Nonparametric Mixture MLEs Under Gaussian-Smoothed Optimal Transport Distance

Fang Han, Zhen Miao, and Yandi Shen

***Abstract*— The Gaussian-smoothed optimal transport (GOT) framework, pioneered by Goldfeld et al. and followed up by a series of subsequent papers, has quickly caught attention among researchers in statistics, machine learning, information theory, and related fields. One key observation made therein is that, by adapting to the GOT framework instead of its unsmoothed counterpart, the curse of dimensionality for using the empirical measure to approximate the true data generating distribution can be lifted. The current paper shows that a related observation applies to the estimation of nonparametric mixing distributions in discrete exponential family models, where under the GOT cost the estimation accuracy of the nonparametric MLE can be accelerated to a polynomial rate. This is in sharp contrast to the classical sub-polynomial rates based on unsmoothed metrics, which cannot be improved from an information-theoretical perspective. A key step in our analysis is the establishment of a new Jackson-type approximation bound of Gaussian-smoothed Lipschitz functions. This insight bridges existing techniques of analyzing the nonparametric MLEs and the new GOT framework.**

***Index Terms*— GOT distance, nonparametric mixture models, nonparametric maximum likelihood estimation, rate of convergence, function approximation.**

I. Introduction

LET $f(x\,|\,\theta)$ be a known parametric density function with respect to a certain (counting or continuous) measure and $X_1, \ldots, X_n$ be n i.i.d. observations drawn from the following *mixture density function*,

$$h_Q(x) := \int f(x\,|\,\theta)\mathrm{d}Q(\theta), \tag{1}$$

where Q is unspecified and termed the *mixing distribution* of θ. Our goal is to estimate the unknown Q based on $X_1, \ldots, X_n$. This is the celebrated nonparametric mixing distribution estimation problem, which has been extensively studied in literature [2]. The focus of this paper is on studying the estimation of Q in the case of (identifiable) *discrete exponential family models* [3], i.e., $f(x\,|\,\theta)$ taking the following form that is known to us:

$$f(x\,|\,\theta) = g(\theta)w(x)\theta^x, \tag{2}$$

with $x = 0, 1, 2, \ldots$, $w(x) > 0$ for all $x \geq 0$, and $0 \leq \theta \leq$ (a known fixed constant) $\theta_* < \theta_r$, where $\theta_r \in (0, \infty]$ is the radius of convergence of the power series $\theta \mapsto \sum_{x=0}^{\infty} w(x)\theta^x$ and $g(\cdot)$ is analytic in a neighborhood of 0. This model includes, among many others, Poisson and negative binomial distributions. We assume throughout the paper that the support of the true mixing distribution Q is contained in $[0, \theta_*]$.

Estimation of Q under the discrete exponential family models has been extensively investigated in literature through, e.g., the use of nonparametric maximum likelihood estimators (MLEs) [4], method of moments [5], Fourier and kernel methods [3], [6], [7], and projection methods [8], [9], [10]. Of particular interest to us is the MLE-based approach, partly due to its asymptotic efficiency under regular parametric models. In the case of nonparametric mixture models, the MLE can be written as

$$\widehat{Q} := \operatorname*{argmax}_{\widetilde{Q} \text{ on } [0,\theta^*]} \sum_{i=1}^{n} \log h_{\widetilde{Q}}(X_i), \tag{3}$$

which is a convex problem with efficient solving algorithms [4].

Although a proof of the consistency of $\widehat{Q}$ has been standard now (cf. [11]), of central importance to statisticians and machine learning scientists in making inference based on $\widehat{Q}$ is its rate of convergence. It is by now well-understood that the estimation of Q and many other deconvolution-related problems suffer from a sub-polynomial rate. In this regard, [3] established the first minimax lower bound, indicating that, at the worst case, it is impossible for the MLEs to achieve a polynomial rate if measured using regular metrics such as the total variation distance and the optimal transport distance (OT; in this paper restricted to the Wasserstein-1 distance W_1); see also [12], [13], [14] for slow rates in other deconvolution problems. More recently, it was shown that for Poisson mixtures, the minimax rate of convergence under the W_1 distance is $\log\log n/\log n$ and could indeed be achieved by MLEs [15]. The aforementioned slow rates demonstrate that the estimation of Q suffers severely from its nonparametric structure.

Interestingly, a similar fundamental "curse" also exists in using the empirical measure P_n of an independently and identically distributed (i.i.d.) sample of size n to approximate the true data generating distribution P in $\mathbb{R}^d$. First studied by Dudley [16] and then refined in a series of recent work [17], [18], [19], [20], it is now well known that under some moment/regularity assumptions on P, the minimax rate under the W_1 distance is $n^{-1/d}$ as $d > 2$. Partly motivated by a problem of estimating information flows in deep neural

Manuscript received 6 December 2021; revised 27 June 2023; accepted 6 July 2023. Date of publication 25 July 2023; date of current version 22 November 2023. The work of Fang Han was supported by NSF under Grant DMS-1712536, Grant SES-2019363, and Grant DMS-2210019. *(Corresponding author: Fang Han.)*

Fang Han and Zhen Miao are with the Department of Statistics, University of Washington, Seattle, WA 98195 USA (e-mail: fanghan@uw.edu; zhenm@uw.edu).

Yandi Shen is with the Department of Statistics, University of Chicago, Chicago, IL 60615 USA (e-mail: ydshen@uchicago.edu).

Communicated by E. Gassiat, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2023.3296380

networks [1], [21] introduced a new distance W_1^σ named the *Gaussian-smoothed OT (GOT)* distance, which is defined by

$$W_1^\sigma(Q_1, Q_2) := W_1(Q_1 * \mathcal{N}_\sigma, Q_2 * \mathcal{N}_\sigma). \tag{4}$$

Here $\mathcal{N}_\sigma$ is short for the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. In words, W_1^σ proceeds by first smoothing the target distributions (P, Q) via Gaussian convolution, and then computing their W_1 distance. The GOT distance, like its unsmoothed counterpart, is a metric on the probability measure space with finite first moment that metrizes the weak topology. In addition, both W_1^σ and the corresponding optimal transport plan converge weakly to the corresponding unsmoothed versions as the smoothing parameter $\sigma \to 0$ (cf. Theorems 2, 3, and 4 in [22]).

Under this new distance and with some further moment conditions on P, [1] was able to prove an upper bound of $W_1^\sigma(P_n, P)$ that is of the best possible root-n order and thus overcomes the curse of dimensionality faced with the classical unsmoothed scenario. Subsequent developments establish the weak convergence of $W_1^\sigma(P_n, P)$ to a functional of a Gaussian process [23], weaken the moment assumption [24], and study high noise limit as $\sigma \to \infty$ [25].

One of the main contributions of this paper is to prove that an observation similar to what was made in [1] occurs to the nonparametric mixture MLEs, i.e., under some conditions on $w(\cdot)$, we have

$$\sup_{Q \text{ on } [0,\theta^*]} \mathbb{E} W_1^\sigma(\widehat{Q}, Q) \le C(\sigma, \theta_*, w) n^{-\eta(\theta_*, w)}, \tag{5}$$

where C and η are two positive constants only depending on $\{\sigma, \theta_*, w\}$ and $\{\theta_*, w\}$, respectively. Our result thus bridges two distinct areas, namely nonparametric mixing distribution estimation and empirical approximation to population distribution; in the earlier case, GOT is shown to boost the convergence rate to polynomial, while in the latter case GOT overcomes the curse of dimensionality.

The main technical step in our proof of (5) is a new Jackson-type bound on the error of degree-k polynomial (for an arbitrary positive integer k) approximation to Gaussian-smoothed Lipschitz functions with a bounded support. Our result thus extends the classic Jackson's Theorem (see [26]; Lemma 15) and paves a way to leverage existing technical tools of analyzing the nonparametric MLEs, devised in an early draft written by some of the authors in this paper [15, Sec. 6].

Notation: For any positive integer n, let $[n] := \{1, \ldots, n\}$. Let $\mathbb{Z}_+$ denote the set of all nonnegative integers. For any two distributions Q_1 and Q_2 over $\mathbb{R}^d$, let $Q_1 * Q_2$ represent the convolution of Q_1 and Q_2, i.e., $Q_1 * Q_2(A) = \int\int \mathbb{1}_A(x + y) \mathsf{d}Q_1(x) \mathsf{d}Q_2(y)$, with $\mathbb{1}_\cdot(\cdot)$ standing for the indicator function. For any two measurable functions f, g on $\mathbb{R}^d$, $f * g$ represents their convolution, i.e., $f * g(x) = \int f(x-y) g(y) \mathsf{d}y$. For any function $f : \mathbb{R} \to \mathbb{R}$ and $\alpha > 0$, let $f^{(\alpha)}$ represent the α-time derivative of f. With the help of Kantorovich-Rubinstein formula, the OT (i.e., Wasserstein W_1) distance between Q_1 and Q_2 is as

$$W_1(Q_1, Q_2) := \sup_{\ell \in \mathrm{Lip}_1} \int \ell(\mathsf{d}Q_1 - \mathsf{d}Q_2),$$

where the supremum is over all 1-Lipschitz functions (under the Euclidean metric $\|\cdot\|$) on $\mathbb{R}^d$. The GOT distance W_1^σ is defined as

$$W_1^\sigma(Q_1, Q_2) := W_1\big(Q_1 * \mathcal{N}(0, \sigma^2 I_d), Q_2 * \mathcal{N}(0, \sigma^2 I_d)\big), \tag{6}$$

where $\mathcal{N}(0, \sigma^2 I_d)$ is the d-dim Gaussian distribution with mean 0 and covariance $\sigma^2 I_d$. Let ϕ_σ denote the density function of $\mathcal{N}(0, \sigma^2 I_d)$, whose dimension will be clear from the context. Let $\mathcal{P}(\mathbb{R}^d)$ represent the set of all Borel probability measures on $\mathbb{R}^d$ and $\mathcal{P}_1(\mathbb{R}^d)$ be the subset of $\mathcal{P}(\mathbb{R}^d)$ with elements of finite first moment. Throughout the paper, C, C', C'', c, c' are used to represent generic positive constants whose values may change in different locations.

Paper Organization: The rest of this paper is organized as follows. Section II gives the preliminaries on the studied nonparametric mixture models and the MLEs. Section III delivers the main results, including the key technical insights to the proof. Section IV collects the main proofs, with auxiliary proofs relegated to Section V.

II. Preliminaries

A. Nonparametric Mixture MLEs

Estimating the mixing distribution is known to be statistically challenging in a variety of nonparametric mixture models including the Gaussian [27], binomial [28], [29], and Poisson [15] ones. Specific to the discrete exponential family models in the form of (2), the following $\log n$-scale information-theoretical lower bound formalized this difficulty under the standard unsmoothed W_1 distance. We present its proof in Section IV-B.

Theorem 1 (Minimax Lower Bounds Under W_1 Distance): Let $n \ge 2$ and $(X_1, \ldots, X_n)$ be an i.i.d. sample generated from the mixture density function h_Q defined in (1).

(a) For any $f(x \,|\, \theta)$ taking the form (2), we have

$$\inf_{\widetilde{Q}} \sup_{Q \text{ on } [0,\theta_*]} \mathbb{E} W_1(Q, \widetilde{Q}) \ge \frac{c}{\log n}, \tag{7}$$

where the infimum is taken over all measurable estimators of the mixing distribution Q with support on $[0, \theta_*]$ and $c = c(\theta_*) > 0$ only depends on θ_*.

(b) (Theorem 6.2 in [15]) Suppose further $f(x \,|\, \theta) = e^{-\theta}\theta^x / x!$ is the probability mass function of the Poisson distribution with mean parameter θ. Then

$$\inf_{\widetilde{Q}} \sup_{Q \text{ on } [0,\theta_*]} \mathbb{E} W_1(Q, \widetilde{Q}) \ge \frac{c' \log\log n}{\log n},$$

for a constant $c' = c'(\theta_*) > 0$ depending only on θ_*.

Remark 2: In (2), the condition "$w(x) > 0$ for all $x \in \mathbb{Z}_+$" is a sufficient condition to ensure the identifiability of Q in (1). Similar conditions were also posed in, e.g., Corollary 1 in [3] and Corollary 1 in [6]. As a matter of fact, Theorem 1(a) in [30] showed that, if there exists some $x_0 \in \mathbb{Z}_+$ such that $f(x \,|\, \theta) = 0$ for all $x \ge x_0$ and $\theta \in [0, \theta_*]$, then Q is not identifiable, i.e., there exist at least two distinct mixing distributions Q_1, Q_2 over $[0, \theta_*]$ such that $h_{Q_1} = h_{Q_2}$. On the other hand, it is straightforward to generalize Theorem 1 to

the case of "$w(x) > 0$ for all $x \geq x_0$ for some $x_0 \in \mathbb{Z}_+$ that is known to us".

In the past several decades, methods that provably (nearly) achieve the above minimax lower bounds have been proposed; cf. [3], [9], and [10] among many others. However, none of the above methods is likelihood-based, partly due to the theoretical challenges faced with analyzing the nonparametric MLEs. A major breakthrough towards understanding the rate of convergence of nonparametric mixture MLEs was made in [29] for the binomial case and [15] for the Poisson case.

The following theorem provides an extension of Theorem 6.1 in [15] to cover general discrete mixture models of the form (2). We present its proof in Section IV-C.

Theorem 3 (Minimax Upper Bounds of MLEs Under W_1 Distance): Let $(X_1, \ldots, X_n)$ be an i.i.d. sample generated from the mixture density function h_Q defined in (1), and $\widehat{Q}$ be the MLE defined in (3). The following are true.

(a) If there exists some $C \geq 1$ such that $1/w(x) \leq C^x$ for all $x \in \mathbb{Z}_+$, then there exists some $C' = C'(\theta_*, C) > 0$ such that

$$\sup_{Q \text{ on } [0,\theta_*]} \mathbb{E} W_1(Q, \widehat{Q}) \leq \frac{C'}{\log n}.$$

(b) If there exists some $C \geq 1$ such that $1/w(x) \leq (Cx)^{Cx}$ for all integers $x \geq 1$, then there exists some $C' = C'(\theta_*, C) > 0$ such that

$$\sup_{Q \text{ on } [0,\theta_*]} \mathbb{E} W_1(Q, \widehat{Q}) \leq \frac{C' \log\log n}{\log n}.$$

Remark 4: Theorem 3 is concerned with two types of tails conditions (exponential and super-exponential) for $1/w(x)$; they are classical and ensure the identifiability of Q as discussed in Remark 2. Similar conditions were posed in Theorem 4 in [3], Corollary 1 in [6], Theorem 1 in [7], and Corollary 1 in [10].

Remark 5: It is straightforward to verify that, after some standard operations including location shift, point mass inflation, and reparametrization, Theorem 3(a) applies to, e.g., the (zero-inflated or C-truncated) negative binomial, the logarithmic [31], the lost games [32], as well as the generalized Poisson, negative binomial, and logarithmic [33] distributions; Theorem 3(b) applies to, e.g., the (zero-inflated or C-truncated) Poisson as well as the Poisson polynomial [34] distributions.

B. The GOT Distance

Theorem 1 suggests that, under the W_1 cost, the sub-polynomial rate in estimating the mixing distribution of a nonparametric mixture model is information-theoretically optimal. As a matter of fact, the conclusion of Theorem 1 goes beyond the discrete exponential family models studied in this paper; cf. Proposition 8 in [27] for a similar phenomenon in the nonparametric Gaussian mixture models.

Revising the Wasserstein distance through convolution/smoothing has a long and rich history. In probability theory, this is interestingly related to heat semigroup operators on Riemannian manifold [35], which reveals its connection to the Ricci curvature. More recently, stemming from the interest in estimating the mutual information of deep networks, [21] initiated the study of GOT distances, introduced as a smoothed alternative to the classic OT metric.

Indeed, the GOT distance is now known to be able to effectively alleviate some undesired issues associated with the OT distances. Let us start with the following simple fact, that the W_1-distance is non-increasing under convolution.

Lemma 6: Consider $\mu_1, \mu_2, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ be arbitrary three Borel probability measures on $\mathbb{R}^d$ with finite first moment. We then have

$$W_1(\mu_1 * \nu, \mu_2 * \nu) \leq W_1(\mu_1, \mu_2).$$

Proof: Recall the duality definition of $W_1(\mu_1, \mu_2)$ as

$$W_1(\mu_1, \mu_2) := \inf \mathbb{E}\|X - Y\|,$$

with the infimum is taken over all couplings of (X, Y) such that $X \sim \mu_1$ and $Y \sim \mu_2$. We then consider any such (X, Y) and assume Z to be independent of (X, Y) and follows the distribution of ν. Then it is immediate that

$$W_1(\mu_1 * \nu, \mu_2 * \nu) \leq \mathbb{E}\|(X + Z) - (Y + Z)\| = \mathbb{E}\|X - Y\|,$$

and accordingly (by taking infimum over all such (X, Y))

$$W_1(\mu_1 * \nu, \mu_2 * \nu) \leq W_1(\mu_1, \mu_2).$$

This completes the proof. □

Lemma 6 confirms that the GOT distance is no greater than the original OT distance, but it does not quantify the difference. For that purpose, the existing literature has provided us with an interesting example, i.e., in approximating the population measure using the empirical one. In detail, suppose P_n is the empirical measure of P, and both are supported on $\mathbb{R}^d$ with some integer $d \geq 3$. Theorem 1 in [19] showed that

$$\sup_{P: \mathbb{E}_P \|X\|^2 < \infty} \mathbb{E} W_1(P, P_n) \asymp n^{-1/d},$$

which is faced with severe curse of dimensionality as the dimension d becomes larger. In a recent paper of [1], the authors showed that, via appealing to the GOT one, this curse can be effectively handled. More specifically, they proved that, as long as P is sub-gaussian with a fixed subgaussian constant, we have

$$\mathbb{E} W_1^\sigma(P, P_n) \lesssim n^{-1/2},$$

which is the parametric rate of convergence. See also [23] for the limiting distribution of $\sqrt{n} W_1^\sigma(P, P_n)$ as well as [24] for the relaxation of the moment conditions on P.

The purpose of this paper is to present the second and also a statistically interesting example, for which adopting the GOT distance can significantly accelerate the convergence rate of a statistical procedure.

III. Main Results

The following theorem is the main result of this paper.

Theorem 7: Let $(X_1, \ldots, X_n)$ be an i.i.d. sample generated from the mixture density function h_Q defined in (1), and $\widehat{Q}$ be the MLE introduced in (3). Suppose that there exist

some positive constants $c_1, c_2, c_3, C_1, C_2, C_3$ depending only on $w(\cdot)$ such that one of the following two conditions holds,

(i) $c_1 c_2^x \le 1/w(x) \le C_1 C_2^x$ for all $x = 1, 2, \ldots$;
(ii) $c_1 c_2^x x^{c_3 x} \le 1/w(x) \le C_1 C_2^x x^{C_3 x}$ for all $x = 1, 2, \ldots$.

Then we have

$$\sup_{Q \text{ on } [0,\theta_*]} \mathbb{E} W_1^\sigma(Q, \widehat{Q}) \le C \cdot n^{-c}.$$

Here $C = C(\sigma, \theta_*, c_1, c_2, c_3, C_1, C_2, C_3)$ and $c = c(\theta_*, c_3, C_2, C_3)$ are two positive constants.

Remark 8: Let us point out some results in the nonparametric mixture model literature that are relevant to ours. Reference [36] studied the convergence of $h_{\widehat{Q}}$ to h_Q in a specific nonparametric Poisson mixture model based on the regular unsmoothed distance. They observed that the convergence rate can be nearly parametric; cf. Proposition 3.1 therein. This observation is particularly relevant to ours as the map $Q \mapsto h_Q$ is intrinsically also "smoothing" the probability measure. The Gaussian analogue of this result was derived in [37] and [38], which showed that the nonparametric MLE achieved near-parametric convergence for mixture density estimation when the mixing distribution has bounded/sub-Gaussian tails. A similar observation was made in [27], which studied the estimation of mixing distributions in (finite) Gaussian mixture models via a method of moments. In particular, their Lemma 8 considers bounding the chi-squared distance between two Gaussian mixtures with sub-gaussian mixing distributions whose first k moments are identical. Their bound suggests a similar exponential-order improvement as ours. However, it is clear from the context that the proof techniques in [36], [37], [38], and [27] are distinct from the current paper, where, as we detail next, the conclusion is arrived via a new Jackson-type bound.

Remark 9: From a methodological perspective, one may interpret Theorem 3.1 as a motivation to use Gaussian smoothing for studying mixing distribution-related statistical properties. In practice, this insight has partly helped the authors in a separate study of autism spectrum disorder [15], [39]. There the problem of interest is to identify differentially expressed genes based on brain single-cell data, which were modeled using nonparametric Poisson mixtures. Reference [15] found that, compared to the unsmoothed version, there are extra biologically plausible genes discovered based on Poisson-smoothed nonparametric MLEs, which are natural counterparts of the Gaussian-smoothed ones investigated in this paper. See the discussions before Theorem 2.1 in [15] for more details.

Remark 10: In Theorem 7 the explicit value of c was not exposed. For readers of interest, considering $\epsilon \in (0, 1)$ to be an arbitrarily small positive constant, the largest possible c we can obtain for the Poisson mixture is

$$1/10 - \epsilon$$

and for negative binomial mixture is

$$\Big[2\Big\{1 + 2 \cdot \frac{\log(e/\theta_*)}{\log(1/\theta_*)}\Big\}\Big]^{-1} - \epsilon,$$

recalling that $\theta_* \in (0, 1)$ in this case.

We present the proof of the above two claims in Section IV-E. While we have not been able to pin down the sharpest possible value of c, it is our conjecture that for any fixed σ, the best possible rate under W_1^σ, at least in the Poisson case, should be the parametric rate $n^{-1/2}$ up to some logarithmic factors. In other words, the parametric rate as was observed in [1] is also (up to some logarithmic terms) recoverable in the setting of nonparametric mixture MLEs considered in this paper. An improvement of the current upper bound analysis calls for the parametric convergence rate of the nonparametric MLE under linear functionals (e.g., polynomial functionals). There has been some recent work (e.g., [27], [40]) that studies the estimation of linear functionals in deconvolution problems, but to our best understanding this is still open for the nonparametric MLEs.

Next we give a proof sketch of Theorem 7, which shares some common steps with that of Theorem 3. Invoking the same argument that was used in the proof of Theorem 6.1(a) in [15], for any given 1-Lipschitz function $\ell(\cdot)$ such that $\ell(0) = 0$, we introduce the following function to approximate it,

$$\widehat{\ell}_k(\theta) := \sum_{x=0}^{k} b_{x,\ell} f(x|\theta), \quad \text{for } b_{x,\ell} \in \mathbb{R} \text{ and } \theta \in [0, \theta_*],$$

where k is an integer to be chosen later. In the unsmoothed case, some straightforward manipulations (see Section IV-C for details) then yield

$$\begin{aligned} W_1(Q, \widehat{Q}) \le \sup_{\ell \in \mathrm{Lip}_1, \ell(0)=0} \Big\{ & 2 \sup_{\theta \in [0,\theta_*]} \Big| \ell(\theta) - \widehat{\ell}_k(\theta) \Big| \\ & + \Big| \sum_{x=0}^{k} b_{x,\ell} \Big(h_Q(x) - h^{\mathrm{obs}}(x) \Big) \Big| \\ & + \Big| \sum_{x=0}^{k} b_{x,\ell} \Big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x) \Big) \Big| \Big\}, \end{aligned} \tag{8}$$

where $h^{\mathrm{obs}}(x) := n^{-1} \sum_{i=1}^{n} \mathbb{1}(x = X_i)$. The last two terms of (8) can be handled by concentration arguments (see Lemma 17): with high probability,

$$\begin{aligned} & \Big| \sum_{x=0}^{k} b_{x,\ell} \Big(h_Q(x) - h^{\mathrm{obs}}(x) \Big) \Big| \vee \Big| \sum_{x=0}^{k} b_{x,\ell} \Big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x) \Big) \Big| \\ & \lesssim \frac{\max_{0 \le x \le k} |b_{x,\ell}|}{n^{1/2-\epsilon}}, \end{aligned}$$

where $\epsilon > 0$ is some small constant. It remains to bound the coefficients $\max_{0 \le x \le k} |b_{x,\ell}|$ and the approximation error in the first term of (8). To this end, we apply a Jackson-type bound (see Lemma 12) and obtain:

$$\sup_{\ell \in \mathrm{Lip}_1, \ell(0)=0} \sup_{\theta \in [0,\theta_*]} \Big| \ell(\theta) - \widehat{\ell}_k(\theta) \Big| \le C/k, \tag{9}$$

and,

$$\sup_{\ell \in \mathrm{Lip}_1, \ell(0)=0} \max_{0 \le x \le k} \Big| b_{x,\ell} \Big| \le C^k \max_{0 \le x \le k} \Big\{ \frac{1}{w(x)} \Big\},$$

where $C > 0$ is independent of k and ℓ. After plugging in the two types of tails of $1/w(\cdot)$, balancing the above two terms yields the optimal choices of $k \sim \log n$ or

$k \sim \log n/\log\log n$, leading to the two sub-polynomial bounds of $\mathbb{E}W_1(Q,\widehat{Q})$ in Theorem 3.

With these concepts in mind, let us move on to the smoothed case where the GOT distance is used. Similar to the derivation of (8) and further noting that $\int \ell \mathsf{d}(Q * \mathcal{N}_\sigma) = \int \ell_\sigma \mathsf{d}Q$ with $\ell_\sigma := \ell * \phi_\sigma$, one can show that

$$W_1^\sigma(Q,\widehat{Q}) \le \sup_{\ell\in \mathrm{Lip}_1, \ell(0)=0} \Big\{ 2 \sup_{\theta\in[0,\theta_*]} \Big| \ell_\sigma(\theta) - \ell_\sigma(0) - \widehat{\ell}_k(\theta) \Big| + \Big| \sum_{x=0}^k b_{x,\ell} \Big(h_Q(x) - h^{\mathrm{obs}}(x) \Big) \Big| + \Big| \sum_{x=0}^k b_{x,\ell} \Big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x) \Big) \Big| \Big\}. \tag{10}$$

The last two terms in (10) can be similarly handled as in (8), and it remains to control the first term. Up to some negligible terms, we prove the following approximation bound (see Lemma 11 for precise statement):

$$\sup_{\theta\in[a,b]} \Big| \ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta) \Big| \lesssim k^{-k/2}.$$

In contrast to the linear convergence in the classical Jackson-type bound (9), the above bound states that approximation to Gaussian-smoothed Lipschitz functions by degree-k polynomials is super-exponentially fast, hinting a substantial gain of convergence speed whence GOT distances are used to quantify the distance. We present the proof of the above bound in Section V-A, which is based on a general recipe of Devore (cf. Lemma 13) that bounds the approximation error by the modulus of continuity of (the derivatives of) the target function. We refer to Section IV-D for the complete proof of Theorem 7.

IV. Proofs of Main Results

In the subsequent proofs, we sometimes drop the track of dependence on C, C' for simplicity.

A. Approximation Results

This subsection collects all the approximation results used in our proof. Lemmas 13-15 are standard in the literature, and we provide the proofs of Lemmas 11 and 12 in Section V ahead.

Lemma 11 (Polynomial Approximation of Gaussian-Smoothed Lipschitz Functions): Let $0 \in [a,b] \subset \mathbb{R}$ be a bounded interval and let $\ell(\cdot)$ be a 1-Lipschitz function over $[a,b]$. For any $\sigma > 0$ and integer $k > 1$, there exist a constant $C = C(a,b) > 0$ only depending on a, b and a polynomial $p_k(\cdot)$ of degree at most k such that

$$\sup_{\theta\in[a,b]} \Big| \ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta) \Big| \le Ce\sigma \cdot \Big[\frac{2\sqrt{e}\sigma\sqrt{k}}{b-a} \Big]^{-k} k^{-1/4},$$

where we recall that $\ell_\sigma := \ell * \phi_\sigma$ with ϕ_σ standing for the density function of $\mathcal{N}(0,\sigma^2)$.

Lemma 12: For any integer $k \ge 1$ and 1-Lipschitz function $\theta \mapsto \ell(\theta)$ on $[0,\theta_*]$ with $\ell(0) = 0$, there exists some $\widehat{\ell}(\theta) = \sum_{x=0}^k b_x f(x|\theta)$ such that $\max_{\theta\in[0,\theta_*]} |\ell(\theta) - \widehat{\ell}(\theta)| \le C/k$, and

$$\max_{x\in[0,k]} |b_x| \le C^k \cdot \max_{0\le x\le k} 1/w(x),$$

where $C > 0$ is independent of k and $\ell(\cdot)$.

Lemma 13 (Theorem 6.2 in Chapter 7, [41]): For any integer $r \ge 1$, let

$$W_\infty^r([-1,1]) := \Big\{ \psi : [-1,1] \to \mathbb{R} : \psi^{(r-1)} \text{ is absolutely continuous and the supremum of } \psi^{(r)} \text{ on } [-1,1] \text{ is finite} \Big\}$$

be the Sobolev space on $[-1,1]$. For functions $f \in W_\infty^r([-1,1])$ and any integer $k > r$, there exists a polynomial p_k of degree at most k such that

$$\sup_{\theta\in[-1,1]} \Big| f(\theta) - p_k(\theta) \Big| \le Ck^{-r} \omega\big(f^{(r)}, k^{-1}\big),$$

where $C > 0$ is a universal constant and

$$\omega(f^{(r)}, k^{-1}) := \sup_{\theta_1,\theta_2:|\theta_1-\theta_2|\le k^{-1}} \Big| f^{(r)}(\theta_1) - f^{(r)}(\theta_2) \Big|.$$

Lemma 14 (Chapter 2.6 Equation 9 in [42]): Suppose $k \in \mathbb{Z}_+$ and $\theta \mapsto p_k(\theta) = \sum_{x=0}^k c_x \theta^x$ with coefficients $\{c_x\}_{x=0}^k \subset \mathbb{R}$. Then

$$|c_x| \le \frac{k^x}{x!} \max_{|\theta|\le 1} |p_k(\theta)| \le e^k \max_{|\theta|\le 1} |p_k(\theta)|.$$

Lemma 15 (Jackson's Theorem, Lemma 10 of [43] or see [41]): Let $k > 0$ be any integer, and $[a,b] \subseteq \mathbb{R}$ be any bounded interval. For any 1-Lipschitz function $\ell(\cdot)$ on $[a,b]$, there exists a constant C independent of k, ℓ such that there exists a polynomial $p_k(\cdot)$ of degree at most k such that

$$|\ell(\theta) - p_k(\theta)| \le C\sqrt{(b-a)(\theta-a)}/k, \ \forall \theta \in [a,b]. \tag{11}$$

In particular, the following norm bound holds:

$$\sup_{\theta\in[a,b]} |\ell(\theta) - p_k(\theta)| \le C(b-a)/k. \tag{12}$$

Combining with Lemma 14, it follows that the coefficients of $p_k(\theta)$ are bounded by $e^k O(|b-a| + \ell(a))$.

B. Proof of Theorem 1

The proof of part (a) is based on Le Cam's two-point method (cf. Chapter 2.3 in [44]) and uses the following proposition.

Proposition 16 (Lemma 3 in [28], Proposition 4.3 in [29]): For any positive integer k and any $M > 0$, there exist two distributions P_1, P_2 with support in $[0,M]$ such that P_1, P_2 have first k moments identical and $W_1(P_1,P_2) \ge M/(2k)$.

We first upper bound $g^{(x)}(0)\theta_*^x/x!$. Note that $g(\theta)$ is an analytic function of θ and $g(0) \ne 0$, so that it must have an analytic inverse in the neighborhood of 0. Therefore, there exist positive constants C' and C depending only on $w(\cdot)$ and θ_* such that

$$|g^{(x)}(0)\theta_*^x/x!| \le C'(C+1)^x, \quad \text{for all } x = 0,1,2,\ldots. \tag{13}$$

We then combine (13) with Proposition 16 to finish the proof. On one hand, for any $k = 1, 2, \ldots$, Proposition 16 guarantees the existence of two distributions Q_1, Q_2 over $[0, \theta_*/(C+3)]$ such that

$$\int \theta^x \mathsf{d}Q_1(\theta) = \int \theta^x \mathsf{d}Q_2(\theta), \quad \text{for all } x \in [k],$$

and $W_1(Q_1, Q_2) \geq \theta_*/(2(C+3)k)$. On the other hand, the total variation distance between h_{Q_1} and h_{Q_2} satisfies

$$\begin{aligned}
&\mathrm{TV}(h_{Q_1}, h_{Q_2}) \\
&= \frac{1}{2}\sum_{x=0}^{\infty}\Big|\int_0^{\theta_*/(C+3)} g(\theta)w(x)\theta^x \mathsf{d}Q_1(\theta) \\
&\quad - \int_0^{\theta_*/(C+3)} g(\theta)w(x)\theta^x \mathsf{d}Q_2(\theta)\Big| \\
&\leq \sum_{x=0}^{\infty} w(x) \sum_{m: m+x\geq k+1} \frac{|g^{(m)}(0)|}{m!}\Big(\frac{\theta_*}{C+3}\Big)^m \\
&\leq \sum_{x=0}^{\infty} w(x)\Big(\frac{\theta_*}{C+3}\Big)^x \sum_{m: m+x\geq k+1} C'\Big(\frac{C+1}{C+3}\Big)^m \\
&\leq C'\Big(\frac{C+1}{C+3}\Big)^k \sum_{x=0}^{\infty} w(x)\theta_*^x = C' g(\theta_*)^{-1}\Big(\frac{C+2}{C+3}\Big)^k.
\end{aligned}$$

Picking $k \asymp \log n$ so that

$$C' g(\theta_*)^{-1}\Big(\frac{C+2}{C+3}\Big)^k = 1/(2n),$$

it follows from Le Cam's lower bound for two hypotheses that, denoting $Q^{\otimes n}$ to be the n-time product measure of Q,

$$\begin{aligned}
\inf_{\widetilde{Q}} \sup_{Q} \mathbb{E} W_1(Q, \widetilde{Q}) &\geq \frac{1}{2} W_1(Q_1, Q_2)\Big\{1 - \mathrm{TV}(h_{Q_1}^{\otimes n}, h_{Q_2}^{\otimes n})\Big\} \\
&\geq \frac{1}{2} W_1(Q_1, Q_2)\{1 - n/(2n)\} \\
&= \frac{1}{4} W_1(Q_1, Q_2),
\end{aligned}$$

with $W_1(Q_1, Q_2) \geq \theta_*/(2(C+3)k)$ by the construction. Note that here we use the fact that for any discrete distributions P, Q, $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \leq n \cdot \mathrm{TV}(P, Q)$; see, e.g., Lemma B.8(i) in [45]. This completes the proof of part (a). Part (b) has been proved in Theorem 6.2 in [15]. □

C. Proof of Theorem 3

First we need a technical lemma, whose proof will be given in Section V ahead.

Lemma 17 (A generalized version of Lemma B.2 in the supplemental of [15]): Let $\{X_i, i \in [n]\}$ be an i.i.d. sample generated from the mixture distribution h_Q in (1). Let $\widehat{Q}$ be defined in (3), and $h^{\mathrm{obs}}(x) = n^{-1}\sum_{i=1}^n \mathbf{1}_{X_i = x}$ for $x \in \mathbb{N}$. Then for any $\delta \in (0,1)$ and $\epsilon \in (0,1)$, there exists some $C = C(\epsilon, \theta_*) > 0$ such that for any $n \geq 1$,

$$\Big|\sum_{x=0}^{\infty} b_x \big(h^{\mathrm{obs}}(x) - h_Q(x)\big)\Big| \leq C \max_{x\geq 0}|b_x| \sqrt{\frac{1}{n^{1-\epsilon}\delta^{1+\epsilon}}}$$

and

$$\Big|\sum_{x=0}^{\infty} b_x \Big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x)\Big)\Big| \leq C \max_{x\geq 0}|b_x| \sqrt{\frac{1}{n^{1-\epsilon}\delta^{1+\epsilon}}}$$

hold with probability at least $1-\delta$ uniformly over $\{b_x\} \subset \mathbb{R}$.

Proof of Theorem 3: By definition of W_1, we have

$$\begin{aligned}
W_1(Q_1, Q_2) &= \sup_{\ell \in \mathrm{Lip}_1} \int \ell(\mathsf{d}Q_1 - \mathsf{d}Q_2) \\
&= \sup_{\ell \in \mathrm{Lip}_1, \ell(0)=0} \int \ell(\mathsf{d}Q_1 - \mathsf{d}Q_2).
\end{aligned}$$

To control each $\int \ell(\mathsf{d}Q_1 - \mathsf{d}Q_2)$, define the following approximation function of $\ell(\theta)$:

$$\theta \mapsto \widehat{\ell}(\theta) := \sum_{x=0}^{k} b_x f(x|\theta), \text{ where } b_x \in \mathbb{R} \text{ and } \theta \in [0, \theta_*];$$

here the integer k and the values $\{b_x\}_{x=0}^k$ will be specified later. Recall that $h_Q(x) = \int f(x|\theta)\mathsf{d}Q(\theta)$. Then since Q and $\widehat{Q}$ are both supported on $[0, \theta_*]$, direct calculation yields that

$$\begin{aligned}
&\int \ell(\theta)\mathsf{d}\big(Q(\theta) - \widehat{Q}(\theta)\big) \\
&= \int_0^{\theta_*} \ell(\theta)\mathsf{d}\big(Q(\theta) - \widehat{Q}(\theta)\big) \\
&= \int_0^{\theta_*} \big(\ell(\theta) - \widehat{\ell}(\theta)\big)\mathsf{d}\big(Q(\theta) - \widehat{Q}(\theta)\big) \\
&\quad + \sum_{x=0}^{k} b_x \big(h_Q(x) - h_{\widehat{Q}}(x)\big) \\
&\leq 2\|\ell - \widehat{\ell}\|_\infty + \Big|\sum_{x=0}^{k} b_x \big(h_Q(x) - h^{\mathrm{obs}}(x)\big)\Big| \\
&\quad + \Big|\sum_{x=0}^{k} b_x \big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x)\big)\Big|,
\end{aligned}$$

where $\|\ell - \widehat{\ell}\|_\infty = \sup_{\theta\in[0,\theta_*]} |\ell(\theta) - \widehat{\ell}(\theta)|$ and $h^{\mathrm{obs}}(x) = n^{-1}\sum_{i=1}^n \mathbf{1}_{X_i = x}$. This implies

$$\begin{aligned}
W_1(Q, \widehat{Q}) \leq \sup_{\ell \in \mathrm{Lip}(1)} \Big\{&2\|\ell - \widehat{\ell}\|_\infty + \Big|\sum_{x=0}^{k} b_x \big(h_Q(x) - h^{\mathrm{obs}}(x)\big)\Big| \\
&+ \Big|\sum_{x=0}^{k} b_x \big(h^{\mathrm{obs}}(x) - h_{\widehat{Q}}(x)\big)\Big|\Big\}. \qquad (14)
\end{aligned}$$

By Lemma 17, for any $\delta \in (0, 1/2)$ and $\epsilon \in (0,1)$, there exists some $C_1 = C_1(\epsilon, \theta_*) > 0$ such that the sum of the last two terms in (14) is upper bounded by

$$C_1 \max_{x\geq 0}|b_x| / \sqrt{n^{1-\epsilon}\delta^{1+\epsilon}}$$

uniformly over $\{b_x\}$ with probability at least $1 - 2\delta$. The bound on $\max_{x\geq 0}|b_x|$, as we discuss next, depends on the tail of $1/w(x)$.

(i) If $1/w(x) \leq C_2^x$ for some constant $C_2 > 1$ and all $x \geq 0$, it follows from Lemma 12 that for any $k \in \mathbb{Z}_+$ and 1-Lipschitz function $\ell(\theta)$ on $[0, \theta_*]$, there exists an approximation $\widehat{\ell}(\theta) = \sum_{x=0}^k b_x f(x|\theta)$ with $\{b_x\}_{x=0}^k$, such that $\max_{\theta\in[0,\theta_*]} |\ell(\theta) - \widehat{\ell}(\theta)| \leq C_3/k$ and

$$\max_{x\in[0,k]} |b_x| \leq C_3^k / w(k) \leq (C_2 C_3)^k,$$

where $C_3 > 0$ is independent of k and ℓ. Hence it follows from (14) that

$$W_1(Q, \widehat{Q}) \le 2C_3/k + C_1(C_2C_3)^k/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}},$$

for any $n \ge 1$ with probability at least $1-2\delta$. Taking $k = k(n)$ such that $(C_2C_3)^k = n^c$ for some small positive constant c specified later, it follows that

$$\begin{aligned} W_1(Q, \widehat{Q}) &\le 2C_3/k(n) + C_1 n^c/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}} \\ &= 2C_3/k(n) + C_1 n^{c+\epsilon/2-1/2}/\sqrt{\delta^{1+\epsilon}}. \end{aligned}$$

Note that $(C_2C_3)^{k(n)} = n^c$ implies $k(n) = c\log n/\log(C_2C_3)$. Letting $\epsilon = 1/4$ and $c = 1/8$, it follows that

$$\begin{aligned} W_1(Q, \widehat{Q}) &\le 2C_3\log(C_2C_3)/(c\log n) + C_1 n^{c+\epsilon/2-1/2}/\sqrt{\delta^{1+\epsilon}} \\ &\le 16C_3\log(C_2C_3)/\log n + C_1 n^{-1/4}/\delta^{5/8}. \end{aligned}$$

Therefore, for sufficiently large n (depending on θ_*), there exists a positive constant $C_4 = C_4(\theta_*)$ such that $\mathbb{E}W_1(Q, \widehat{Q}) \le C_4/\log n$ by integrating the tail estimate.

(ii) If $1/w(x) \le (C_5x)^{C_5x}$ for some $C_5 > 0$ and all $x \ge 1$, it follows from Lemma 12 that any 1-Lipschitz function $\ell(\theta)$ on $[0, \theta_*]$ can be approximated by $\widehat{\ell}(\theta) = \sum_{x=0}^k b_x f(x|\theta)$ such that $\max_{\theta\in[0,\theta_*]} |\ell(\theta) - \widehat{\ell}(\theta)| \le C_3/k$, and

$$\max_x |b_x| \le C_3^k/w(k) \le (C_5(C_3)^{1/C_5}k)^{C_5k} \le (C_6k)^{C_6k}$$

for $k \ge 1$, where $C_6 = C_6(\theta_*)$ is a constant. Hence it follows that

$$W_1(Q, \widehat{Q}) \le 2C_3/k + (C_6k)^{C_6k}C_1/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}},$$

for any $n \ge 1$ with probability at least $1 - 2\delta$. Taking $k = k(n)$ satisfying $(C_6k)^{C_6k} = n^c$ for a small positive constant c specified later, it follows that

$$W_1(Q, \widehat{Q}) \le 2C_3/k(n) + C_1 n^{c+\epsilon/2-1/2}/\sqrt{\delta^{1+\epsilon}}.$$

Since $(C_6k)^{C_6k} = n^c$ is equivalent to $\log(C_6k)\exp(\log(C_6k)) = c\log n$, it follows that $\log(C_6k(n)) = W(c\log n)$ and hence $k(n) = \exp(W(c\log n))/C_6$, where $W(\cdot)$ is the Lambert W function. Using the expansion

$$W(x) = \log x - \log\log x + o(1), \text{ as } x \to \infty,$$

there exists a constant $C_7 > 0$ such that

$$\exp(W(x)) \ge x/(2\log x) \text{ for } x \ge C_7.$$

Therefore, for sufficiently large n, we have

$$k(n) \ge \frac{c\log n}{2C_6\log(c\log n)}. \tag{15}$$

As a result,

$$\begin{aligned} W_1(Q, \widehat{Q}) \le\ & \{4C_3C_6\log(c\log n)\}/(c\log n) \\ & + C_1 n^{c+\epsilon/2-1/2}/\sqrt{\delta^{1+\epsilon}}, \end{aligned}$$

with probability at least $1 - 2\delta$. Letting $c = 1/8, \epsilon = 1/4$, we have

$$W_1(Q, \widehat{Q}) \lesssim \log\log n/\log n + n^{-1/4}\delta^{-5/8}.$$

Therefore, for sufficiently large n (depending on θ_*), it follows that $\mathbb{E}W_1(Q, \widehat{Q}) \lesssim \log\log n/\log n$ by integrating the tail estimate. □

D. Proof of Theorem 7

Proof of Theorem 7: This proof is divided into three steps.

Step 1: In the first step, we prove that for any $\sigma > 0$, integer $k > 1$, and any $\ell \in \mathrm{Lip}(1)$ on $[-\theta_*, \theta_*]$ with $\ell(0) = 0$, there exist a positive constant $C_4 = C_4(\theta_*, \sigma)$ and a set of coefficients

$$\Big\{b_x \in \mathbb{R}, x = 0, \ldots, 2k\Big\}$$

such that

$$\begin{aligned} &\sup_{\theta\in[0,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - \sum_{0\le x\le 2k} b_x f(x|\theta)\Big| \\ &\le C_4\Big\{\Big[\theta_*\sigma\sqrt{ek}\Big]^{-k} + \sum_{x\ge k+1} w(x)\theta_*^x\Big\}, \end{aligned}$$

where we recall that $\ell_\sigma(\theta) := [\ell * \phi_\sigma](\theta)$ and ϕ_σ is the probability density function of $\mathcal{N}_\sigma$.

For any $k = 1, 2, \ldots$, let $q_k(\theta) := \sum_{x=0}^k w(x)\theta^x$ be a truncation of the function $\theta \mapsto 1/g(\theta) = \sum_{x=0}^\infty w(x)\theta^x$ on $[0, \theta_*]$. Noting that $1/g(\theta)$ is monotonically non-decreasing with respect to θ, one can then readily verify that

$$\begin{aligned} R_k(\theta) &:= g(\theta)\cdot\Big\{\frac{1}{g(\theta)} - q_k(\theta)\Big\} \\ &= g(\theta)\cdot\sum_{x\ge k+1} w(x)\theta^x \\ &\le g(0)\cdot\sum_{x\ge k+1} w(x)\theta_*^x \end{aligned}$$

whenever $\theta \in [0, \theta_*]$.

Let $p_k(\theta)$ be the degree-k polynomial achieving the approximation bound in Lemma 11. We then have

$$\sup_{\theta\in[-\theta_*,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)\Big| \le C_5e\sigma\cdot\Big[2\theta_*^{-1}\sigma\sqrt{ek}\Big]^{-k}, \tag{16}$$

where $C_5 = C_5(\theta_*) > 0$. Then there exists a set of coefficients $\{b_x \in \mathbb{R}, x = 0, 1, \ldots, 2k\}$ such that

$$p_k(\theta)q_k(\theta) = \sum_{x=0}^{2k} b_x w(x)\theta^x.$$

Then we have $g(\theta)p_k(\theta)q_k(\theta) = \sum_{x=0}^{2k} b_x f(x|\theta)$, and the proof in this step is complete by noting that

$$\begin{aligned} &\sup_{\theta\in[0,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)q_k(\theta)g(\theta)\Big| \\ &= \sup_{\theta\in[0,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)\big[1 - R_k(\theta)\big]\Big| \\ &\le 2\sup_{\theta\in[0,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)\Big| \\ &\quad + \sup_{\theta\in[0,\theta_*]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0)\Big|\cdot\sup_{\theta\in[0,\theta_*]} \Big|R_k(\theta)\Big| \\ &\overset{(*)}{\le} C_5 2e\sigma\cdot\Big[2\theta_*^{-1}\sigma\sqrt{ek}\Big]^{-k} + 2(\theta_*+\sigma)g(0)\cdot\sum_{x\ge k+1} w(x)\theta_*^x. \end{aligned} \tag{17}$$

Here in $(*)$ we use the fact that, as $\ell(0)=0$ and $\ell\in\mathrm{Lip}(1)$,

$$\begin{aligned}\sup_{\theta\in[0,\theta_*]}|\ell_\sigma(\theta)| &\le \sup_{|\theta|\le\theta_*}|\ell_\sigma(\theta)|\\ &= \sup_{|\theta|\le\theta_*}\Big|\int \ell(\theta-\theta_1)\phi_\sigma(\theta_1)\mathsf{d}\theta_1\Big|\\ &\le \int(\theta_*+|\theta_1|)\phi_\sigma(\theta_1)\mathsf{d}\theta_1\le\theta_*+\sigma.\end{aligned}$$

Step 2: In this step, we upper bound $\max_{x\in\{0,1,\dots,2k\}}|b_x|$. Let

$$\widetilde{r}(\theta):=p_k(\theta_*\theta)q_k(\theta_*\theta):=\sum_{x=0}^{2k}\widetilde{b}_x w(x)\theta^x$$

be a rescaled version of $p_k(\theta)q_k(\theta)$, so that $\widetilde{b}_x=\theta_*^x b_x$. Then by Lemma 14, it holds that for each $0\le x\le 2k$,

$$\begin{aligned}|\widetilde{b}_x|w(x)&\le\frac{(2k)^x}{x!}\sup_{|\theta|\le 1}|\widetilde{r}(\theta)|\\ &\le\frac{(2k)^x}{x!}\sup_{|\theta|\le\theta_*}p_k(\theta)\cdot\sup_{|\theta|\le\theta_*}q_k(\theta).\end{aligned}$$

Since

$$\sup_{|\theta|\le\theta_*}q_k(\theta)\le 1/g(\theta_*)$$

and by (16),

$$\sup_{|\theta|\le\theta_*}p_k(\theta)\le C$$

for some positive constant C only depending on θ_* and σ, it follows that

$$\begin{aligned}\max_{0\le x\le 2k}|b_x|&\le C_6\max_{0\le x\le 2k}\frac{(2k)^x}{w(x)\theta_*^x x!}\\ &\le C_6\max_{0\le x\le 2k}\frac{1}{w(x)}\cdot\max_{1\le x\le 2k}\frac{1}{\theta_*^x}\cdot\max_{0\le x\le 2k}\frac{(2k)^x}{x!}\end{aligned}$$

where $C_6=C_6(\theta_*,\sigma)>0$. Combining the above inequality with

$$\max_{0\le x\le 2k}1/\theta_*^x\le\Big(\max\{1,1/\theta_*\}\Big)^{2k}$$

and

$$\max_{0\le x\le 2k}(2k)^x/x!\le e^{2k},$$

it follows that

$$\max_{0\le x\le 2k}|b_x|\le C_6\cdot\Big(e\cdot\max\{1,1/\theta_*\}\Big)^{2k}\cdot\max_{0\le x\le 2k}\frac{1}{w(x)}.$$

Step 3: In this step we prove the claim of the theorem. Recall that

$$W_1^\sigma(\widehat{Q},Q)=\sup_\ell\int\ell\mathsf{d}[\widehat{Q}*\mathcal{N}_\sigma]-\ell\mathsf{d}[Q*\mathcal{N}_\sigma],$$

where $\ell\in\mathrm{Lip}(1)$ with $\ell(0)=0$. It further holds that

$$\begin{aligned}&W_1^\sigma(\widehat{Q},Q)\\ &=\sup_{\ell\in\mathrm{Lip}(1):\ell(0)=0}\int\ell\mathsf{d}[\widehat{Q}*\mathcal{N}_\sigma]-\ell\mathsf{d}[Q*\mathcal{N}_\sigma]\\ &=\sup_{\ell\in\mathrm{Lip}(1):\ell(0)=0}\int(\ell_\sigma(\theta)-\ell_\sigma(0))[\mathsf{d}\widehat{Q}-\mathsf{d}Q]\\ &=\sup_{\ell\in\mathrm{Lip}(1):\ell(0)=0}\int\Big\{\ell_\sigma(\theta)-\ell_\sigma(0)-\sum_{0\le x\le 2k}b_x f(x|\theta)\Big\}\\ &\quad[\mathsf{d}\widehat{Q}-\mathsf{d}Q]\\ &\quad+\sup_{\ell\in\mathrm{Lip}(1):\ell(0)=0}\int\sum_{0\le x\le 2k}b_x f(x|\theta)[\mathsf{d}\widehat{Q}-\mathsf{d}Q]\\ &:=(I)+(II).\end{aligned}$$

By Step 1, we have

$$(I)\le 2C_4\Big\{\Big[2\theta_*^{-1}\sigma\sqrt{ek}\Big]^{-k}+\sum_{x\ge k+1}w(x)\theta_*^x\Big\}.\tag{18}$$

Next we bound (II). Recall that

$$h^{\mathrm{obs}}(x):=\sum_{i=1}^n\mathbb{1}(X_i=x)/n.$$

We have

$$\begin{aligned}&\int\sum_{0\le x\le 2k}b_x f(x|\theta)[\mathsf{d}\widehat{Q}(\theta)-\mathsf{d}Q(\theta)]\\ &\le\Big|\sum_{0\le x\le 2k}b_x[h_{\widehat{Q}}(x)-h^{\mathrm{obs}}(x)]\Big|\\ &\quad+\Big|\sum_{0\le x\le 2k}b_x[h^{\mathrm{obs}}(x)-h_Q(x)]\Big|.\end{aligned}$$

It follows from Lemma 17 that for an arbitrary $\delta\in(0,1)$ and an arbitrary $\epsilon\in(0,1)$, there exists a constant $C_7=C_7(\epsilon,\theta_*)>0$ such that with probability at least $1-\delta$

$$\begin{aligned}&\Big|\sum_{0\le x\le 2k}b_x\left[h^{\mathrm{obs}}(x)-h_Q(x)\right]\Big|+\Big|\sum_{0\le x\le 2k}b_x\left[h^{\mathrm{obs}}(x)-h_{\widehat{Q}}(x)\right]\Big|\\ &\le C_7\max_{0\le x\le 2k}|b_x|\sqrt{\frac{1}{n^{1-\epsilon}\delta^{1+\epsilon}}}\end{aligned}$$

uniformly over all $\{b_x\}_{x=0}^{2k}$. Consequently, we have

$$(II)\le C_8\max_{0\le x\le 2k}|b_x|\Big/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}}$$

with probability at least $1-\delta$ for some constant $C_8=C_8(\epsilon,\theta_*)>0$. Note that $\max_{0\le x\le 2k}|b_x|$ has been upper bounded in Step 2.

Putting together the estimates for (I) and (II), we have that with probability at least $1-\delta$, $W_1^\sigma(Q,\widehat{Q})$ is upper bounded by

$$\begin{aligned}&\Big[2\theta_*^{-1}\sigma\sqrt{ek}\Big]^{-k}+\sum_{x\ge k+1}w(x)\theta_*^x\\ &+\Big(e\cdot\max\{1,1/\theta_*\}\Big)^{2k}\cdot\frac{\max_{1\le x\le 2k}1/w(x)}{\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}}}\end{aligned}\tag{19}$$

up to a constant depending on σ,θ_* and ϵ.

(i) If $1/w(x)\le C_1C_2^x$, (19) becomes

$$[2\theta_*^{-1}\sigma\sqrt{ek}]^{-k}+\sum_{x\ge k+1}w(x)\theta_*^x+C_9^{2k}/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}},$$

where $C_9 := e \cdot \max\{1, 1/\theta_*\} \cdot \max\{1, C_2\}$ is a positive constant. For the second term, it follows from Corollary 1.1.10 in [46] that for any $R \in (\theta_*, \theta_r)$ there exists some constant $C_{10} = C_{10}(R) > 0$ such that $w(x) \le C_{10}/R^x$ for all $x = 0, 1, 2 \ldots$, and hence

$$\sum_{x \ge k+1} w(x)\theta_*^x \le C_{10} \sum_{x \ge k+1} (\theta_*/R)^x \le C_{10} \cdot (\theta_*/[R - \theta_*]) \cdot [\theta_*/R]^k$$

for any $k = 1, 2, \ldots$. Therefore, the second term dominates the first term in (19), and (19) becomes

$$[\theta_*/R]^k + C_9^{2k}/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}}.$$

The proof is then complete by letting $C_9^{2k} = n^\alpha$ for some $\alpha \in (0, 1/2-\epsilon/2)$. The final bound is then $n^{-\frac{(1-\epsilon)\log(R/\theta_*)}{2\log(R/\theta_*)+4\log C_9}}$ for any $\epsilon \in (0, 1)$.

(ii) If $c_1 c_2^x x^{c_3 x} \le 1/w(x) \le C_1 C_2^x x^{C_3 x}$, (19) becomes

$$[2\theta_*^{-1}\sigma\sqrt{ek}]^{-k} + (C_{11}k)^{-c_3 k} + (C_{12}k)^{2C_3 k}/\sqrt{n^{1-\epsilon}\delta^{1+\epsilon}}$$

for some positive constants C_{11} and C_{12} and the proof is then complete by letting $(C_{12}k)^{2C_3 k} = n^\alpha$ for some $\alpha \in (0, 1/2 - \epsilon/2)$. The final bound is then $n^{-\frac{(1-\epsilon)/2}{1+\max\{4C_3, 2C_3/c_3\}}}$. □

E. Proof of Remark 10

Proof of Remark 10: (i) For the Poisson mixture, we have $1/w(x) = x!$ with $\sqrt{2\pi x}\,(x/e)^x e^{\frac{1}{12x+1}} < x! < \sqrt{2\pi x}(x/e)^x e^{\frac{1}{12x}}$ from Stirling's approximation. Therefore, it satisfies the assumption (ii) in Theorem 7 with $c_3 = C_3 = 1$. It then follows from the arguments in the end of the proof of Theorem 7 that the rate is

$$\frac{(1-\epsilon)/2}{1 + \max\{4C_3, 2C_3/c_3\}} = \frac{1-\epsilon}{10}.$$

(ii) For the negative binomial mixture with

$$f(x|\theta) = (1-\theta)^r \binom{x+r-1}{x} \theta^x, r > 0,$$

we have

$$1/w(x) = 1/\binom{x+r-1}{x} \le 1 \quad \text{and} \quad \theta_r = 1.$$

Therefore, it satisfies the assumption (i) in Theorem 7 with $C_2 = 1$. By taking

$$R = (\theta_* + \theta_r)/2 = (\theta_* + 1)/2,$$

it then follows from the arguments in the end of the proof of Theorem 7 that the rate is

$$\frac{(1-\epsilon)\log[(\theta_*+1)/(2\theta_*)]}{2\log[(\theta_*+1)/(2\theta_*)] + 4\log C_9} \le \frac{(1-\epsilon)}{2 + 4\log[C_9]/\log[1/\theta_*]}.$$

The proof is then completed by noting that $C_9 = e \cdot \max\{1, 1/\theta_*\} \cdot \max\{1, C_2\} = e/\theta_*$. □

V. Proofs of Auxiliary Results

A. Proof of Lemma 11

By rescaling, we assume that $a = -1$ and $b = 1$. For any integer $r \ge 1$, let

$$W_\infty^r([a,b]) := \Big\{\psi : [a,b] \to \mathbb{R} : \psi^{(r-1)} \text{ is absolutely continuous and the essential supremum of } \psi^{(r)} \text{ on } [a,b] \text{ is finite}\Big\}$$

be the Sobolev space on $[a, b]$. Then it is readily verifiable that for any $\ell \in \mathrm{Lip}(1)$ and $\sigma^2 > 0$,

$$\ell_\sigma(\theta) - \ell_\sigma(0) = (\ell * \phi_\sigma)(\theta) - (\ell * \phi_\sigma)(0),$$

when restricted on $[a, b]$, belongs to $W_\infty^r([a, b])$. Hence by Lemma 13, we have that for any integer $k > r$, there exists some polynomial p_k of degree k such that

$$\sup_{\theta \in [a,b]} \Big|\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)\Big| \le C_1 k^{-r} \omega\big(\ell_\sigma^{(r)}, k^{-1}\big),$$

In the above inequality, $C_1 = C_1(a, b) > 0$ is a constant and

$$\omega(\psi, t) := \sup_{\theta_1, \theta_2 : |\theta_1 - \theta_2| \le t} |\psi(\theta_1) - \psi(\theta_2)|$$

is the modulus of continuity of function ψ at radius t. To bound the righthand side of the above display, note that, with $H_n(\cdot)$ denoting the n-th Hermite polynomial (page 775 in [47]), we have

$$\begin{aligned}\ell_\sigma^{(r)}(\theta) &= \int \ell(\theta_1)\phi_\sigma^{(r)}(\theta - \theta_1)\mathrm{d}\theta_1 \\ &= \sigma^{-r}(-1)^r \int \ell(\theta - \theta_1)\phi_\sigma(\theta_1) H_r\big(\theta_1/\sigma\big)\mathrm{d}\theta_1,\end{aligned}$$

where the second equality follows from the derivative identity for the normal density (page 785 in [47]). Hence for any θ_1, θ_2 such that $|\theta_1 - \theta_2| \le k^{-1}$, we have

$$\begin{aligned}&\big|\ell_\sigma^{(r)}(\theta_1) - \ell_\sigma^{(r)}(\theta_2)\big| \\ &\le \sigma^{-r} \int |\ell(\theta_1 - \theta) - \ell(\theta_2 - \theta)|\phi_\sigma(\theta)|H_r(\theta/\sigma)|\mathrm{d}\theta \\ &\le \sigma^{-r} k^{-1} \int \phi_\sigma(\theta)|H_r(\theta/\sigma)|\mathrm{d}\theta \\ &= \sigma^{-r} k^{-1} \int \phi_1(\theta)|H_r(\theta)\big|\mathrm{d}\theta \\ &\le \sigma^{-r} k^{-1} [\int \phi_1(\theta) H_r^2(\theta)\mathrm{d}\theta]^{1/2} \\ &= \sigma^{-r} k^{-1}\sqrt{r!}.\end{aligned}$$

It further follows from the Stirling's formula $\sqrt{r!} \le \sqrt{e r^{r+1/2} e^{-r}}$ that

$$\big|\ell_\sigma^{(r)}(\theta_1) - \ell_\sigma^{(r)}(\theta_2)\big| \le \sigma^{-r} k^{-1}\sqrt{e r^{r+1/2} e^{-r}}.$$

Using $r < k$, we hence obtain

$$\begin{aligned}\sup_{\theta \in [a,b]} |\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)| &\le C_1\sqrt{e}(\sqrt{e}\sigma k/\sqrt{r})^{-r} r^{1/4} k^{-1} \\ &\le C_1\sqrt{e}(\sqrt{e}\sigma\sqrt{k})^{-r} k^{-3/4}.\end{aligned}$$

By rescaling, we then have for any $a \le 0, b \ge 0$ it follows that

$$\begin{aligned}&\sup_{\theta\in[a,b]} |\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)| \\ &\le C_1([b-a]/2)^{r+1}\sqrt{e}(\sqrt{e}\sigma\sqrt{k})^{-r}k^{-3/4} \\ &\le \frac{C_1(b-a)\sqrt{e}}{2}\cdot\Big[\frac{2\sqrt{e}\sigma\sqrt{k}}{b-a}\Big]^{-r}k^{-3/4}.\end{aligned}$$

Now taking $r = k-1$, we have

$$\sup_{\theta\in[a,b]} |\ell_\sigma(\theta) - \ell_\sigma(0) - p_k(\theta)| \le C_1 e\sigma\cdot\Big[\frac{2\sqrt{e}\sigma\sqrt{k}}{b-a}\Big]^{-k}k^{-1/4}$$

and accordingly complete the proof. □

B. Proof of Lemma 17

Whenever there is no ambiguity, let $h^{\rm obs}$, $h_{\widehat{Q}}$, and h_Q also represent distributions with respect to corresponding probability mass functions $x \mapsto h^{\rm obs}(x)$, $x \mapsto h_{\widehat{Q}}(x)$, and $x \mapsto h_Q(x)$. This proof consists of two steps. In the first step, we prove that both

$$\left|\sum_{x=0}^{\infty} b_x\left(h^{\rm obs}(x) - h_Q(x)\right)\right| \text{ and } \left|\sum_{x=0}^{\infty} b_x\left(h^{\rm obs}(x) - h_{\widehat{Q}}(x)\right)\right|$$

can be upper bounded by $\mathrm{KL}(h^{\rm obs}, h_Q)$, where KL is the Kullback-Leibler divergence. In the second step, we upper bound $\mathrm{KL}(h^{\rm obs}, h_Q)$ by truncation arguments.

Step 1: It follows from the triangle inequality and Pinsker's inequality that

$$\begin{aligned}\left|\sum_{x=0}^{\infty} b_x\left(h^{\rm obs}(x) - h_Q(x)\right)\right| &\le \max_{x\ge 0}|b_x|\cdot\left\|h^{\rm obs} - h_Q\right\|_1 \\ &\le \max_{x\ge 0}|b_x|\sqrt{\frac{1}{2}\cdot\mathrm{KL}(h^{\rm obs}, h_Q)},\end{aligned}$$

where $\|h^{\rm obs} - h_Q\|_1$ represents the total variation distance between distributions $h^{\rm obs}$ and h_Q. Analogously, we have

$$\begin{aligned}\left|\sum_{x=0}^{\infty} b_x\left(h^{\rm obs}(x) - h_{\widehat{Q}}(x)\right)\right| &\le \max_{x\ge 0}|b_x|\sqrt{\frac{1}{2}\cdot\mathrm{KL}(h^{\rm obs}, h_{\widehat{Q}})} \\ &\le \max_{x\ge 0}|b_x|\sqrt{\frac{1}{2}\cdot\mathrm{KL}(h^{\rm obs}, h_Q)},\end{aligned}$$

where the last inequality follows from the fact that, by definition,

$$\begin{aligned}\widehat{Q} &= \operatorname*{argmax}_{\widetilde{Q}}\sum_{i=1}^{n}\log h_{\widetilde{Q}}(X_i) \\ &= \operatorname*{argmax}_{\widetilde{Q}}\sum_{x=0}^{\infty}h^{\rm obs}(x)\log h_{\widetilde{Q}}(x) = \operatorname*{argmin}_{\widetilde{Q}}\mathrm{KL}(h^{\rm obs}, h_{\widetilde{Q}}).\end{aligned}$$

Step 2: Suppose $C_1 = C_1(\theta_*)$ is the smallest positive integer larger than $\theta_* g(0)(1/g)'(\theta_*)$. Define

$$T_i := X_i\mathbb{1}(X_i \le C_1 - 1) + C_1\mathbb{1}(X_i \ge C_1) \text{ for all } i \in [n].$$

Let t_Q be the probability mass function of T_1 and let $t^{\rm obs}$ be the sample version of t_Q, i.e.

$$x \mapsto t_Q(x) := P(T_1 = x) \text{ and } x \mapsto t^{\rm obs}(x) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(T_i = x),$$

for $x \in \{0, \ldots, C_1\}$. Note that

$$t_Q(x) = h_Q(x) \text{ and } t^{\rm obs}(x) = h^{\rm obs}(x) \text{ for } x = 0, \ldots, C_1 - 1$$

and

$$t_Q(C_1) = \sum_{x\ge C_1} h_Q(x), \quad t^{\rm obs}(C_1) = \sum_{x\ge C_1} h^{\rm obs}(x).$$

Hence it follows that

$$\begin{aligned}&\mathrm{KL}(h^{\rm obs}, h_Q) \\ &= \sum_{x=0}^{C_1-1} t^{\rm obs}(x)\log\frac{t^{\rm obs}(x)}{t_Q(x)} + \sum_{x\ge C_1} h^{\rm obs}(x)\log\frac{h^{\rm obs}(x)}{h_Q(x)} \\ &= \mathrm{KL}(t^{\rm obs}, t_Q) - t^{\rm obs}(C_1)\log\frac{t^{\rm obs}(C_1)}{t_Q(C_1)} \\ &\quad + \sum_{x\ge C_1} h^{\rm obs}(x)\log\frac{h^{\rm obs}(x)}{h_Q(x)},\end{aligned}$$

where $t^{\rm obs}$ and t_Q are viewed as distributions with respect to corresponding probability mass functions of $x \mapsto t_Q(x)$ and $x \mapsto t^{\rm obs}(x)$.

If $t^{\rm obs}(C_1) = 0$, then

$$t^{\rm obs}(C_1)\log\frac{t^{\rm obs}(C_1)}{t_Q(C_1)} = 0.$$

Otherwise it follows from the inequality

$$\log(1+x) \le x \text{ for } x > 0$$

that

$$-t^{\rm obs}(C_1)\log\frac{t^{\rm obs}(C_1)}{t_Q(C_1)} \le \sum_{x\ge C_1}\Big\{h_Q(x) - h^{\rm obs}(x)\Big\}.$$

Analogously, we have

$$\begin{aligned}&\sum_{x\ge C_1} h^{\rm obs}(x)\log\frac{h^{\rm obs}(x)}{h_Q(x)} \\ &\le \sum_{x\ge C_1}\frac{(h^{\rm obs}(x) - h_Q(x))^2}{h_Q(x)} + \sum_{x\ge C_1}\Big\{h^{\rm obs}(x) - h_Q(x)\Big\}\end{aligned}$$

and hence

$$\begin{aligned}&-t^{\rm obs}(C_1)\log\frac{t^{\rm obs}(C_1)}{t_Q(C_1)} + \sum_{x\ge C_1} h^{\rm obs}(x)\log\frac{h^{\rm obs}(x)}{h_Q(x)} \\ &\le \sum_{x\ge C_1}\frac{(h^{\rm obs}(x) - h_Q(x))^2}{h_Q(x)}.\end{aligned}$$

Step 2(a): We first upper bound $\sum_{x\ge C_1}(h^{\rm obs}(x) - h_Q(x))^2/h_Q(x)$. Fix an arbitrary $\epsilon \in (0,1)$ and choose a $\gamma > 0$ in $(1-\epsilon, 1)$. Define $A := \alpha^{(1-\gamma)/3}$, where $\alpha := (\theta_* + \theta_r)/(2\theta_*) > 1$. Note that $\alpha\theta_* < \theta_r$ and we have

$1/g(\theta) = \sum_{x=0}^{\infty} w(x)\theta^x < \infty$ for all $\theta \in [0, \alpha\theta_*]$. It then follows from Hölder's inequality that

$$\begin{aligned} & n^{1-\epsilon} \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} \\ &= n^{1-\epsilon} \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{-x} A^x \\ &\leq n^{1-\epsilon} \left(\sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{-x/\gamma} \right)^{\gamma} \\ &\quad \cdot \left(\sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{x/(1-\gamma)} \right)^{1-\gamma}. \end{aligned}$$

It further follows from $A > 1$ that

$$\begin{aligned} & n \cdot \mathbb{E}\Big\{ \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{-x/\gamma} \Big\} \\ &= \sum_{x\geq C_1} (1 - h_Q(x)) A^{-x/\gamma} \\ &\leq \sum_{x\geq C_1} A^{-x/\gamma} = \frac{A^{-C_1/\gamma}}{1 - A^{-1/\gamma}} < \infty \end{aligned}$$

and hence for an arbitrary $\delta \in (0,1)$, we have

$$n \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{-x/\gamma} \leq \frac{A^{-C_1/\gamma}}{1 - A^{-1/\gamma}} \frac{1}{\delta}$$

with probability at least $1 - \delta$. Therefore, with probability at least $1 - \delta$, we have

$$\begin{aligned} & \left(\sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{-x/\gamma} \right)^{\gamma} \\ &\leq \left(\frac{A^{-C_1/\gamma}}{1 - A^{-1/\gamma}} \frac{1}{n\delta} \right)^{\gamma} \leq \frac{1}{\left(\alpha^{\frac{1-\gamma}{3\gamma}} - 1\right)^{\gamma}} \frac{1}{(n\delta)^{\gamma}}, \end{aligned}$$

where the last inequality follows from $C_1 \geq 1$. On the other hand,

$$\begin{aligned} & \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{x/(1-\gamma)} \\ &\leq 2 \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x))^2}{h_Q(x)} \alpha^{x/3} + 2 \sum_{x\geq C_1} h_Q(x) \alpha^{x/3}. \end{aligned}$$

We first show that the second term on the righthand side is bounded, which is true if

$$\sum_{x\geq C_1} h_Q(x)\alpha^x \leq g(\theta_*) \Big/ g(\alpha\theta_*).$$

Since $\sum_{x=0}^{\infty} g(\theta) w(x) \theta^x = 1$ and $1/g(\theta) = \sum_{x=0}^{\infty} w(x)\theta^x$, it follows from

$$(1/g)'(\theta) = \sum_{x=1}^{\infty} x w(x) \theta^{x-1} > 0$$

and

$$(1/g)''(\theta) = \sum_{x=2}^{\infty} x(x-1) w(x) \theta^{x-2} > 0$$

that $g(\cdot)$ is monotonically decreasing on $[0, \theta_*]$ and $(1/g)'(\cdot)$ is monotonically increasing on $[0, \theta_*]$. Therefore, it follows from

$$\log f(x|\theta) = -\log(1/g(\theta)) + x \log\theta + \log w(x)$$

that

$$\begin{aligned} \frac{\mathsf{d}(\log f(x|\theta))}{\mathsf{d}\theta} &= \frac{1}{\theta} \left(x - \theta g(\theta) (1/g)'(\theta) \right) \\ &\geq \frac{1}{\theta} \left(x - \theta_* g(0) (1/g)'(\theta_*) \right) \geq \frac{1}{\theta} (x - C_1) \geq 0 \end{aligned}$$

for all $x \geq C_1$. Therefore we have

$$h_Q(x) = \int_0^{\theta_*} f(x|\theta) \mathsf{d}Q(\theta) \leq \sup_{\theta\in[0,\theta_*]} f(x|\theta) = f(x|\theta_*)$$

and

$$\begin{aligned} \sum_{x\geq C_1} h_Q(x)\alpha^x &\leq \sum_{x\geq C_1} f(x|\theta_*)\alpha^x \\ &\leq \sum_{x\geq 0} f(x|\theta_*)\alpha^x = \frac{g(\theta_*)}{g(\alpha\theta_*)} < \infty. \end{aligned}$$

We now switch to the first term. For any fixed $k > 0$, define A_n to be the event

$$A_n := \Big\{ h^{\text{obs}}(x) > k h_Q(x) \alpha^{x/3} \text{ for some } x \geq C_1 \Big\}.$$

Then, it follows from Markov's inequality that

$$\begin{aligned} P(A_n) &\leq \sum_{x\geq C_1} P(h^{\text{obs}}(x) > k h_Q(x) \alpha^{x/3}) \\ &\leq \frac{1}{k} \sum_{x\geq C_1} \mathbb{E}\{h^{\text{obs}}(x)\} \frac{1}{h_Q(x)\alpha^{x/3}} \leq \frac{1}{k} \frac{1}{\alpha^{1/3} - 1}. \end{aligned}$$

Thus, $P(A_n)$ can be made arbitrarily small by choosing k large enough and on the complement of A_n we have

$$\sum_{x\geq C_1} \frac{(h^{\text{obs}}(x))^2}{h_Q(x)} \alpha^{x/3} \leq k^2 \sum_{x\geq C_1} h_Q(x)\alpha^x \leq k^2 \frac{g(\theta_*)}{g(\alpha\theta_*)}.$$

Therefore, for an arbitrary $\delta \in (0,1)$, we have

$$\sum_{x\geq C_1} \frac{(h^{\text{obs}}(x))^2}{h_Q(x)} \alpha^{x/3} \leq \frac{g(\theta_*)}{g(\alpha\theta_*)} \left(\frac{1}{\delta} \frac{1}{\alpha^{1/3} - 1} \right)^2$$

with probability at least $1 - \delta$. Thus, for an arbitrary $\delta \in (0,1)$, with probability at least $1 - \delta$, it follows that

$$\begin{aligned} & \Big\{ \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} A^{\frac{x}{1-\gamma}} \Big\}^{1-\gamma} \\ &\leq 2^{1-\gamma} \cdot \Big\{ \frac{g(\theta_*)}{g(\alpha\theta_*)} \left(\frac{1}{\delta} \frac{1}{\alpha^{1/3} - 1} \right)^2 + \frac{g(\theta_*)}{g(\alpha\theta_*)} \Big\}^{1-\gamma} \leq \frac{C_2}{\delta^{2-2\gamma}}, \end{aligned}$$

where $C_2 = C_2(\theta_*) = \{2 \cdot g(\theta_*)[1/(\alpha^{1/3} - 1)^2 + 1]/g(\alpha\theta_*)\}^{1-\gamma}$ is a constant. For an arbitrary $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$, it follows that

$$\begin{aligned} & n^{1-\epsilon} \sum_{x\geq C_1} \frac{(h^{\text{obs}}(x) - h_Q(x))^2}{h_Q(x)} \\ &\leq n^{1-\epsilon} \frac{1}{\left(\alpha^{\frac{1-\gamma}{3\gamma}} - 1\right)^{\gamma}} \frac{1}{(n\delta)^{\gamma}} \frac{C_2}{\delta^{2-2\gamma}} \\ &= n^{1-\epsilon-\gamma} \frac{C_2}{\left(\alpha^{\frac{1-\gamma}{3\gamma}} - 1\right)^{\gamma}} \frac{1}{\delta^{2-\gamma}}. \end{aligned}$$

Thus, by letting γ go to $1-\epsilon$, we have

$$\sum_{x\geq C_1}\frac{(h^{\rm obs}(x)-h_Q(x))^2}{h_Q(x)}\leq\frac{C_2}{\left(\alpha^{\frac{\epsilon}{3(1-\epsilon)}}-1\right)^{1-\epsilon}}\frac{1}{n^{1-\epsilon}}\frac{1}{\delta^{1+\epsilon}}.$$

As a result, for arbitrary $\delta\in(0,1/2)$ and $\epsilon\in(0,1)$, with probability at least $1-2\delta$, we have

$$\begin{aligned}\mathrm{KL}(h^{\rm obs},h_Q)&\leq\mathrm{KL}(t^{\rm obs},t_Q)+\sum_{x\geq C_1}\frac{(h^{\rm obs}(x)-h_Q(x))^2}{h_Q(x)}\\&\leq\mathrm{KL}(t^{\rm obs},t_Q)+C_3\frac{1}{n^{1-\epsilon}}\frac{1}{\delta^{1+\epsilon}},\end{aligned}$$

where $C_3=C_3(\epsilon,\theta_*)=C_2/(\alpha^{\frac{\epsilon}{3(1-\epsilon)}}-1)^{1-\epsilon}$.

Step 2(b): We then upper bound $\mathrm{KL}(t^{\rm obs},t_Q)$. It follows from [48] that with probability at least $1-\delta$,

$$\mathrm{KL}(t^{\rm obs},t_Q)\leq\frac{C_1+1}{2n}\log\frac{4n}{C_1+1}+\frac{1}{n}\log\frac{3e}{\delta},$$

and hence for any $\epsilon\in(0,1)$ and $\delta\in(0,1/3)$, with probability at least $1-3\delta$,

$$\mathrm{KL}(h^{\rm obs},h_Q)\leq\frac{1}{n\delta^{1+\epsilon}}\left(3C_1\log(2n)+C_3n^{\epsilon}\right).$$

Therefore, it follows that there exists a constant $C_4=C_4(\epsilon,\theta_*)$ such that for any $n\geq1$

$$\mathrm{KL}(h^{\rm obs},h_Q)\leq\frac{C_4}{n^{1-\epsilon}\delta^{1+\epsilon}}$$

holds with probability at least $1-3\delta$ for any $\epsilon\in(0,1)$ and $\delta\in(0,1/3)$. Therefore,

$$\left|\sum_{x=0}^{\infty}b_x\left(h^{\rm obs}(x)-h_{\widehat{Q}}(x)\right)\right|\leq\max_{x\geq0}|b_x|\sqrt{\frac{C_4}{2n^{1-\epsilon}\delta^{1+\epsilon}}}$$

holds for all $n\geq n_1$ with probability at least $1-3\delta$ for any $\epsilon\in(0,1)$ and $\delta\in(0,1/3)$. □

C. Proof of Lemma 12

This proof consists of two steps. In the first step, we prove the existence of $\widehat{\ell}$ and upper bound $\sup_{\theta\in[0,\theta_*]}|\widehat{\ell}(\theta)-\ell(\theta)|$. In the second step, we upper bound coefficients of $\widehat{\ell}$, i.e., $\max_{0\leq x\leq k}|b_x|$.

Step 1: It follows from $\sum_{x=0}^{\infty}f(x|\theta)=1$ that $\sum_{x=0}^{\infty}g(\theta)w(x)\theta^x=1$ and hence $g(\theta)>0$ for $\theta\in[0,\theta_*]$. As a consequence, $1/g(\theta)=\sum_{x=0}^{\infty}w(x)\theta^x$ on $[0,\theta_*]$.

Since $\theta\mapsto\sum_{x=0}^{\infty}w(x)\theta^x$ is a continuous function on $[-\theta_*,\theta_*]$ taking the value $w(0)>0$ at $\theta=0$ (recall the convention $0^0=1$), there exists some $\theta_0\in(0,\theta_*]$ such that $\theta\mapsto\sum_{x=0}^{\infty}w(x)\theta^x$ is strictly positive on $[-\theta_0,\theta_*]$. For $\theta\in[-\theta_0,0)$, define $1/g(\theta):=\sum_{x=0}^{\infty}w(x)\theta^x$ and $\ell(\theta):=-\ell(-\theta)$. Then $\theta\mapsto\ell(\theta)$ is a 1-Lipschitz function on $[-\theta_0,\theta_*]$ and for any $\theta_1,\theta_2\in[-\theta_0,\theta_*]$, we have

$$\begin{aligned}&|\ell(\theta_1)/g(\theta_1)-\ell(\theta_2)/g(\theta_2)|\\&\leq|\ell(\theta_1)/g(\theta_1)-\ell(\theta_2)/g(\theta_1)|+|\ell(\theta_2)/g(\theta_1)-\ell(\theta_2)/g(\theta_2)|\\&\leq|\theta_1-\theta_2|\{1/g(\theta_*)+\theta_*(1/g)'(\theta_*)\},\end{aligned}$$

using the fact that both $\theta\mapsto1/g(\theta)$ and $\theta\mapsto(1/g)'(\theta)$ achieve their maxima at θ_*. This implies $\theta\mapsto\ell(\theta)/g(\theta)$ is Lipschitz with constant $1/g(\theta_*)+\theta_*(1/g)'(\theta_*)$. Therefore, it follows from Jackson's theorem (see Lemma 15) that there exists a polynomial $\sum_{x=0}^{k}v_x\theta^x$ of degree $k\geq1$ such that

$$\sup_{\theta\in[-\theta_0,\theta_*]}|\ell(\theta)/g(\theta)-\sum_{x=0}^{k}v_x\theta^x|\leq C_1/k,$$

where $C_1>0$ is independent of k and ℓ and $v_x\in\mathbb{R}$ for all $x=0,\ldots,k$. Let $b_x:=v_x/w(x)$ for $0\leq x\leq k$ (this is well-defined since $w(x)>0$ for $x\in\mathbb{Z}$). Then with $\widehat{\ell}(\theta):=\sum_{x=0}^{k}b_xg(\theta)w(x)\theta^x=\sum_{x=0}^{k}b_xf(x|\theta)$, it holds that

$$\sup_{\theta\in[-\theta_0,\theta_*]}\left|\ell(\theta)-\widehat{\ell}(\theta)\right|\leq\frac{C_1}{k}\cdot\sup_{\theta\in[-\theta_0,\theta_*]}g(\theta)\leq\frac{C_2}{k},$$

where $C_2>0$ does not depend on k and ℓ.

Step 2: To bound the coefficients $\{b_x\}_{0\leq x\leq k}$, we first define a polynomial

$$\theta\mapsto r(\theta):=\sum_{x=0}^{k}v_x(\theta_0\theta)^x\ \text{ on }[-1,1]$$

and note that

$$\begin{aligned}\sup_{\theta\in[-1,1]}|r(\theta)|&=\sup_{\theta\in[-\theta_0,\theta_0]}\Big|\sum_{x=0}^{k}v_x\theta^x\Big|\\&\leq C_1/k+\sup_{\theta\in[-\theta_0,\theta_0]}|\ell(\theta)/g(\theta)|\\&\leq C_1/k+\theta_0/g(\theta_0),\end{aligned}$$

using the fact that $\theta\mapsto1/g(\theta)$ achieves its maximum at θ_0. We then apply Lemma 14 on the polynomial $r(\theta)$, and it follows that

$$|v_x|\theta_0^x\leq\max_{|\theta|\leq1}|r(\theta)|\cdot k^x/x!\leq C_3k^x/x!,$$

where $C_3>0$ does not depend on k and ℓ. Hence

$$|b_x|=|v_x|/w(x)\leq C_3\cdot(k/\theta_0)^x/\big(x!w(x)\big)$$

and

$$\begin{aligned}\max_{x\in[0,k]}|b_x|&\leq C_3\cdot\max_{x\in[0,k]}\frac{(k/\theta_0)^x}{x!w(x)}\\&\leq C_4\cdot\frac{(k/\theta_0)^k}{k!}\cdot\max_{0\leq x\leq k}1/w(x)\\&\leq C_4\cdot(e/\theta_0)^k\cdot\max_{0\leq x\leq k}1/w(x),\end{aligned}$$

where $C_4>0$ does not depend on k and ℓ. It follows from Corollary 1.1.10 in [46] that $w(x)\leq C_5/\theta_*^x$ for all $x\in\mathbb{N}$ and some constant $C_5\geq1$ depending only on $w(\cdot)$ and θ_* and hence

$$(e/\theta_0)^k/w(k)\geq(\theta_*e/\theta_0)^k/C_5\geq e^k/C_5>1$$

for all sufficiently large k. □

ACKNOWLEDGMENT

The authors would like to thank the associate editor and two anonymous referees for their helpful comments.

REFERENCES

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

[2] B. G. Lindsay, "Mixture models: Theory, geometry and applications," in *Proc. NSF-CBMS Regional Conf. Ser. Probab. Statist.*, 1995, pp. 1–163.

[3] C.-H. Zhang, "On estimating mixing densities in discrete exponential family models," *Ann. Statist.*, vol. 23, no. 3, pp. 929–945, Jun. 1995.

[4] L. Simar, "Maximum likelihood estimation of a compound Poisson process," *Ann. Statist.*, vol. 4, no. 6, pp. 1200–1209, Nov. 1976.

[5] H. G. Tucker, "An estimate of the compounding distribution of a compound Poisson distribution," *Theory Probab. Appl.*, vol. 8, no. 2, pp. 195–200, Jan. 1963.

[6] W.-L. Loh and C.-H. Zhang, "Global properties of kernel estimators for mixing densities in discrete exponential family models," *Statistica Sinica*, vol. 6, no. 3, pp. 561–578, 1996.

[7] W. Loh and C. Zhang, "Estimating mixing densities in exponential family models for discrete variables," *Scandin. J. Statist.*, vol. 24, no. 1, pp. 15–32, Mar. 1997.

[8] G. G. Walter and G. G. Hamedani, "Bayes empirical Bayes estimation for natural exponential families with quadratic variance functions," *Ann. Statist.*, vol. 19, no. 3, pp. 1191–1224, Sep. 1991.

[9] N. W. Hengartner, "Adaptive demixing in Poisson mixture models," *Ann. Statist.*, vol. 25, no. 3, pp. 917–928, Jun. 1997.

[10] F. Roueff and T. Rydén, "Nonparametric estimation of mixing densities for discrete distributions," *Ann. Statist.*, vol. 33, no. 5, pp. 2066–2108, Oct. 2005.

[11] J. Chen, "Consistency of the MLE under mixture models," *Stat. Sci.*, vol. 32, no. 1, pp. 47–63, Feb. 2017.

[12] F. d. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1879–1892, Mar. 2018.

[13] P. Rigollet and J. Weed, "Uncoupled isotonic regression via minimum Wasserstein deconvolution," *Inf. Inference, A, J. IMA*, vol. 8, no. 4, pp. 691–717, Dec. 2019.

[14] S. Jana, Y. Polyanskiy, and Y. Wu, "Extrapolating the profile of a finite population," in *Proc. Conf. Learn. Theory*, 2020, pp. 2011–2033.

[15] Z. Miao, W. Kong, R. K. Vinayak, W. Sun, and F. Han, "Fisher–Pitman permutation tests based on nonparametric Poisson mixtures with application to single cell genomics," *J. Amer. Stat. Assoc.*, pp. 1–13, Nov. 2022.

[16] R. M. Dudley, "The speed of mean Glivenko–Cantelli convergence," *Ann. Math. Statist.*, vol. 40, no. 1, pp. 40–50, Feb. 1969.

[17] S. Dereich, M. Scheutzow, and R. Schottstedt, "Constructive quantization: Approximation by empirical measures," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 49, no. 4, pp. 1183–1203, Nov. 2013.

[18] E. Boissard and T. Le Gouic, "On the mean speed of convergence of empirical and occupation measures in Wasserstein distance," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 50, no. 2, pp. 539–563, May 2014.

[19] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probab. Theory Rel. Fields*, vol. 162, nos. 3–4, pp. 707–738, Aug. 2015.

[20] J. Weed and F. Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, vol. 25, no. 4A, pp. 2620–2648, Nov. 2019.

[21] Z. Goldfeld et al., "Estimating information flow in deep neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2299–2308.

[22] Z. Goldfeld and K. Greenewald, "Gaussian-smoothed optimal transport: Metric structure and statistical efficiency," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3327–3337.

[23] R. Sadhu, Z. Goldfeld, and K. Kato, "Limit distribution theory for the smooth 1-Wasserstein distance with applications," 2021, *arXiv:2107.13494*.

[24] Y. Zhang, X. Cheng, and G. Reeves, "Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2422–2430.

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

[26] D. Jackson, "The general theory of approximation by polynomials and trigonometric sums," *Bull. Amer. Math. Soc.*, vol. 27, no. 9, pp. 415–431, 1921.

[27] Y. Wu and P. Yang, "Optimal estimation of Gaussian mixtures via denoised method of moments," *Ann. Statist.*, vol. 48, no. 4, pp. 1981–2007, Aug. 2020.

[28] K. Tian, W. Kong, and G. Valiant, "Learning populations of parameters," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5778–5787.

[29] R. K. Vinayak, W. Kong, G. Valiant, and S. Kakade, "Maximum likelihood estimation for learning populations of parameters," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6448–6457.

[30] J. Stoyanov and G. D. Lin, "Mixtures of power series distributions: Identifiability via uniqueness in problems of moments," *Ann. Inst. Stat. Math.*, vol. 63, no. 2, pp. 291–303, Apr. 2011.

[31] A. Noack, "A class of random variables with discrete distributions," *Ann. Math. Statist.*, vol. 21, no. 1, pp. 127–132, Mar. 1950.

[32] R. C. Gupta, "Estimating the probability of winning (losing) in a gambler's ruin problem with applications," *J. Stat. Planning Inference*, vol. 9, no. 1, pp. 55–62, Jan. 1984.

[33] K. G. Janardan, "A new discrete exponential family of distributions: Properties and application to power series distributions," *Amer. J. Math. Manag. Sci.*, vol. 2, no. 2, pp. 145–158, Feb. 1982.

[34] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[35] M.-K. Von Renesse and K.-T. Sturm, "Transport inequalities, gradient estimates, entropy and Ricci curvature," *Commun. Pure Appl. Math.*, vol. 58, no. 7, pp. 923–940, 2005.

[36] D. Lambert and L. Tierney, "Asymptotic properties of maximum likelihood estimates in the mixed Poisson model," *Ann. Statist.*, vol. 12, no. 4, pp. 1388–1399, Dec. 1984.

[37] C.-H. Zhang, "Empirical Bayes and compound estimation of normal means," *Statistica Sinica*, vol. 7, no. 1, pp. 181–193, 1997.

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

[39] M. Zhang, S. Liu, Z. Miao, F. Han, R. Gottardo, and W. Sun, "IDEAS: Individual level differential expression analysis for single-cell RNA-seq data," *Genome Biol.*, vol. 23, no. 1, pp. 1–17, Jan. 2022.

[40] M. Pensky, "Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity," *Ann. Statist.*, vol. 45, no. 4, pp. 1516–1541, Aug. 2017.

[41] R. A. DeVore, "Degree of approximation," *Approximation Theory II*, vol. 241, no. 242, pp. 117–161, 1976.

[42] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*. Amsterdam, The Netherlands: Elsevier, 2014.

[43] Y. Han and K. Shiragur, "On the competitive analysis and high accuracy optimality of profile maximum likelihood," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2021, pp. 1317–1336.

[44] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics). New York, NY, USA: Springer, 2009, doi: 10.1007/b13794.

[45] S. Ghosal and A. Van Der Vaart, *Fundamentals of Nonparametric Bayesian Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[46] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*. Berlin, Germany: Springer, 2002.

[47] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, vol. 55. Washington, DC, USA: U.S. Government Printing Office, 1964.

[48] J. Mardia, J. Jiao, E. Tánczos, R. D. Nowak, and T. Weissman, "Concentration inequalities for the empirical distribution of discrete distributions: Beyond the method of types," *Inf. Inference, A, J. IMA*, vol. 9, no. 4, pp. 813–850, Dec. 2020.

Fang Han received the Ph.D. degree from the Department of Biostatistics, Johns Hopkins University, in 2015. He is currently an Associate Professor of statistics and economics at the University of Washington. His main research interests are in rank- and graph-based statistics, statistical optimal transport, mixture models, nonparametric and semi-parametric regressions, time series analysis, and random matrix theory.

Zhen Miao received the B.S. degree in statistics from the University of Science and Technology of China in 2017 and the M.S. and Ph.D. degrees in statistics from the University of Washington in 2019 and 2023, respectively. He is currently with Microsoft Corporation.

Yandi Shen received the bachelor's degree in mathematics from Zhejiang University, China, in 2016, and the Ph.D. degree in statistics from the University of Washington, Seattle, WA, USA, in 2021. He is currently a William H. Kruskal Instructor with the Department of Statistics, The University of Chicago. He is broadly interested in nonparametric and semi-parametric statistics, high dimensional inference, and applied probability.