2(H(-a/2) + 1 - H(a/2)). \tag{41}$$

Let $\pi(x) = c(1+\sin(\omega_0 x))1\{|x| \le a\}$ with $c > 0$ chosen to make π a probability density. Then $h * \pi$ has at least $\frac{\omega_0 a}{2\pi}$ local maxima on $[-a/2, a/2]$.

Proof. Let $\pi_0(x) = 1+\sin(\omega_0 x)$. Then $0 \le \pi_0 \le 0$. Then $(\pi_0 * h)(x) = |\widehat{h}(\omega_0)| \sin(\omega_0 x - \arg\widehat{h}(\omega_0)) + 1$, which is a shifted and scaled sinusoid. Let S_+ and S_- be the sets of global maxima and minima of $\pi_0 * h$ (which are lattices with step $\frac{2\pi}{\omega_0}$). Let $\pi_1(x) = \pi_0(x)1\{|x| \le a\}$. Define $\Delta \triangleq h * \pi_0 - h * \pi_1$. Then $\Delta \ge 0$ everywhere. Furthermore,

$$\Delta(x) = \int_{|y|>a} \pi_0(y) dh(x-y) \le 2(H(x-a) + 1 - H(x+a)). \tag{42}$$

[5] It would be even more spectacular if the unconstrained NPMLE achieved the same number of atoms, but we are not willing to conjecture this.

[6] A different construction using $\Omega(a^2)$ equally weighted and equally spaced Gaussians is given in the independent work [KK20] in response to a conjecture of authors of [DYPS20], cf. *arxiv:1901.03264v4*, that their bound can be improved to $O(A)$.

Thus, by assumption (41), for any $|x| \le a/2$ we have $\Delta(x) \le |\widehat{h}(\omega_0)|$. Consequently, for any $x \in S_+ \cap [-a/2, a/2]$ we have $(\pi_1 * h)(x) = (\pi_0 * h)(x) - \Delta(x) > 1$ and for any $x \in S_- \cap [-a/2, a/2]$ we have $(\pi_1 * h)(x) < 1 - |\widehat{h}(\omega_0)|$. Thus, the level $1 - \frac{1}{2}|\widehat{h}(0)|$ must be crossed in between any two consecutive points from S_+ and S_-, implying the statement. □

Corollary 10. *There exists a compactly supported density π on $[-a, a]$ so that $\pi * \varphi$ has $\Omega(a^2)$ local maxima on $[-a/2, a/2]$.*

Proof. By the Gaussian tail bound, we have $H(-a/2) + 1 - H(a/2) \le 2e^{-a^2/8}$. Choosing $\omega_0 = a/4$, the claim follows from Lemma 9 for sufficiently large a. □

5.5 Mixture of log-concave densities

Consider the following question: Given a convex combination of k unimodal densities, how many modes can it have? A moment of thought shows that the answer is trivial as the sum of two unimodal densities, e.g. $f(x) + f(x-1)$ with

$$f(x) = (1 - |x|)\mathbf{1}_{\{|x| \le 1\}}, \tag{43}$$

can have infinitely many modes; the same example also applies even if unimodality is replaced by log-concavity. A natural question is what happens to strongly log-concave densities.[7] By replacing (43) with

$$f(x) = \begin{cases} 0, & |x| > 1, \\ 1 - |x|, & \epsilon < |x| \le 1 \\ -x^2/(2\epsilon) + 1 - \epsilon/2, & |x| \le \epsilon \end{cases}$$

which is strongly log-concave, we again see that $f(x) + f(x-1)$ can have a flat piece. Furthermore, it is possible to construct infinitely differentiable f (by convolving with a mollifier) with the same property; however, such a density is not analytic. Thus, we ask the question:

> *Given a convex combination of k analytic densities that are strongly log-concave, how many modes can it have?*

The following result gives an $\Omega(k^2)$ lower bound. Whether this is tight is an open question.

Corollary 11. *There exist strongly log-concave analytic densities $f_1, \dots, f_k$ on $\mathbb{R}$ and weights $\alpha_1, \dots, \alpha_k$ such that $\alpha_1 f_1 + \dots + \alpha_k f_k$ has $\Omega(k^2)$ local maxima.*

Proof. Take the π supported on $[-a, a]$ from Corollary 10. Partition $[-a, a]$ into $k = 4a$ consecutive intervals $I_1, \dots, I_k$ of length $1/4$. Let π_i denote the conditional version of π on I_i and set $\alpha_i = \pi(I_i)$. Recall the fact that $(\log(\mu * \varphi))'' \ge 1 - b^2$ for any probability measure μ supported on an interval of length $2b$; this follows from the well-known identity $(\log(\mu * \varphi))''(y) = 1 - \mathrm{Var}(X | X + Z = y)$, where $X \sim \mu$ and $Z \sim N(0,1)$ are independent. Then $f_i \triangleq \pi_i * \varphi$ is strongly log-concave satisfying $(\log f_i)'' \ge 3/4$. Since $\pi * \varphi = \sum_{i=1}^k \alpha_i f_i$, the desired conclusion then follows from Corollary 10. □

5.6 Further open problems

In addition to those on exponential mixtures and constrained NPMLE mentioned in Section 5.3 and Section 5.2, we end the paper by describing some further open problems on the structure of NPMLE:

[7] Recall that (cf. [SW14, Definition 2.9]) a density f is called c-strongly log-concave strongly convex if $\log f$ is strongly concave, i.e., $\log f((1-\alpha)x + \alpha y) \ge (1-\alpha)\log f(x) + \alpha \log f(y) + \frac{c}{2}\alpha(1-\alpha)\|x - y\|_2^2$ for all x, y and all $\alpha \in [0,1]$ for some constant $c > 0$ In the case of twice-differentiable h, this is equivalent to $\nabla^2(\log f) \preceq -cI$.

Lower bound for NPMLE A particular consequence of Theorem 1 is the following: when the sample are generated from a finite Gaussian mixture, say, $N(0,1)$, with high probability the NPMLE outputs a Gaussian mixture with at most $O(\log n)$ components. To understand the NPMLE from the perspective of overparameterization, it is of great interest to determine whether this bound is tight. (Note that the reasoning in Remark 1 only shows that this is tight when the true density is $N(0,\sigma^2)$ for any $\sigma^2 > 1$.) If so, it would show that the unpenalized NPMLE indeed selects a slightly inflated model (at the price of being fully automatic) and the model selection criterion, such as BIC [Ler92, Ker00], is genuinely needed for achieving consistency in estimating the order of the mixture.

As mentioned in Section 1, such lower bound is known to hold for the Grenander estimator (NPMLE for monotone density): if the true density f is uniform, then the number of pieces in the Grenander estimator is asymptotically $N(\log n, \log n)$ [GL93]. This is a direct consequence of a celebrated result of Sparre Andersen on the least concave majorant of empirical CDF [SA54, Gro20], whose discontinuity in slope correspond to the atoms of Grenander estimator. For the NPMLE in mixture models, no such simple characterization is known other than the first-order condition (6).

Multivariate models Compared to the univariate case, the structure of the NPMLE is far less well understood for multivariate models. Indeed, the general theory developed in [Lin95] relies on the parameter space being one-dimensional. For instance, for the simplest Gaussian location mixture, even the uniqueness of the solution is open in dimension $d \geq 2$. Similar to the analysis in the current paper, bounding the number of atoms in the NPMLE boils down to counting the critical points of a Gaussian mixture (39) with centers being the individual observations, which, if drawn from a subgaussian distribution, lie in a hypercube of size $O(\sqrt{\log n})$ with high probability. The construction in [KK20, Proposition 3] shows that there exists a mixing distribution on $[-a,a]^d$ whose Gaussian location mixture has $\Omega(a^{2d})$ modes. However, it is unclear whether this is tight and directly extending the complex-analytic technique in this paper to multiple dimensions appears challenging.

On the other hand, although the uniqueness of the NPMLE is not settled, the usual analysis of maximal likelihood (zeroth-order optimality) yields statistical guarantees that apply to any solution of the NPMLE [DZ16, SG20]. For example, extending the work of [Zha09], [SG20, Corollary 2.2] showed that if the mixing distribution is compactly supported, then the estimated mixture density has squared Hellinger accuracy of $O_d((\log n)^{d+1}/n)$.

In view of the above results, we conjecture that the solution to the NPMLE for multivariate Gaussian mixtures is unique and, furthermore, given a subgaussian sample of size n it is typically $(\log n)^{C(d)}$-atomic when the dimension d is not too big.

Log-concave NPMLE The NPMLE for log-concave densities is well-studied in nonparametric statistics literature. Basic properties (such as the almost sure existence and uniqueness) and computational algorithms are obtained in [PWM07, DR09] in one dimension and extended to multiple dimensions [CSS10]. In particular, similar to the NPMLE for monotone density (Grenander estimator) which is piecewise constant, the logarithmic of the NPMLE for log-concave density is piecewise affine with at most n pieces; however, unlike the Grenander estimator, its typical structure (e.g. the number of pieces) is little understood, partly because the optimal condition is more complicated.

In terms of statistical results, in one dimension the minimax squared Hellinger rate is shown to be $\Theta(n^{-4/5})$ [DW16, KS16]. For dimension $d \geq 2$, [KS16] proved the minimax lower bound $\Omega(n^{-2/(d+1)})$ and showed it can be attained by the NPMLE up to logarithmic factors for $d = 2$ and 3. This near-optimality of NPMLE is recently extended to any dimension in [KDR19, Han19]. In

view of the corresponding results for the Grenander estimator, if one interprets the minimax rate as the effective dimension divided by the sample size, it is reasonable to conjecture that the typical number of pieces in the log-concave NPMLE is $O(n^{1/5})$ and $O(n^{(d-1)/(d+1)})$ for $d \geq 2$.

Acknowledgment

This work was partially completed when the authors were visiting the Information Processing Group at the School of Computer and Communication Sciences of EPFL, whose generous support is gratefully acknowledged and whose seminar, canceled due to COVID-19, nevertheless brought the independent work [DYPS20] to our attention. The authors thank Pengkun Yang for helpful discussions at the onset of the project and for informing us [KK20]. The authors are also grateful to Roger Koenker for helpful discussion on [KG19] and providing numerical simulation.

Y. Wu is supported in part by the NSF Grant CCF-1900507, NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. Y. Polyanskiy is supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, and the MIT-IBM Watson AI Lab.

References

[Bir89] Lucien Birgé. The Grenader estimator: A nonasymptotic approach. *The Annals of Statistics*, pages 1532–1549, 1989.

[CSS10] Madeleine Cule, Richard Samworth, and Michael Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607, 2010.

[DR09] Lutz Dümbgen and Kaspar Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[DW16] Charles R Doss and Jon A Wellner. Global rates of convergence of the MLEs of log-concave and s-concave densities. *The Annals of Statistics*, 44(3):954, 2016.

[DYPS20] A. Dytso, S. Yagli, H. V. Poor, and S. Shamai (Shitz). The capacity achieving distribution for the amplitude constrained additive Gaussian channel: An upper bound on the number of mass points. *IEEE Transactions on Information Theory*, 66(4):2006–2022, 2020. arxiv:1901.03264v4.

[DZ16] Lee H Dicker and Sihai D Zhao. High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

[Egg58] H. G. Eggleston. *Convexity*, volume 47 of *Tracts in Math and Math. Phys.* Cambridge University Press, 1958.

[GG71] IJ Good and Ray A Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[GJ14] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014.

[GL93] Piet Groeneboom and HP Lopuhaa. Isotonic estimators of monotone densities and distribution functions: basic facts. *Statistica Neerlandica*, 47(3):175–183, 1993.

[Gre56] Ulf Grenander. On the theory of mortality measurement. Part II. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.

[Gre81] Ulf Grenander. *Abstract inference*. John Wiley & Sons, New York, 1981.

[Gro11] Piet Groeneboom. Vertices of the least concave majorant of brownian motion with parabolic drift. *Electronic Journal of Probability*, 16:2334–2358, 2011.

[Gro20] Piet Groeneboom. Grenander functionals and Cauchy's formula. *Scandinavian Journal of Statistics*, pages 1–20, 2020. arXiv preprint arXiv:1902.08806.

[GvdV01] 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*, 29(5):1233–1263, 2001.

[GvdV07] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

[GW92] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

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

[Han19] Qiyang Han. Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions. *arXiv preprint arXiv:1905.12823*, 2019.

[Jew82] Nicholas P Jewell. Mixtures of exponential distributions. *The Annals of Statistics*, 10(2):479–484, 1982.

[JZ09] Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

[JZB+16] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems*, pages 4116–4124, 2016.

[KDR19] Gil Kur, Yuval Dagan, and Alexander Rakhlin. Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*, 2019.

[Ker00] Christine Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62(1):49–66, 2000.

[KG19] Roger Koenker and Jiaying Gu. Comment: Minimalist g-modeling. *Statistical Science*, 34(2):209–213, 2019.

[Kim14] Arlene KH Kim. Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802–1818, 2014.

[KK20] Navin Kashyap and Manjunath Krishnapur. How many modes can a constrained Gaussian mixture have? *Arxiv preprint arXiv:2005.01580*, April 2020.

[KM14] Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

[KS16] Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, 44(6):2756–2779, 2016.

[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.

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

[Ler92] Brian G Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[Lin83a] Bruce G Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

[Lin83b] Bruce G Lindsay. The geometry of mixture likelihoods, part II: the exponential family. *The Annals of Statistics*, 11(3):783–792, 1983.

[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.

[LR93] Bruce G Lindsay and Kathryn Roeder. Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, 21(2):139–147, 1993.

[MM11] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, 2011.

[PW20] Yury Polyanskiy and Yihong Wu. Note on approximating the Laplace transform of a Gaussian on a unit disk. *arXiv preprint arXiv:2008.13372*, 2020.

[PWM07] Jayanta Kumar Pal, Michael Woodroofe, and Mary Meyer. *Estimating a Polya frequency function$_2$*, volume Volume 54 of *Lecture Notes–Monograph Series*, pages 239–249. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

[Rob50] Herbert Robbins. A generalization of the method of maximum likelihood: Estimating a mixing distribution (Abstract). In *Annals of Mathematical Statistics*, volume 21, pages 314–315, 1950.

[SA54] Erik Sparre Andersen. On the fluctuations of sums of random variables. *Mathematica Scandinavica*, pages 263–285, 1954.

[SG20] Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, 48(2):738–762, 2020.

[Sil82] Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

[Sim76] Léopold Simar. Maximum likelihood estimation of a compound poisson process. *The Annals of Statistics*, pages 1200–1209, 1976.

[SW14] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45–114, 2014.

[Tij71] R. Tijdeman. On the number of zeros of general exponential polynomials. *Indagationes Mathematicae (Proceedings)*, 74:1 – 7, 1971.

[vdG00] Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

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

[WY20] Yihong Wu and Pengkun Yang. Optimal estimation of Gaussian mixtures with denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007, 2020. arxiv:1807.07237.

[Zha09] Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.