

# Minimax Estimation of Discrete Distributions under $\ell_1$ Loss

Yanjun Han, *Student Member, IEEE*, Jiantao Jiao, *Student Member, IEEE*, and Tsachy Weissman, *Fellow, IEEE*.

**Abstract**—We consider the problem of discrete distribution estimation under  $\ell_1$  loss. We provide tight upper and lower bounds on the maximum risk of the empirical distribution (the maximum likelihood estimator), and the minimax risk in regimes where the support size  $S$  may grow with the number of observations  $n$ . We show that among distributions with bounded entropy  $H$ , the asymptotic maximum risk for the empirical distribution is  $2H/\ln n$ , while the asymptotic minimax risk is  $H/\ln n$ . Moreover, we show that a hard-thresholding estimator oblivious to the unknown upper bound  $H$ , is essentially minimax. However, if we constrain the estimates to lie in the simplex of probability distributions, then the asymptotic minimax risk is again  $2H/\ln n$ . We draw connections between our work and the literature on density estimation, entropy estimation, total variation distance ( $\ell_1$  divergence) estimation, joint distribution estimation in stochastic processes, normal mean estimation, and adaptive estimation.

**Index Terms**—Distribution estimation, entropy estimation, minimax risk, hard-thresholding, high dimensional statistics

## I. INTRODUCTION AND MAIN RESULTS

Given  $n$  independent samples from an unknown discrete probability distribution  $P = (p_1, p_2, \dots, p_S)$ , with *unknown* support size  $S$ , we would like to estimate the distribution  $P$  under  $\ell_1$  loss. Equivalently, the problem is to estimate  $P$  based on the Multinomial random vector  $(X_1, X_2, \dots, X_S) \sim \text{Multi}(n; p_1, p_2, \dots, p_S)$ .

A natural estimator of  $P$  is the Maximum Likelihood Estimator (MLE), which in this problem setting coincides with the empirical distribution  $P_n$ , where  $P_n(i) = X_i/n$  is the number of occurrences of symbol  $i$  in the sample divided by the sample size  $n$ . This paper is devoted to analyzing the performances of the MLE, and the minimax estimators in various regimes. Specifically, we focus on the following three regimes:

- 1) Classical asymptotics: the dimension  $S$  of the unknown parameter  $P$  remains fixed, while the number of observations  $n$  grows.

Manuscript received Month 00, 0000; revised Month 00, 0000; accepted Month 00, 0000. Date of current version Month 00, 0000. This work was supported in part by the Center for Science of Information (CSol), an NSF Science and Technology Center, under grant agreement CCF-0939370. The material in this paper was presented in part at the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA. Copyright (c) 2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Yanjun Han, Jiantao Jiao and Tsachy Weissman are with the Department of Electrical Engineering, Stanford University, CA, USA. Email: {yjhan, jiantao, tsachy}@stanford.edu

This work was supported in part by the NSF Center for Science of Information under grant agreement CCF-0939370.

- 2) High dimensional asymptotics: we let the support size  $S$  and the number of observations  $n$  grow together, characterize the scaling under which consistent estimation is feasible, and obtain the minimax rates.
- 3) Infinite dimensional asymptotics: the distribution  $P$  may have infinite support size, but is constrained to have bounded entropy  $H(P) \leq H$ , where the entropy [1] is defined as

$$H(P) \triangleq \sum_{i=1}^S -p_i \ln p_i. \quad (1)$$

We remark that results for the first regime follow from the well-developed theory of asymptotic statistics [2, Chap. 8], and we include them here for completeness and comparison with other regimes. One motivation for considering the high dimensional and infinite dimensional asymptotics is that the modern era of big data gives rise to situations in which we cannot assume that the number of observations is much larger than the dimension of the unknown parameter. It is particularly true for the distribution estimation problem, e.g., the Wikipedia page on the Chinese characters showed that the number of Chinese sinograms is at least 80,000. Meanwhile, for distributions with extremely large support sizes (such as the Chinese language), the number of frequent symbols are considerably smaller than the support size. This observation motivates the third regime, in which we focus on distributions with finite entropy, but possibly extremely large support sizes. Another key result that motivates the problem of discrete distribution estimation under bounded entropy constraint is Marton and Shields [3], who essentially showed that the entropy rate dictates the difficulty in estimating discrete distributions under  $\ell_1$  loss in stochastic processes.

We denote by  $\mathcal{M}_S$  the set of all distributions of support size  $S$ . The  $\ell_1$  loss for estimating  $P$  using  $Q$  is defined as

$$\|P - Q\|_1 \triangleq \sum_{i=1}^S |p_i - q_i|, \quad (2)$$

where  $Q$  is not necessarily a probability mass function. The risk function for an estimator  $\hat{P}$  in estimating  $P$  under  $\ell_1$  loss is defined as

$$R(P; \hat{P}) \triangleq \mathbb{E}_P \|\hat{P} - P\|_1, \quad (3)$$

where the expectation is taken with respect to the measure  $P$ . The maximum  $\ell_1$  risk of an estimator  $\hat{P}$ , and the minimax

risk in estimating  $P$  are respectively defined as

$$R_{\text{maximum}}(\mathcal{P}; \hat{P}) \triangleq \sup_{P \in \mathcal{P}} R(P; \hat{P}) \quad (4)$$

$$R_{\text{minimax}}(\mathcal{P}) \triangleq \inf_{\hat{P}} \sup_{P \in \mathcal{P}} R(P; \hat{P}), \quad (5)$$

where  $\mathcal{P}$  is a given collection of probability measures  $P$ , and the infimum is taken over all estimators  $\hat{P}$ .

This paper is dedicated to investigating the maximum risk of the MLE  $R_{\text{maximum}}(\mathcal{P}; P_n)$  and the minimax risk  $R_{\text{minimax}}(\mathcal{P})$  for various  $\mathcal{P}$ . There are good reasons for focusing on the  $\ell_1$  loss, as we do in this paper. Other loss functions in distribution estimation, such as the squared error loss, have been extensively studied in a series of papers [4]–[7], while less is known for the  $\ell_1$  loss. For the squared error loss, the minimax estimator is unique and depends on the support size  $S$  [8, Pg. 349]. Since the support size  $S$  is unknown in our setting, this estimator is highly impractical. This fact partially motivates our focus on the  $\ell_1$  loss, which turns out to bridge our understanding of both parametric and nonparametric models. The  $\ell_1$  loss, being proportional to the variational distance, is often a more natural measure of discrepancy between distributions than the  $\ell_2$  loss. Moreover, the  $\ell_1$  loss in discrete distribution estimation is compatible with and is a degenerate case of the  $L_1$  loss in density estimation, which is the only loss that satisfies certain natural properties [9].

All logarithms in this paper are assumed to be in the natural base.

#### A. Main results

We investigate the maximum risk of the MLE  $R_{\text{maximum}}(\mathcal{P}; P_n)$  and the minimax risk  $R_{\text{minimax}}(\mathcal{P})$  in the aforementioned three different regimes separately.

Understanding  $R_{\text{maximum}}(\mathcal{P}; P_n) = \sup_{P \in \mathcal{P}} R(P; P_n)$  follows from an understanding of  $R(P; P_n) = \mathbb{E}_P \|P_n - P\|_1$ . This problem can be decomposed into analyzing the Binomial mean absolute deviation defined as

$$\mathbb{E} \left| \frac{X}{n} - p \right|, \quad (6)$$

where  $X \sim \text{B}(n, p)$  follows a Binomial distribution. Complicated as it may seem, De Moivre obtained an explicit expression for this quantity. Diaconis and Zabell provided a nice historical account for De Moivre's discovery in [10].

Berend and Kontorovich [11] provided tight upper and lower bounds on the Binomial mean absolute deviation, and we summarize some key results in Lemma 4 of the Appendix. A well-known result to recall first is the following.

**Theorem 1.** *The maximum  $\ell_1$  risk of the empirical distribution  $P_n$  satisfies*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{\frac{S-1}{n}}, \quad (7)$$

where  $\mathcal{M}_S$  denotes the set of distributions with support size  $S$ .

In fact, a tighter upper bound on the worst-case  $\ell_1$  risk of

the empirical distribution  $P_n$  has been obtained in [12]:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{\frac{2(S-1)}{\pi n}} + \frac{2S^{\frac{1}{2}}(S-1)^{\frac{1}{4}}}{n^{\frac{3}{4}}}. \quad (8)$$

In the present paper we show that MLE is minimax rate-optimal in all the three regimes we consider, but possibly suboptimal in terms of constants in high dimensional and infinite dimensional settings.

*1) Classical Asymptotics:* The next corollary is an immediate result from Theorem 1.

**Corollary 1.** *The empirical distribution  $P_n$  achieves the worst-case convergence rate  $O(n^{-\frac{1}{2}})$ . Specifically,*

$$\limsup_{n \rightarrow \infty} \sqrt{n} \cdot \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{S-1} < \infty. \quad (9)$$

Regarding the lower bound, the well-known Hájek-Le Cam local asymptotic minimax theorem [13] (Theorem 6 in Appendix A) and corresponding achievability theorems [2, Lemma 8.14] show that the MLE is optimal (even in constants) in classical asymptotics. Concretely, one corollary of Theorem 6 in the Appendix shows the following.

**Corollary 2.** *Fixing  $S \geq 2$ , we have*

$$\liminf_{n \rightarrow \infty} \sqrt{n} \cdot \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \sqrt{\frac{2(S-1)}{\pi}} > 0, \quad (10)$$

where the infimum is taken over all estimators.

Note that the combination of (8) and Corollary 2 actually yields that the MLE is asymptotically minimax, and

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sqrt{n} \cdot \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \\ &= \lim_{n \rightarrow \infty} \sqrt{n} \cdot \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 = \sqrt{\frac{2(S-1)}{\pi}}, \end{aligned} \quad (11)$$

where the infimum is taken over all estimators. This result was also proved in [12] via a different approach to obtain the exact constant for the lower bound in the classical asymptotics setting.

*2) High-dimensional Asymptotics:* Theorem 1 also implies the following:

**Corollary 3.** *For  $S = n/c$ , the empirical distribution  $P_n$  achieves the worst-case convergence rate  $O(c^{-\frac{1}{2}})$ , i.e.,*

$$\limsup_{c \rightarrow \infty} \sqrt{c} \cdot \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq 1 < \infty. \quad (12)$$

Now we show that  $S = n/c$  is the *critical scaling* in high dimensional asymptotics. In other words, if  $n = o(S)$ , then no estimator for the distribution  $P$  is consistent under  $\ell_1$  loss. This phenomenon has been observed in several papers, such as [11] and [14], to name a few.

The following theorem presents a non-asymptotic minimax lower bound.

**Theorem 2.** For any  $\zeta \in (0, 1]$ , we have

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{1}{8} \sqrt{\frac{eS}{(1+\zeta)n}} \mathbb{1} \left( \frac{(1+\zeta)n}{S} > \frac{e}{16} \right) + \exp \left( -\frac{2(1+\zeta)n}{S} \right) \mathbb{1} \left( \frac{(1+\zeta)n}{S} \leq \frac{e}{16} \right) - \exp \left( -\frac{\zeta^2 n}{24} \right) - 12 \exp \left( -\frac{\zeta^2 S}{32(\ln S)^2} \right), \quad (13)$$

where the infimum is taken over all estimators.

Theorem 2 implies the following minimax lower bound in high dimensional asymptotics, if we take  $\zeta \rightarrow 0^+$ .

**Corollary 4.** For any constant  $c > 0$ , if  $S = n/c$ , the convergence rate of the maximum  $\ell_1$  risk is  $\Omega(c^{-\frac{1}{2}})$ . Specifically,

$$\liminf_{c \rightarrow \infty} \sqrt{c} \cdot \liminf_{n \rightarrow \infty} \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{\sqrt{e}}{8} > 0, \quad (14)$$

where the infimum is taken over all estimators.

Corollaries 3 and 4 imply that MLE achieves the optimal convergence rate  $\Theta(c^{-\frac{1}{2}})$  in high dimensional linear scaling.

3) *Infinite-dimensional Asymptotics:* The performance of MLE in the regime of bounded entropy is characterized in the following theorem.

**Theorem 3.** The empirical distribution  $P_n$  satisfies that, for any  $H > 0$  and  $\eta > 1$ ,

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 \leq \frac{2H}{\ln n - 2\eta \ln \ln n} + \frac{1}{(\ln n)^\eta}. \quad (15)$$

Further, for any  $c \in (0, 1)$  and  $n > \max\{(1-c)^{-\frac{1}{1-c}}, e^H\}$ ,

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 \geq \frac{2cH}{\ln n} \left( 1 - ((1-c)n)^{-\frac{1}{c}} \right)^n. \quad (16)$$

The next corollary follows from Theorem 3 after taking  $c \rightarrow 1^-$ .

**Corollary 5.** For any  $H > 0$ , the MLE  $P_n$  satisfies

$$\lim_{n \rightarrow \infty} \frac{\ln n}{H} \cdot \sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 = 2. \quad (17)$$

It implies that we not only have obtained the  $\Theta((\ln n)^{-1})$  convergence rate of the asymptotic  $\ell_1$  risk of MLE, but also shown that the multiplicative constant is exactly  $2H$ . We note that this logarithmic convergence rate is really slow, implying that the sample size needs to be squared to reduce the maximum  $\ell_1$  risk by a half. Also note that the maximum  $\ell_1$  risk is proportional to the entropy  $H$ , thus the smaller the entropy of a distribution is known to be, the easier it is to estimate.

However, given this slow rate  $\Theta((\ln n)^{-1})$ , it is of utmost importance to obtain estimators such that the corresponding multiplicative constant is as small as possible. We show that MLE does not achieve the optimal constant. In the following theorem, an essentially minimax estimator is explicitly con-

structed.

**Theorem 4.** For any  $\eta > 1$ , denote

$$\Delta_n \triangleq \frac{(\ln n)^{2\eta}}{n}, \quad (18)$$

then for the hard-thresholding estimator defined as  $\hat{P}^*(\mathbf{X}) = (g_n(X_1), g_n(X_2), \dots, g_n(X_S))$  with

$$g_n(X_i) = \frac{X_i}{n} \mathbb{1} \left( \frac{X_i}{n} > e^2 \Delta_n \right), \quad (19)$$

we have

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P}^* - P\|_1 \leq \frac{H}{\ln n - \ln(2e^2) - 2\eta \ln \ln n} + (\ln n)^{-\eta} + n^{1-\frac{\eta}{4}}. \quad (20)$$

Moreover, for any  $c \in (0, 1)$  and  $n \geq e^H$ , we have

$$\inf_{\hat{P}} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{cH}{\ln n} \cdot \left( 1 - n^{1-\frac{1}{c}} (1-c)^{-\frac{1}{c}} \right), \quad (21)$$

where the infimum is taken over all estimators.

Theorem 4 presents both a non-asymptotic achievable maximum  $\ell_1$  risk and a non-asymptotic lower bound of the minimax  $\ell_1$  risk, and it is straightforward to verify that the upper bound and lower bound coincide asymptotically by choosing  $c \rightarrow 1^-$ . As a result, the asymptotic minimax  $\ell_1$  risk is characterized in the following corollary.

**Corollary 6.** For any  $H > 0$ , the asymptotic minimax risk is  $\frac{H}{\ln n}$ , i.e.,

$$\lim_{n \rightarrow \infty} \frac{\ln n}{H} \cdot \inf_{\hat{P}} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 = 1, \quad (22)$$

and the estimator  $\hat{P}$  in Theorem 4 is asymptotically minimax.

In light of Corollaries 5 and 6, the asymptotic minimax  $\ell_1$  risk for the distribution estimation with bounded entropy is exactly  $\frac{H}{\ln n}$ , half of that obtained by MLE. Since the convergence rate is  $\Theta((\ln n)^{-1})$ , the performance of the asymptotically minimax estimator with  $n$  samples in this problem is nearly that of MLE with  $n^2$  samples, which is a significant improvement.

Note that the asymptotically minimax estimator given in Theorem 4 is a hard-thresholding estimator which neglects all symbols with frequency less than  $e^2 \Delta_n = e^2 (\ln n)^{2\eta}/n$ , and its risk depends on  $\eta$  through two factors. On one hand, there is a loss due to ignoring low frequency symbols, which increases with  $\eta$  and corresponds to the first term in the right-hand side of (20). On the other hand, the risk incurred by the involvement of high frequency symbols decreases with  $\eta$  and corresponds to the  $(\ln n)^{-\eta}$  term in (20). Taking derivative with respect to  $\eta$  yields that the optimal parameter  $\eta^*$  to handle this trade-off takes the form  $\eta^* = c_n \ln n / \ln \ln n$  with coefficients  $c_n \rightarrow 0$  (but larger than  $\Theta(\ln \ln n / \ln n)$ ), and then the upper bound

of the  $\ell_1$  risk in (20) becomes

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P}^* - P\|_1 \leq \frac{H}{(1 - 2c_n) \ln n - \ln(2e^2)} + n^{-c_n} + n^{1 - \frac{c^2}{4}}. \quad (23)$$

However, one may notice that the asymptotically minimax estimator in Theorem 4 outputs estimates that are not necessarily probability distributions. What if we constrain the estimator  $\hat{P}$  to output estimates confined to  $\mathcal{M}_\infty \triangleq \cup_{S=1}^\infty \mathcal{M}_S$ , i.e., are bona fide probabilities? In this case, the next theorem shows that the MLE is asymptotically minimax again.

**Theorem 5.** *For any  $c \in (0, 1)$  and  $n \geq e^H$ , we have*

$$\inf_{\hat{P} \in \mathcal{M}_\infty} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{2cH}{\ln n} \cdot \left(1 - n^{1 - \frac{1}{c}} (1 - c)^{-\frac{1}{c}}\right), \quad (24)$$

where the infimum is taken over all estimators with outputs confined to the probability simplex.

The next corollary follows immediately from Corollary 5 upon taking  $c \rightarrow 1^-$ :

**Corollary 7.** *For any  $H > 0$ , the asymptotic minimax risk is  $\frac{2H}{\ln n}$  when the estimates are restricted to the probability simplex:*

$$\lim_{n \rightarrow \infty} \frac{\ln n}{H} \cdot \inf_{\hat{P} \in \mathcal{M}_\infty} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 = 2, \quad (25)$$

and the MLE is asymptotically minimax.

Hence, the MLE still enjoys the asymptotically minimax property in the bounded entropy case if the estimates have to be probability mass functions. We can explain this peculiar phenomenon from the Bayes viewpoint as follows. By the minimax theorem [15], any minimax estimator can be always approached by a sequence of Bayes estimators under some priors. However, for general loss functions including the  $\ell_1$  loss, the corresponding Bayes estimator (without any restrictions) may not belong to the probability simplex  $\mathcal{M}_\infty$  even if the prior is supported on  $\mathcal{M}_\infty$ . Hence, the space of all Bayes estimators which form probability masses is strictly smaller than the space of all free Bayes estimators without any constraints. For example, it is well-known that under the  $\ell_1$  loss, the Bayes response is actually the median vector of the posterior distribution, and thus does not necessarily form a probability distribution. We also remark that for some special loss functions such as the squared error loss, the Bayes estimator under any prior supported on  $\mathcal{M}_\infty$  will always belong to  $\mathcal{M}_\infty$ . Broadly speaking, if the loss function is a Bregman divergence [16], the Bayes estimator under any prior will always be the conditional expectation, which is supported on the convex hull of the parameter space.

## B. Discussion

Now we draw various connections between our results and the literature.

1) *Density estimation under  $L_1$  loss:* There is an extensive literature on density estimation under  $L_1$  loss, and we refer to the book by Devroye and Györfi [9] for an excellent overview. This problem is also very popular in theoretical computer science, e.g. [17].

The problem of discrete distribution estimation under  $\ell_1$  loss has been the subject of recent interest in the theoretical computer science community. For example, Daskalakis, Diakonikolas and Servedio considered the problem of  $k$ -modal distributions [18], and a very recent talk by Diakonikolas [14] provided a literature survey. The conclusion that it is necessary and sufficient to use  $n = \Theta(S)$  samples to consistently estimate an arbitrary discrete distribution with support size  $S$  has essentially appeared in the literature [14], but we did not find an explicit reference giving non-asymptotic results, and for completeness we have included proofs corresponding to the high dimensional asymptotics in this paper. We remark that, a very detailed analysis of the discrete distribution estimation problem under  $\ell_1$  loss may be insightful and instrumental in future breakthroughs in density estimation under  $L_1$  loss.

2) *Entropy estimation:* It was shown that the entropy is nearly Lipschitz continuous under the  $\ell_1$  norm [19, Thm. 17.3.3] [20, Lemma 2.7], i.e., if  $\|P - Q\|_1 \leq 1/2$ , then

$$|H(P) - H(Q)| \leq -\|P - Q\|_1 \ln \frac{\|P - Q\|_1}{S}, \quad (26)$$

where  $S$  is the support size. At first glance, it seems to suggest that the estimation of entropy can be reduced to estimation of discrete distributions under  $\ell_1$  loss. However, this question is far more complicated than it appears.

First, people have already noticed that this near-Lipschitz continuity result is valid only for finite alphabets [21]. The following result by Antos and Kontoyiannis [22] addresses entropy estimation over countably infinite support sizes.

**Remark 1.** *Among all discrete sources with finite entropy and  $\text{Var}(-\ln p(X)) < \infty$ , for any sequence  $\{H_n\}$  of estimators for the entropy, and for any sequence  $\{a_n\}$  of positive numbers converging to zero, there is a distribution  $P$  (supported on at most countably infinite symbols) with  $H = H(P) < \infty$  such that*

$$\limsup_{n \rightarrow \infty} \frac{\mathbb{E}_P |H_n - H|}{a_n} = \infty. \quad (27)$$

Remark 1 shows that, among all sources with finite entropy and finite varentropy, no rate-of-convergence results can be obtained for any sequence of estimators. Indeed, if the entropy is still nearly-Lipschitz continuous with respect to  $\ell_1$  distance in the infinite alphabet setting, then Corollary 5 immediately implies that the MLE plug-in estimator for entropy attains a universal convergence rate. That there is no universal convergence rate of entropy estimators for sources with bounded entropy is particularly interesting in light of the fact that the minimax rates of convergence of distribution estimation with bounded entropy is  $\Theta((\ln n)^{-1})$ .

Second, it is very interesting and deserves pondering that along high dimensional asymptotics, the minimax sample complexity for estimating entropy is  $n = \Theta(\frac{S}{\ln S})$  samples, a result first discovered by Valiant and Valiant [23], then

recovered by Jiao *et al.* using a different approach in [24], and Wu and Yang in [25]. Since it is shown in Corollary 4 that we need  $n = \Theta(S)$  samples to consistently estimate the distribution, this result shows that we can consistently estimate the entropy without being able to consistently estimate the underlying distribution under  $L_1$  loss. Note that if the plug-in approach is used for entropy estimation, i.e., if we use  $H(P_n)$  to estimate  $H(P)$ , it has been shown in [26], [27] that this estimator again requires  $n = \Theta(S)$  samples. In fact, for a wide class of functionals of discrete distributions, it is shown in [24] that the MLE is strictly suboptimal, and the performance of the optimal estimators with  $n$  samples is essentially that of the MLE with  $n \ln n$  samples, a phenomenon termed “effective sample size enlargement” in [24]. Jiao *et al.* [28] showed that the improved estimators introduced in [24] can lead to consistent and substantial performance boosts in various machine learning algorithms.

### 3) $\ell_1$ divergence estimation between two distributions:

Now we turn to the estimation problem for the  $\ell_1$  divergence  $\|P - Q\|_1$  between two discrete distributions  $P, Q$  with support size at most  $S$ . At first glance, by setting one of the distributions to be deterministic, the problem of  $\ell_1$  divergence estimation seems a perfect dual to the distribution estimation problem under  $\ell_1$  loss. However, compared to the required sample complexity  $n = \Theta(S)$  in the distribution estimation problem under  $\ell_1$  loss, the minimax sample complexity for estimating the  $\ell_1$  divergence between two arbitrary distributions is  $n = \Theta(\frac{S}{\ln S})$  samples, a result of Valiant and Valiant [29]. Hence, it is easier to estimate the  $\ell_1$  divergence than to estimate the distribution with a vanishing  $\ell_1$  risk. Note that for distribution estimation, for each symbol we need to obtain a good estimate for  $p_i$  in terms of the  $\ell_1$  risk, while for  $\ell_1$  divergence estimation we do not need to estimate each  $p_i$  and  $q_i$  separately.

### 4) Joint $d$ -block distribution estimation in stochastic processes:

Insights can be gleaned from the comparison of Corollary 5 and the result obtained by Marton and Shields [3]. Marton and Shields showed that, in a stationary ergodic stochastic process with sample size  $n$  and entropy rate  $H$ , the joint  $d$ -tuple distribution of the process can be consistently estimated using the empirical distribution if  $d \leq \frac{(1-\epsilon)\ln n}{H}$ , where  $\epsilon > 0$  is an arbitrary constant. Moreover, the empirical distribution is not consistent if  $d \geq \frac{(1+\epsilon)\ln n}{H}$ . Now we treat the joint  $d$ -tuple distribution as a single distribution with a large support size, and consider the corresponding estimation problem. To be precise, we assume without loss of generality that the original process consists of  $n$  observations and merge all disjoint blocks containing  $d = \frac{(1-\epsilon)\ln n}{H}$  symbols into supersymbols. Consequently, we obtain a sample size of  $n/d$  from which we would like to learn a new distribution over an alphabet of size  $S^d$  and entropy nearly  $dH$ . Considering a special case where the stationary ergodic process is nearly i.i.d. and applying our result on the minimum sample complexity of the distribution estimation problem with bounded support size, we need  $\Theta(S^d) = \Theta(n^{\frac{(1-\epsilon)\ln S}{H}})$  samples to estimate the new distribution, which cannot be achieved by  $n$  samples unless  $\ln S = H$ , i.e., the distribution is uniform. Furthermore, under

the regime of distribution estimation with bounded entropy, the asymptotic minimax risk of MLE will be  $\frac{2dH}{\ln n} = 2(1-\epsilon)$ , which does not vanish as  $n \rightarrow \infty$ . Hence, we conclude that the minimax distribution estimation under i.i.d. samples is indeed harder than the joint  $d$ -tuple distribution estimation in a stationary ergodic process. To clarify the distinction, we remark that the ergodicity plays a crucial role in the latter case: for  $d \rightarrow \infty$ , the Shannon-McMillan-Breiman theorem [30] guarantees that there are approximately  $e^{dH}$  typical sequences of length  $d$ , each of which occurs with probability about  $e^{-dH}$ . Then it turns out that we need only to estimate a uniform distribution with support size  $e^{dH}$ , and applying the previous conclusion  $n/d = \Theta(e^{dH})$  yields the desired result  $d \approx \frac{\ln n}{H}$ . On the other hand, for the distribution estimation problem with a large support size, the equipartition property does not necessarily hold, and our scheme focuses on the worst-case risk over all possible distributions. Hence, it is indeed harder to handle the distribution estimation problem under i.i.d. samples in the large alphabet regime, and it indicates that directly applying results from a large alphabet regime to stochastic processes with a large memory may not be a fruitful route. Before closing the discussion, we mention that Marton and Shields did not show that the  $d \sim \frac{\ln n}{H}$  scaling is minimax optimal. It remains an interesting question to investigate whether there is a better estimator to estimate the  $d$ -tuple joint distribution in stochastic processes.

### 5) Hard-thresholding estimation is asymptotically minimax:

Corollary 6 shows that in the infinite dimensional asymptotics, MLE is far from asymptotically minimax, and a hard-thresholding estimator achieves the asymptotic minimax risk. The phenomenon that thresholding methods are needed in order to obtain minimax estimators for high dimensional parameters in a  $\ell_p$  ball under  $\ell_q$  error,  $p > 0, p < q, q \in [1, \infty)$ , was first noticed by Donoho and Johnstone [31]. Following the rationale of the James-Stein shrinkage estimator [32], Donoho and Johnstone proposed the soft- and hard-thresholding estimators for the normal mean given that we know *a priori* that the mean  $\theta$  lies in a  $\ell_p$  ball,  $p \in (0, \infty)$ . Later, Donoho and Johnstone applied this idea to nonparametric estimation in Besov spaces, and obtained the famous *wavelet shrinkage* estimator for denoising [33]. Note that the set  $\{P : H(P) \leq H\}$  forms a ball similar to the  $\ell_p$  ball, and the loss function is  $\ell_1$ , so it is not surprising that hard-thresholding leads to an asymptotically minimax estimator. The asymptotic minimax estimators under other constraints on the distribution  $P$  remain to be explored.

### 6) Adaptive estimation:

Note that in the infinite dimensional asymptotics, for a sequence of problems  $\{H(P) \leq H\}$  with different upper bounds  $H$ , the asymptotically minimax estimator in Theorem 4 achieves the minimax risk over all “entropy balls” without knowing its “radius”  $H$ . It is very important in practice, since we do not know *a priori* an upper bound on the entropy of the distribution. This estimator belongs to a general collections of estimators called adaptive estimators. For details we refer to a survey paper by Cai [34].

The rest of this paper is organized as follows. Section II provides outlines of the proofs of the main theorems, and some useful auxiliary lemmas are listed in Appendix A.

Complete proofs of some lemmas and corollaries are provided in Appendix B.

## II. OUTLINES OF PROOFS OF MAIN THEOREMS

### A. Analysis of MLE

For the analysis of the performance of the MLE  $P_n$ , the key is to obtain a good approximation of  $\mathbb{E}|X/n - p|$  with  $X \sim \text{B}(n, p)$ , i.e., the Binomial mean absolute deviation. Lemma 4 in the Appendix lists some sharp approximations, which together with the concavity of  $\sqrt{x(1-x)}$  yield

$$\mathbb{E}_P \|P - P_n\|_1 = \sum_{i=1}^S \mathbb{E}_P \left| \frac{X_i}{n} - p_i \right| \quad (28)$$

$$\leq \sum_{i=1}^S \sqrt{\frac{p_i(1-p_i)}{n}} \quad (29)$$

$$\leq \sqrt{\frac{S-1}{n}}, \quad (30)$$

which completes the proof of Theorem 1. For the upper bound in Theorem 3, we use Lemma 4 again and obtain

$$\mathbb{E}_P \|P - P_n\|_1 \leq \sum_{i=1}^S \min \left\{ \sqrt{\frac{p_i}{n}}, 2p_i \right\} \quad (31)$$

$$\leq 2 \sum_{p_i \leq \frac{(\ln n)^{2\eta}}{n}} p_i + \frac{1}{\sqrt{n}} \sum_{p_i > \frac{(\ln n)^{2\eta}}{n}} \sqrt{p_i} \quad (32)$$

$$\leq \frac{2}{\ln n - 2\eta \ln \ln n} \sum_{p_i \leq \frac{(\ln n)^{2\eta}}{n}} (-p_i \ln p_i) + \frac{1}{\sqrt{n}} \cdot \sqrt{\left| \left\{ i : p_i > \frac{(\ln n)^{2\eta}}{n} \right\} \right|} \quad (33)$$

$$\leq \frac{2H}{\ln n - 2\eta \ln \ln n} + \frac{1}{(\ln n)^\eta}. \quad (34)$$

For the lower bound, we consider the distribution  $P = (\delta/S', \dots, \delta/S', 1 - \delta)$  with entropy  $H$ , then for  $c \in (0, 1)$ ,  $\delta \leq c$ , due to the monotone decreasing property of  $(1-x)^{\frac{1}{x}}$  with respect to  $x \in (0, 1)$ ,

$$S' = \delta \exp \left( \frac{H}{\delta} \right) \cdot \left( (1 - \delta)^{\frac{1}{\delta}} \right)^{1-\delta} \quad (35)$$

$$\geq \delta \exp \left( \frac{H}{\delta} \right) \cdot (1 - c)^{\frac{1}{c}}. \quad (36)$$

Note that  $S' + 1$  is the support size, and we assume without loss of generality that  $S'$  is an integer. For  $c \in (0, 1)$ , since  $n \geq e^H$ , we choose  $\delta = cH / \ln n \leq c$ , then since  $n > (1 - c)^{-\frac{1}{1-c}}$ ,

$$\frac{\delta}{S'} \leq (1 - c)^{-\frac{1}{c}} \exp \left( -\frac{H}{\delta} \right) = ((1 - c)n)^{-\frac{1}{c}} < \frac{1}{n}, \quad (37)$$

and the identity in Lemma 4 can be applied to obtain

$$\mathbb{E}_P \|P - P_n\|_1 \geq \sum_{i=1}^{S'} \mathbb{E}_P \left| p_i - \frac{X_i}{n} \right| \quad (38)$$

$$= 2\delta \left( 1 - \frac{\delta}{S'} \right)^n \quad (39)$$

$$\geq \frac{2cH}{\ln n} \left( 1 - ((1 - c)n)^{-\frac{1}{c}} \right)^n. \quad (40)$$

### B. Analysis of the Estimator in Theorem 4

For the achievability result, we first establish some lemmas on the properties of  $g_n(X)$ .

**Lemma 1.** *If  $X \sim \text{B}(n, p)$ ,  $p \leq \Delta_n$ , we have*

$$\mathbb{E} |g_n(X) - p| \leq p + \left( \frac{p}{e\Delta_n} \right)^{e^2 \ln n}. \quad (41)$$

*Proof:* It is clear from the triangle inequality that

$$\mathbb{E} |g_n(X) - p| \leq p + \mathbb{E} \left[ \frac{X}{n} \mathbb{1} \left( \frac{X}{n} > e^2 \Delta_n \right) \right] \quad (42)$$

$$\leq p + \mathbb{P}(X > e^2 n \Delta_n) \quad (43)$$

$$\leq p + \left( \frac{np}{en\Delta_n} \right)^{e^2 (\ln n)^{2\eta}} \quad (44)$$

where Lemma 5 in Appendix A is used in the last step. The proof is completed by noticing that  $\frac{p}{e\Delta_n} \leq \frac{1}{e} < 1$ . ■

**Lemma 2.** *If  $X \sim \text{B}(n, p)$ ,  $p \geq 2e^2 \Delta_n$ , we have*

$$\mathbb{E} |g_n(X) - p| \leq \sqrt{\frac{p}{n}} + n^{-\frac{e^2}{4}}. \quad (45)$$

*Proof:* It follows from the identity

$$g_n(X) = \frac{X}{n} - \frac{X}{n} \mathbb{1} \left( \frac{X}{n} \leq e^2 \Delta_n \right) \quad (46)$$

and the triangle inequality that

$$\mathbb{E} |g_n(X) - p| \leq \mathbb{E} \left| \frac{X}{n} - p \right| + \mathbb{E} \left[ \frac{X}{n} \mathbb{1} \left( \frac{X}{n} \leq e^2 \Delta_n \right) \right] \quad (47)$$

$$\leq \mathbb{E} \left| \frac{X}{n} - p \right| + \mathbb{P}(X \leq e^2 n \Delta_n) \quad (48)$$

$$\leq \sqrt{\frac{p}{n}} + e^{-\frac{e^2 (\ln n)^{2\eta}}{4}}. \quad (49)$$

Then the proof is completed by noticing that  $e^{-\frac{e^2 (\ln n)^{2\eta}}{4}} \leq e^{-\frac{e^2 \ln n}{4}} = n^{-\frac{e^2}{4}}$ . ■

**Lemma 3.** *If  $X \sim \text{B}(n, p)$ ,  $\Delta_n < p < 2e^2 \Delta_n$ , we have*

$$\mathbb{E} |g_n(X) - p| \leq \sqrt{\frac{p}{n}} + p. \quad (50)$$

*Proof:* It is clear that

$$\mathbb{E} |g_n(X) - p| \leq \mathbb{E} \left[ \left| \frac{X}{n} - p \right| \mathbb{1} \left( \frac{X}{n} > e^2 \Delta_n \right) \right] + \mathbb{E} \left[ |0 - p| \mathbb{1} \left( \frac{X}{n} \leq e^2 \Delta_n \right) \right] \quad (51)$$

$$\leq \mathbb{E} \left| \frac{X}{n} - p \right| + p \quad (52)$$

$$\leq \sqrt{\frac{p}{n}} + p. \quad (53)$$

Combining these lemmas, the upper bound of Theorem 4 is given on the bottom of this page.

### C. Proof of the Lower Bounds in Theorem 2, 4 and 5

To obtain a lower bound for the minimax risk, an effective way is to use the Bayes risk to serve as the lower bound [8], where the prior can be arbitrarily chosen. Hence, our target is to find an unfavorable prior and compute the corresponding Bayes risk. For computational simplicity, in the proof we will assign the product of independent priors to the whole probability vector  $P$  based on the Poissonized model  $X_i \sim \text{Poi}(np_i)$ , and then use some concentration inequalities

such as the Hoeffding bound to ensure that the vector  $P$  is close to a probability distribution with overwhelming probability. Then the relationship between the minimax risk of the Poissonized model and that of the Multinomial model needs to be established. The rigorous proofs are detailed as follows.

1) *Lower Bound in Theorem 2:* We denote the uniform distribution on two points  $\{\frac{1-\eta}{S}, \frac{1+\eta}{S}\}$  by  $\mu_0$ , with  $\eta \in (0, 1)$  to be specified later, and assign the product measure  $\mu_0^S$  to the probability vector  $P$ . Under the Poissonized model  $X_i \sim \text{Poi}(np_i)$ , it is straightforward to see that all  $p_i (1 \leq i \leq S)$  are conditionally independent given  $\mathbf{X}$ . Hence, the Bayes estimator  $\hat{P}^B(\mathbf{X})$  under prior  $\mu_0^S$  can be decomposed into  $\hat{P}^B(\mathbf{X}) = (f(X_1), f(X_2), \dots, f(X_S))$ , for some function  $f(\cdot)$ . Then the Bayes risk is shown on the bottom of this page, where  $d_{\text{TV}}(\text{Poi}(u), \text{Poi}(v))$  is the variational distance between two Poisson distributions:

$$d_{\text{TV}}(\text{Poi}(u), \text{Poi}(v)) \triangleq \sup_{A \subset \mathbb{N}} |\mathbb{P}\{\text{Poi}(u) \in A\} - \mathbb{P}\{\text{Poi}(v) \in A\}|. \quad (65)$$

$$\mathbb{E}_P \|P - \hat{P}\|_1 \leq \sum_{p_i \leq \Delta_n} \left( p_i + \left( \frac{p_i}{e\Delta_n} \right)^{e^2 \ln n} \right) + \sum_{\Delta_n < p_i < 2e^2 \Delta_n} \left( \sqrt{\frac{p_i}{n}} + p_i \right) + \sum_{p_i \geq 2e^2 \Delta_n} \left( \sqrt{\frac{p_i}{n}} + n^{-\frac{\epsilon^2}{4}} \right) \quad (54)$$

$$= \sum_{p_i < 2e^2 \Delta_n} p_i + \sum_{p_i > \Delta_n} \sqrt{\frac{p_i}{n}} + \sum_{p_i \leq \Delta_n} \left( \frac{p_i}{e\Delta_n} \right)^{e^2 \ln n} + \sum_{p_i \geq 2e^2 \Delta_n} n^{-\frac{\epsilon^2}{4}} \quad (55)$$

$$\leq \left( \ln \frac{1}{2e^2 \Delta_n} \right)^{-1} \cdot H + \frac{1}{\sqrt{n\Delta_n}} + e^{-e^2 \ln n} \cdot \frac{n}{(\ln n)^{2\eta}} + n^{-\frac{\epsilon^2}{4}} \cdot \frac{1}{2e^2 \Delta_n} \quad (56)$$

$$= \frac{H}{\ln n - \ln(2e^2) - 2\eta \ln \ln n} + \frac{1}{(\ln n)^\eta} + \frac{1}{n^{e^2-1} (\ln n)^{2\eta}} + \frac{1}{2e^2 n^{\frac{\epsilon^2}{4}-1} (\ln n)^{2\eta}} \quad (57)$$

$$\leq \frac{H}{\ln n - \ln(2e^2) - 2\eta \ln \ln n} + (\ln n)^{-\eta} + n^{1-\frac{\epsilon^2}{4}} \quad (58)$$

$$R_B(S, n, \mu_0^S) = \int \mathbb{E}_P \|P - \hat{P}^B\|_1 \mu(dP) \quad (59)$$

$$= \sum_{i=1}^S \int \sum_{k=0}^{\infty} |p_i - f(k)| e^{-np_i} \frac{(np_i)^k}{k!} \mu_0(dp_i) \quad (60)$$

$$\geq \sum_{i=1}^S \int \sum_{k=0}^{\infty} |p_i - f(k)| \min \left\{ \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1-\eta)}{S} \right) = k \right\}, \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1+\eta)}{S} \right) = k \right\} \right\} \mu_0(dp_i) \quad (61)$$

$$\geq S \cdot \frac{\eta}{S} \sum_{k=0}^{\infty} \min \left\{ \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1-\eta)}{S} \right) = k \right\}, \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1+\eta)}{S} \right) = k \right\} \right\} \quad (62)$$

$$= \eta \cdot \sum_{k=0}^{\infty} \min \left\{ \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1-\eta)}{S} \right) = k \right\}, \mathbb{P}\left\{ \text{Poi}\left( \frac{n(1+\eta)}{S} \right) = k \right\} \right\} \quad (63)$$

$$= \eta \cdot \left( 1 - d_{\text{TV}} \left( \text{Poi}\left( \frac{n(1-\eta)}{S} \right), \text{Poi}\left( \frac{n(1+\eta)}{S} \right) \right) \right), \quad (64)$$

Adell and Jodra [35] gives an upper bound for this distance:

$$d_{TV}(\text{Poi}(t), \text{Poi}(t+x)) \leq \min \left\{ 1 - e^{-x}, \sqrt{\frac{2}{e}}(\sqrt{t+x} - \sqrt{t}) \right\}, \quad t, x \geq 0 \quad (66)$$

then

$$d_{TV} \left( \text{Poi} \left( \frac{n(1-\eta)}{S} \right), \text{Poi} \left( \frac{n(1+\eta)}{S} \right) \right) \quad (67)$$

$$\leq \min \left\{ 1 - \exp \left( -2\eta \cdot \frac{n}{S} \right), \sqrt{\frac{2n}{eS}}(\sqrt{1+\eta} - \sqrt{1-\eta}) \right\} \quad (68)$$

$$\leq \min \left\{ 1 - \exp \left( -2\eta \cdot \frac{n}{S} \right), 2\eta \cdot \sqrt{\frac{n}{eS}} \right\}. \quad (69)$$

Hence, by setting

$$\eta = \min \left\{ 1, \frac{1}{4} \sqrt{\frac{eS}{n}} \right\} \quad (70)$$

we can obtain

$$R_B(S, n, \mu_0^S) \geq \begin{cases} \exp \left( -\frac{2n}{S} \right), & \frac{n}{S} \leq \frac{e}{16} \\ \frac{1}{8} \sqrt{\frac{eS}{n}}, & \frac{n}{S} > \frac{e}{16}. \end{cases} \quad (71)$$

The combination of Lemma 7 and 8 in Appendix A yields, for any  $\zeta \in (0, 1]$  and  $\epsilon \in (0, \frac{\zeta}{2(1+\zeta)}]$ ,

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq R_B(S, (1+\zeta)n, \mu_0^S) - \exp \left( -\frac{\zeta^2 n}{24} \right) - 6\mu_0^S(\mathcal{M}_S(\epsilon)^c). \quad (72)$$

Setting  $\epsilon = \frac{\zeta}{4 \ln S}$ , Lemma 6 in Appendix A yields

$$\mu_0^S(\mathcal{M}_S(\epsilon)^c) = \mu_0^S \left\{ \left| \sum_{i=1}^S p_i - \mathbb{E} \left[ \sum_{i=1}^S p_i \right] \right| \geq \epsilon \right\} \quad (73)$$

$$\leq 2 \exp \left( -\frac{2\epsilon^2}{S \cdot (2/S)^2} \right) \quad (74)$$

$$= 2 \exp \left( -\frac{\zeta^2 S}{32(\ln S)^2} \right). \quad (75)$$

The proof of Theorem 2 is completed by the combination of (71), (72) and (73).

2) *Lower Bound in Theorem 4 and 5:* For the proof of the lower bound in Theorem 4, we need a different prior. Specifically, we fix some  $\delta \in (0, 1)$  with its value to be specified later, and consider  $S'$  with

$$\delta \ln S' - \delta \ln \delta - (1-\delta) \ln(1-\delta) = H. \quad (76)$$

Now define  $S = kS' + 1$  with parameter  $k \geq 2$  to be specified later, and consider the following collection  $\mathcal{N}_{S,H}$  of  $S$ -dimensional non-negative vectors:  $P \in \mathcal{N}_{S,H}$  if and only if  $p_S = 1 - \delta, p_i \in \{0, \frac{\delta}{S'}\}, 1 \leq i \leq S-1 = kS'$ , and

$$\left| \left\{ i : 1 \leq i \leq kS', p_i = \frac{\delta}{S'} \right\} \right| = S'. \quad (77)$$

By the preceding two equalities, we can easily verify that  $P \in \mathcal{N}_{S,H}$  implies  $H(P) = H$ . Now we assign the uniform

distribution  $\mu$  on  $\mathcal{N}_{S,H}$  to the distribution vector  $P$ , and consider the Bayes estimator  $\hat{P}^B$  of  $P$  under the prior  $\mu$ . We first compute the posterior distribution of  $P$  given an observation vector  $\mathbf{X}$ . Denote by  $N(\mathbf{X})$  the number of different symbols (excluding the last symbol) appearing in the observation vector  $\mathbf{X}$ , and we assume without loss of generality that the first  $N(\mathbf{X})$  symbols appear in the sample. Then it is straightforward to show that the posterior distribution of  $P$  on  $\mathbf{X}$  is uniform in the following set  $\mathcal{N}_{\mathbf{X}} \subset \mathcal{N}_{S,H}$ :  $P \in \mathcal{N}_{\mathbf{X}}$  if and only if  $p_S = 1 - \delta, p_i = \frac{\delta}{S'}, 1 \leq i \leq N(\mathbf{X}), p_j \in \{0, \frac{\delta}{S'}\}, N(\mathbf{X}) + 1 \leq j \leq kS'$ , and

$$\left| \left\{ j : N(\mathbf{X}) + 1 \leq j \leq kS', p_j = \frac{\delta}{S'} \right\} \right| = S' - N(\mathbf{X}). \quad (78)$$

Hence, the Bayes estimator  $\hat{P}^B(\mathbf{X}) = (a_1, a_2, \dots, a_S)$  should minimize the posterior  $\ell_1$  risk given  $\mathbf{X}$  expressed as

$$R_B(\mathbf{X}, H, n, \mu) \triangleq \sum_{i=1}^{N(\mathbf{X})} \left| \frac{\delta}{S'} - a_i \right| + |1 - \delta - a_S| + \sum_{j=N(\mathbf{X})+1}^{kS'} \left[ \frac{(k-1)S' |a_j|}{kS' - N(\mathbf{X})} + \frac{S' - N(\mathbf{X})}{kS' - N(\mathbf{X})} \left| \frac{\delta}{S'} - a_j \right| \right], \quad (79)$$

and the solution is  $a_1 = \dots = a_{N(\mathbf{X})} = \frac{\delta}{S'}, a_{N(\mathbf{X})+1} = \dots = a_{kS'} = 0$  and  $a_S = 1 - \delta$ , and

$$R_B(\mathbf{X}, H, n, \mu) = \left( 1 - \frac{N(\mathbf{X})}{S'} \right) \delta. \quad (80)$$

In light of this, the Bayes risk can be expressed as  $R_B(H, n, \mu) = \mathbb{E}[R_B(\mathbf{X}, H, n, \mu)]$ , where the expectation is taken with respect to  $\mathbf{X}$ , and by (80) we only need to compute  $\mathbb{E}N(\mathbf{X})$ . Due to the symmetry of the Multinomial distribution, we can assume without loss of generality that  $\mathbf{X} \sim \text{Multi}(n; \frac{\delta}{S'}, \dots, \frac{\delta}{S'}, 0, \dots, 0, 1 - \delta)$ , then

$$\begin{aligned} \mathbb{E}N(\mathbf{X}) &= \mathbb{E} \left[ \sum_{i=1}^{S'} \mathbb{1}(X_i > 0) \right] = \sum_{i=1}^{S'} \mathbb{P}(X_i > 0) \\ &= S' \cdot \left[ 1 - \left( 1 - \frac{\delta}{S'} \right)^n \right] \leq n\delta \end{aligned} \quad (81)$$

which yields

$$\begin{aligned} R_B(H, n, \mu) &= \mathbb{E}[R_B(\mathbf{X}, H, n, \mu)] \\ &= \left( 1 - \frac{\mathbb{E}N(\mathbf{X})}{S'} \right) \delta = \left( 1 - \frac{n\delta}{S'} \right) \delta. \end{aligned} \quad (82)$$

Fix  $c \in (0, 1)$ , we set  $\delta = \frac{cH}{\ln n} \leq c$ , and

$$S' = \delta \exp \left( \frac{H}{\delta} \right) \left( (1 - \delta)^{\frac{1}{\delta}} \right)^{1-\delta} \quad (83)$$

$$\geq \frac{cH}{\ln n} \cdot n^{\frac{1}{c}} \cdot (1 - c)^{\frac{1}{c}}. \quad (84)$$

Since the minimax risk is lower bounded by any Bayes risk,



we have

$$\inf_{\hat{P}} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 \geq R_B(H, n, \mu) \quad (85)$$

$$\geq \frac{cH}{\ln n} \cdot \left(1 - n^{1-\frac{1}{c}}(1-c)^{-\frac{1}{c}}\right), \quad (86)$$

which completes the proof of the lower bound in Theorem 4.

For the lower bound in Theorem 5, note that we need to constrain that the Bayes estimator  $\hat{P}^B$  form a probability mass, i.e.,  $\sum_{i=1}^S a_i = 1$  in the minimization process of (79). Defining  $\lambda \triangleq \frac{(k-2)S' + N(\mathbf{X})}{kS' - N(\mathbf{X})} \in [0, 1]$ , the derivations on the bottom of this page show that the solution becomes  $a_1 = \dots = a_{N(\mathbf{X})} = \frac{\delta}{S'}$ ,  $a_{N(\mathbf{X})+1} = \dots = a_{kS'} = \frac{\delta(S' - N(\mathbf{X}))}{S'(kS' - N(\mathbf{X}))}$  and  $a_S = 1 - \delta$ . Hence, the corresponding Bayes risk given  $\mathbf{X}$  is

$$R_B(\mathbf{X}, H, n, \mu) = \frac{2(k-1)S'}{kS' - N(\mathbf{X})} \left(1 - \frac{N(\mathbf{X})}{S'}\right) \delta \quad (93)$$

$$\geq \frac{2(k-1)\delta}{k} \left(1 - \frac{N(\mathbf{X})}{S'}\right). \quad (94)$$

Applying the similar steps, we can show that the overall Bayes risk is

$$R_B(H, n, \mu) = \mathbb{E} R_B(\mathbf{X}, H, n, \mu) \quad (95)$$

$$\geq \frac{2(k-1)cH}{k \ln n} \cdot \left(1 - n^{1-\frac{1}{c}}(1-c)^{-\frac{1}{c}}\right). \quad (96)$$

Since the minimax risk is lower bounded by any Bayes risk, we have

$$\inf_{\hat{P} \in \mathcal{M}} \sup_{P: H(P) \leq H} \mathbb{E}_P \|\hat{P} - P\|_1 \geq R_B(H, n, \mu) \quad (97)$$

$$\geq \frac{2(k-1)cH}{k \ln n} \cdot \left(1 - n^{1-\frac{1}{c}}(1-c)^{-\frac{1}{c}}\right), \quad (98)$$

which completes the proof of the lower bound in Theorem 5 by letting  $k \rightarrow \infty$ .

### III. ACKNOWLEDGEMENT

We would like to thank I. Sason, the associate editor, and three anonymous reviewers for their very helpful suggestions. J. Jiao is partially supported by a Stanford Graduate Fellowship.

### APPENDIX A AUXILIARY LEMMAS

The following lemma gives a sharp estimate of the Binomial mean absolute deviation.

**Lemma 4.** [11] For  $X \sim \mathcal{B}(n, p)$ , we have

$$\mathbb{E} \left| \frac{X}{n} - p \right| \leq \min \left\{ \sqrt{\frac{p(1-p)}{n}}, 2p \right\}. \quad (99)$$

Moreover, for  $p < 1/n$ , there is an identity

$$\mathbb{E} \left| \frac{X}{n} - p \right| = 2p(1-p)^n. \quad (100)$$

The following lemma gives some tail bounds for Poisson or Binomial random variables.

**Lemma 5.** [36, Exercise 4.7] If  $X \sim \text{Poi}(\lambda)$  or  $X \sim \mathcal{B}(n, \frac{\lambda}{n})$ , then for any  $0 < \delta < 1$ , we have

$$\mathbb{P}(X \geq (1+\delta)\lambda) \leq \left( \frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\lambda, \quad (101)$$

$$\mathbb{P}(X \leq (1-\delta)\lambda) \leq \left( \frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right)^\lambda \leq e^{-\delta^2 \lambda / 2}. \quad (102)$$

The following lemma presents the Hoeffding bound.

**Lemma 6.** [37] For independent and identically distributed random variables  $X_1, \dots, X_n$  with  $a \leq X_i \leq b$  for  $1 \leq i \leq n$ , denote  $S_n = \sum_{i=1}^n X_i$ , we have for any  $t > 0$ ,

$$\mathbb{P} \{|S_n - \mathbb{E}[S_n]|\} \geq t \leq 2 \exp \left( -\frac{2t^2}{n(b-a)^2} \right). \quad (103)$$

---


$$R_B(\mathbf{X}, H, n, \mu) \triangleq \sum_{i=1}^{N(\mathbf{X})} \left| \frac{\delta}{S'} - a_i \right| + \sum_{j=N(\mathbf{X})+1}^{kS'} \left[ \frac{(k-1)S'}{kS' - N(\mathbf{X})} |0 - a_j| + \frac{S' - N(\mathbf{X})}{kS' - N(\mathbf{X})} \left| \frac{\delta}{S'} - a_j \right| \right] + |1 - \delta - a_S| \quad (87)$$

$$\geq \sum_{i=1}^{N(\mathbf{X})} \left| \frac{\delta}{S'} - a_i \right| + \sum_{j=N(\mathbf{X})+1}^{kS'} \left| \frac{(k-1)S'}{kS' - N(\mathbf{X})} \cdot a_j + \frac{S' - N(\mathbf{X})}{kS' - N(\mathbf{X})} \cdot \left( \frac{\delta}{S'} - a_j \right) \right| + |1 - \delta - a_S| \quad (88)$$

$$= \sum_{i=1}^{N(\mathbf{X})} \left| \frac{\delta}{S'} - a_i \right| + \sum_{j=N(\mathbf{X})+1}^{kS'} \left| \lambda a_j + \frac{\delta}{kS' - N(\mathbf{X})} \cdot \left( 1 - \frac{N(\mathbf{X})}{S'} \right) \right| + |1 - \delta - a_S| \quad (89)$$

$$\geq \sum_{i=1}^{N(\mathbf{X})} \lambda \left| a_i - \frac{\delta}{S'} \right| + \sum_{j=N(\mathbf{X})+1}^{kS'} \left| \lambda a_j + \frac{\delta}{kS' - N(\mathbf{X})} \cdot \left( 1 - \frac{N(\mathbf{X})}{S'} \right) \right| + \lambda |a_S - 1 + \delta| \quad (90)$$

$$\geq \left| \sum_{i=1}^{N(\mathbf{X})} \lambda \left( a_i - \frac{\delta}{S'} \right) + \sum_{j=N(\mathbf{X})+1}^{kS'} \left( \lambda a_j + \frac{\delta}{kS' - N(\mathbf{X})} \cdot \left( 1 - \frac{N(\mathbf{X})}{S'} \right) \right) + \lambda (a_S - 1 + \delta) \right| \quad (91)$$

$$= \frac{2(k-1)S'}{kS' - N(\mathbf{X})} \left( 1 - \frac{N(\mathbf{X})}{S'} \right) \delta, \quad (92)$$

A non-negative loss function  $l(\cdot)$  on  $\mathbb{R}^p$  is called *bowl-shaped* iff  $l(u) = l(-u)$  for all  $u \in \mathbb{R}^p$  and for any  $c \geq 0$ , the sublevel set  $\{u : l(u) \leq c\}$  is convex. The following theorem is one of the key theorems in the definition of asymptotic efficiency.

**Theorem 6.** [2, Thm. 8.11] *Let the experiment  $(P_\theta, \theta \in \Theta)$  be differentiable in quadratic mean at  $\theta$  with nonsingular Fisher information matrix  $I_\theta$ . Let  $\psi(\cdot)$  be differentiable at  $\theta$ . Let  $\{T_n\}$  be any estimator sequence in the experiments  $(P_\theta^n, \theta \in \mathbb{R}^k)$ . Then for any bowl-shaped loss function  $l$ ,*

$$\sup_I \liminf_{n \rightarrow \infty} \sup_{h \in I} \mathbb{E}_{\theta + \frac{h}{\sqrt{n}}} l \left( \sqrt{n} \left( T_n - \psi \left( \theta + \frac{h}{\sqrt{n}} \right) \right) \right) \geq \mathbb{E} l(X), \quad (104)$$

where  $\mathcal{L}(X) = \mathcal{N}(0, \psi'(\theta) I_\theta^{-1} \psi'(\theta)^T)$ , and the first supremum is taken over all finite subsets  $I \subset \mathbb{R}^k$ .

The next lemma relates the minimax risk under the Poissonized model of an approximate probability distribution and that under the Multinomial model of a true probability distribution, where the set of approximate probability distribution is defined by

$$\mathcal{M}_S(\epsilon) \triangleq \left\{ P = (p_1, p_2, \dots, p_S) : p_i \geq 0, \left| \sum_{i=1}^S p_i - 1 \right| < \epsilon \right\}. \quad (105)$$

We define the minimax risk for Multinomial model with  $n$  observations on support size  $S$  for estimating  $P$  as

$$R(S, n) = \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_{\text{Multinomial}} \|\hat{P} - P\|_1, \quad (106)$$

and the corresponding minimax risk for Poissonized model for estimating an approximate distribution as

$$R_P(S, n, \epsilon) = \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S(\epsilon)} \mathbb{E}_{\text{Poissonized}} \|\hat{P} - P\|_1. \quad (107)$$

**Lemma 7.** *The minimax risks under the Poissonized model and the Multinomial model are related via the following inequality: for any  $\zeta \in (0, 1]$  and  $0 < \epsilon \leq \frac{\zeta}{2(1+\zeta)}$ , we have*

$$R(S, n) \geq R_P(S, (1 + \zeta)n, \epsilon) - \exp\left(-\frac{\zeta^2 n}{24}\right) - \epsilon. \quad (108)$$

The following lemma establishes the relationship of the  $R_P(S, n, \epsilon)$  and the Bayes risk under some prior  $\mu$ .

**Lemma 8.** *Assigning prior  $\mu$  to a non-negative vector  $P$ , denote the corresponding Bayes risk for estimating  $P$  under  $\ell_1$  loss by  $R_B(S, n, \mu)$ . If there exists a constant  $A > 0$  such that*

$$\mu \left\{ P : \sum_{i=1}^S p_i \leq A \right\} = 1, \quad (109)$$

then the following inequality holds:

$$R_P(S, n, \epsilon) \geq R_B(S, n, \mu) - 3A \cdot \mu(\mathcal{M}_S(\epsilon)^c). \quad (110)$$

## APPENDIX B

### PROOF OF COROLLARIES AND AUXILIARY LEMMAS

#### A. Proof of Corollary 2

Consider the discrete distribution  $P = (p_1, p_2, \dots, p_S)$  with cardinality  $S$ , and we take  $\theta = (p_1, p_2, \dots, p_{S-1})$  to be the free parameter. By definition, we know that the Fisher information matrix is

$$I_{i,j}(\theta) = \mathbb{E}_\theta \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p(x|\theta) \right], \quad 1 \leq i, j \leq S-1. \quad (111)$$

It is straightforward to obtain that

$$\begin{aligned} I_{i,j}(\theta) &= p_i \left[ -\frac{\partial^2}{\partial p_i^2} \ln p_i \right] \delta_{i,j} \\ &\quad + \left( 1 - \sum_{k=1}^{S-1} p_k \right) \left[ -\frac{\partial^2}{\partial p_i \partial p_j} \ln \left( 1 - \sum_{k=1}^{S-1} p_k \right) \right] \end{aligned} \quad (112)$$

$$= \frac{\delta_{i,j}}{p_i} + \frac{1}{p_S}. \quad (113)$$

where  $\delta_{i,j}$  equals one if  $i = j$  and zero otherwise. Hence, in matrix form we have

$$\mathbf{I}(\theta) = \mathbf{\Lambda} + \frac{1}{p_S} \mathbf{1} \mathbf{1}^T, \quad (114)$$

where  $\mathbf{\Lambda} \triangleq \text{diag}(p_1^{-1}, \dots, p_{S-1}^{-1})$ , and  $\mathbf{1} \triangleq (1, 1, \dots, 1)^T$  is a  $(S-1) \times 1$  column vector. According to the Woodbury matrix identity

$$(\mathbf{A} + \mathbf{U} \mathbf{C} \mathbf{V})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{U} (\mathbf{C}^{-1} + \mathbf{V} \mathbf{A}^{-1} \mathbf{U})^{-1} \mathbf{V} \mathbf{A}^{-1}, \quad (115)$$

we can take  $\mathbf{A} = \mathbf{\Lambda}$ ,  $\mathbf{U} = \mathbf{1}$ ,  $\mathbf{C} = p_S^{-1}$ ,  $\mathbf{V} = \mathbf{1}^T$  to obtain

$$\mathbf{I}(\theta)^{-1} = (\mathbf{\Lambda} + \frac{1}{p_S} \mathbf{1} \mathbf{1}^T)^{-1} \quad (116)$$

$$= \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1} \mathbf{1} (p_S + \mathbf{1} \mathbf{\Lambda}^{-1} \mathbf{1}^T)^{-1} \mathbf{1}^T \mathbf{\Lambda}^{-1} \quad (117)$$

$$= \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1} \mathbf{1} \mathbf{1}^T \mathbf{\Lambda}^{-1}. \quad (118)$$

After some algebra we can show that

$$[\mathbf{I}(\theta)^{-1}]_{i,j} = -p_i p_j + p_i \delta_{i,j}, \quad (119)$$

then by choosing  $l : \mathbb{R}^S \mapsto \mathbb{R}_+$  defined by  $l(\mathbf{X}) \triangleq \sum_{i=1}^S |X_i|$  and  $\psi((p_1, p_2, \dots, p_{S-1})) = (p_1, p_2, \dots, p_{S-1}, 1 - \sum_{i=1}^{S-1} p_i)$  in Theorem 6, for  $\mathcal{L}(\mathbf{X}) = \mathcal{N}(0, \psi'(\theta) \mathbf{I}(\theta)^{-1} \psi'(\theta)^T)$ ,

$$\mathbb{E} l(\mathbf{X}) = \sqrt{\frac{2}{\pi}} \sum_{i=1}^S \sqrt{p_i(1-p_i)}. \quad (120)$$

If we choose  $\theta = (1/S, 1/S, \dots, 1/S)$ , then for any estimator sequence  $\{T_n\}_{n=1}^\infty$ , Theorem 6 yields

$$\begin{aligned} \sup_I \liminf_{n \rightarrow \infty} \sup_{h \in I} \mathbb{E}_{\theta + \frac{h}{\sqrt{n}}} \sqrt{n} \left\| T_n - \left( \theta + \frac{h}{\sqrt{n}} \right) \right\|_1 \\ \geq \sqrt{\frac{2}{\pi}} \sum_{i=1}^S \sqrt{p_i(1-p_i)} = \sqrt{\frac{2(S-1)}{\pi}} \end{aligned} \quad (121)$$

and the proof is completed by noticing that

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \sqrt{n} \cdot \inf_{T_n} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|T_n - P\|_1 \\ & \geq \sup_I \liminf_{n \rightarrow \infty} \sup_{h \in I} \mathbb{E}_{\theta + \frac{h}{\sqrt{n}}} \sqrt{n} \left\| T_n - \left( \theta + \frac{h}{\sqrt{n}} \right) \right\|_1. \end{aligned} \quad (122)$$

### B. Proof of Lemma 7

By the definition of the minimax risk under the Multinomial model, for any  $\delta > 0$ , there exists an estimator  $\hat{P}_M(\mathbf{X}, S, n)$  such that

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P}_M(\mathbf{X}, S, n) - P\|_1 < R(S, n) + \delta, \quad \forall n. \quad (123)$$

Now we construct a new estimator under the Poissonized model, i.e., we set  $\hat{P}_P(\mathbf{X}, S) \triangleq \hat{P}_M(\mathbf{X}, S, n')$  where  $n' = \sum_{i=1}^S X_i \sim \text{Poi}(n \sum_{i=1}^S p_i)$ . Then we can obtain that for  $0 < \epsilon < \frac{\zeta}{2(1+\zeta)}$  and  $\zeta \in (0, 1)$ ,

$$R_P(S, n, \epsilon) \leq \sup_{P \in \mathcal{M}_S(\epsilon)} \mathbb{E}_P \|\hat{P}_P(\mathbf{X}, S) - P\|_1 \quad (124)$$

$$= \sum_{m=0}^{\infty} \sup_{P \in \mathcal{M}_S(\epsilon)} \mathbb{E}_P \|\hat{P}_M(\mathbf{X}, S, m) - P\|_1 \cdot \mathbb{P}(n' = m) \quad (125)$$

$$\leq \sum_{m=0}^{\infty} \sup_{P \in \mathcal{M}_S(\epsilon)} \left( \mathbb{E}_P \left\| \hat{P}_M(\mathbf{X}, S, m) - \frac{P}{\sum_{i=1}^S p_i} \right\|_1 + \left\| \frac{P}{\sum_{i=1}^S p_i} - P \right\|_1 \right) \cdot \mathbb{P}(n' = m) \quad (126)$$

$$\leq \sum_{m=0}^{\infty} (R(S, m) + \delta + \epsilon) \cdot \mathbb{P}(n' = m) \quad (127)$$

$$\leq \mathbb{P}(n' \leq \frac{n}{1+\zeta}) + R(S, \frac{n}{1+\zeta}) + \epsilon + \delta \quad (128)$$

$$\leq \exp(-\frac{\zeta^2 n}{24}) + R(S, \frac{n}{1+\zeta}) + \epsilon + \delta, \quad (129)$$

where we have used Lemma 5 in the last step. The desired result follows directly from (129) and the arbitrariness of  $\delta$ .

### C. Proof of Lemma 8

Denote the conditional prior  $\pi$  by

$$\pi(A) = \frac{\mu(A \cap \mathcal{M}_S(\epsilon))}{\mu(\mathcal{M}_S(\epsilon))}, \quad (130)$$

we consider the Bayes estimator  $\hat{P}'$  under prior  $\pi$  and the corresponding Bayes risk  $R'_B(S, n, \pi)$ . Due to our construction of the prior, we know that for all  $\mathbf{X}$ , the sum of all entries of  $\hat{P}'(\mathbf{X})$  will not exceed  $A$  almost surely. Since  $R_B(S, n, \mu)$  is

the Bayes risk under prior  $\mu$ , applying  $\hat{P}'$  yields

$$R_B(S, n, \mu) \leq \int \mathbb{E}_P \|\hat{P}' - P\|_1 \mu(dP) \quad (131)$$

$$= \int_{\mathcal{M}_S(\epsilon)} \mathbb{E}_P \|\hat{P}' - P\|_1 \mu(dP) + \int_{\mathcal{M}_S(\epsilon)^c} \mathbb{E}_P \|\hat{P}' - P\|_1 \mu(dP) \quad (132)$$

$$\leq \frac{1}{\mu(\mathcal{M}_S(\epsilon))} \int_{\mathcal{M}_S(\epsilon)} \mathbb{E}_P \|\hat{P}' - P\|_1 \pi(dP) + \int_{\mathcal{M}_S(\epsilon)^c} 2A \mu(dP) \quad (133)$$

$$= \frac{R'_B(S, n, \pi)}{\mu(\mathcal{M}_S(\epsilon))} + 2A(1 - \mu(\mathcal{M}_S(\epsilon))). \quad (134)$$

Since the Bayes risk serves as a lower bound for the minimax risk, i.e.,  $R_P(S, n, \epsilon) \geq R'_B(S, n, \pi)$ , we have

$$R_P(S, n, \epsilon) \geq R_B(S, n, \mu) - (2A + R_B(S, n, \mu))\mu(\mathcal{M}_S(\epsilon)^c). \quad (135)$$

Then the proof is completed by noticing that  $R_B(S, n, \mu) \leq A$ , for the risk of the null estimator  $\hat{P}(\mathbf{X}) = \mathbf{0}$  under prior  $\mu$  does not exceed  $A$ .

### REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
- [2] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge university press, 2000, vol. 3.
- [3] K. Marton and P. C. Shields, "Entropy and the consistent estimation of joint distributions," *The Annals of Probability*, vol. 22, no. 2, pp. 960–977, 1994.
- [4] H. Steinhaus, "The problem of estimation," *The Annals of Mathematical Statistics*, vol. 28, pp. 633–648, 1957.
- [5] S. Trybula, "Some problems of simultaneous minimax estimation," *The Annals of Mathematical Statistics*, vol. 29, pp. 245–253, 1958.
- [6] M. Rutkowska, "Minimax estimation of the parameters of the multivariate hypergeometric and multinomial distributions," *Zastos. Mat.*, vol. 16, pp. 9–21, 1977.
- [7] I. Olkin and M. Sobel, "Admissible and minimax estimation for the multinomial distribution and for independent binomial distributions," *The Annals of Statistics*, vol. 7, pp. 284–290, 1979.
- [8] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY: Springer-Verlag, 1998.
- [9] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The  $L_1$  View*. New York, NY: John Wiley, 1985.
- [10] P. Diaconis and S. Zabell, "Closed form summation for classical distributions: Variations on a theme of De Moivre," *Statistical Science*, vol. 6, no. 3, pp. 284–302, 1991.
- [11] D. Berend and A. Kontorovich, "A sharp estimate of the binomial mean absolute deviation with applications," *Statistics & Probability Letters*, vol. 83, no. 4, pp. 1254 – 1259, 2013.
- [12] S. Kamath, A. Orlitsky, V. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Proceedings of The 28th Conference on Learning Theory*, 2015, pp. 1066–1100.
- [13] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1972, pp. 175–194.
- [14] I. Diakonikolas, "Beyond histograms: Structure and distribution estimation," in *Workshop of the 46th ACM Symposium on Theory of Computing*, 2014.
- [15] A. Wald, *Statistical Decision Functions*. Wiley, 1950.
- [16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.
- [17] S. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun, "Efficient density estimation via piecewise polynomial approximation," in *46th ACM Symposium on Theory of Computing*, 2014, pp. 604–613.

- [18] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning k-modal distributions via testing," in *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, 2012, pp. 1371–1385.
- [19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.
- [20] I. Csiszar and J. Körner, "Information theory: Coding theorems for discrete memoryless channels," *Budapest: Akadémiai Kiadó*, 1981.
- [21] J. F. Silva and P. A. Parada, "Shannon entropy convergence results in the countable infinite case," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, July 2012, pp. 155–159.
- [22] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Struct. Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.
- [23] G. Valiant and P. Valiant, "Estimating the unseen: an  $n/\log n$ -sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the 43rd ACM Symposium on Theory of Computing*, 2011, pp. 685–694.
- [24] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *Information Theory, IEEE Transactions on*, vol. 61, no. 5, pp. 2835–2885, 2015.
- [25] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *available on arXiv*, 2014.
- [26] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [27] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *submitted to IEEE Transactions on Information Theory*, 2015.
- [28] —, "Beyond maximum likelihood: from theory to practice," *available on arXiv*, 2014.
- [29] G. Valiant and P. Valiant, "The power of linear estimators," in *IEEE 52nd Annual Symposium on Foundations of Computer Science*, Oct 2011, pp. 403–412.
- [30] P. H. Algoet and T. M. Cover, "A sandwich proof of the shannon-mcmillan-breiman theorem," *The Annals of Probability*, vol. 16, no. 2, pp. 899–909, 1988.
- [31] D. L. Donoho and J. M. Johnstone, "Minimax risk over  $\ell_p$ -balls for  $\ell_q$ -error," *Probability Theory and Related Fields*, vol. 99, pp. 277–303, 1994.
- [32] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 399, 1956, pp. 197–206.
- [33] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
- [34] T. T. Cai, "Minimax and adaptive inference in nonparametric function estimation," *Statistical Science*, vol. 27, no. 1, pp. 31–50, 2012.
- [35] J. A. Adell and P. Jodra, "Exact Kolmogorov and total variation distances between some familiar discrete distributions," *Journal of Inequalities and Applications*, vol. 2006, no. 1, pp. 1–8, 2006.
- [36] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.
- [37] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

**Jiantao Jiao** (S'13) received his B.Eng. degree with the highest honor in Electronic Engineering from Tsinghua University, Beijing, China in 2012, and a Master's degree in Electrical Engineering from Stanford University in 2014. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at Stanford University. He is a recipient of the Stanford Graduate Fellowship (SGF). His research interests include information theory and statistical signal processing, with applications in communication, control, computation, networking, data compression, and learning.

**Tsachy Weissman** (S'99-M'02-SM'07-F'13) graduated summa cum laude with a B.Sc. in electrical engineering from the Technion in 1997, and earned his Ph.D. at the same place in 2001. He then worked at Hewlett Packard Laboratories with the information theory group until 2003, when he joined Stanford University, where he is currently Professor of Electrical Engineering and incumbent of the STMicroelectronics chair in the School of Engineering. He has spent leaves at the Technion, and at ETH Zurich.

Tsachy's research is focused on information theory, compression, communication, statistical signal processing, the interplay between them, and their applications. He is recipient of several best paper awards, and prizes for excellence in research and teaching. He served on the editorial board of the IEEE TRANSACTIONS ON INFORMATION THEORY from Sept. 2010 to Aug. 2013, and currently serves on the editorial board of Foundations and Trends in Communications and Information Theory. He is Founding Director of the Stanford Compression Forum.

**Yanjun Han** (S'14) received his B.Eng. degree with the highest honor in Electronic Engineering from Tsinghua University, Beijing, China in 2015. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at Stanford University. His research interests include information theory and statistics, with applications in communications, data compression, and learning.