Minimax Estimation of Discrete Distributions

Yanjun Han
Tsinghua University
hanyj11@mails.tsinghua.edu.cn

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—We analyze the problem of discrete distribution estimation under ℓ_1 loss. We provide non-asymptotic upper and lower bounds on the maximum risk of the empirical distribution (the maximum likelihood estimator), and the minimax risk in regimes where the alphabet 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, a hard-thresholding estimator, whose threshold does not depend on the unknown upper bound H, is asymptotically minimax. We draw connections between our work and the literature on density estimation, entropy estimation, total variation distance (ℓ_1 divergence) estimation, joint distribution estimation in stochastic processes, normal mean estimation, and adaptive estimation.**

I. Introduction and Main Results

Given n independent samples from an unknown discrete probability distribution $P = (p_1, p_2, \cdots, p_S)$, with *unknown* support size S, we would like to estimate the distribution P under ℓ_1 loss. Equivalently, the problem is to estimate P based on the Multinomal random vector $(X_1, X_2, \ldots, X_S) \sim \mathsf{Multi}(n; p_1, p_2, \ldots, p_S)$.

A natural estimator of P is the Maximum Likelihood Estimator (MLE), also known as 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, and we analyze the estimation problem when the number of observations n goes to infinity.
2) *High dimensional asymptotics:* we let the alphabet size S and the number of observations n grow together, analyze the scaling under which consistent estimation is possible, and obtain the minimax rates.
3) *Infinite dimensional asymptotics:* the distribution P may have infinite alphabet size, but is constrained to have bounded entropy $H(P) \le H$, where

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

We remark that results in the first regime follow from the well-developed theory of asymptotic statistics [1, Chap. 8], and we include it here for completeness and comparison with other regimes. One motivation for considering the high dimensional and infinite dimensional asymptotics is that, in the modern era of big data, we can no longer 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 alphabet of Chinese language is at least $80,000$. Meanwhile, for distributions with extremely large alphabet sizes (such as the Chinese language), the number of frequent symbols are considerably smaller than the alphabet size. It motivates the third regime, in which we focus on distributions with finite entropy, but possibly extremely large alphabet sizes. Another key result that motivates the problem of discrete distribution estimation under bounded entropy constraint is Marton and Shields [2], who essentially showed that the entropy dictates the difficulty in estimating discrete distributions under ℓ_1 loss in stochastic processes.

We denote by $\mathcal{M}_S$ the set of all distributions of support size S. The ℓ_1 loss for estimating P using Q is defined as

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

where Q is not necessarily a probability mass function. The risk function for an estimator $\hat{P}$ in estimating P under ℓ_1 loss is defined as

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

where the expectation is taken with respect to the measure P. The maximum ℓ_1 risk of an estimator $\hat{P}$, and the minimax risk in estimating P are respectively defined as

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

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

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

Throughout this paper, we investigate the maximum risk of the MLE $R_{\text{maximum}}(\mathcal{P}; P_n)$ and the minimax risk $R_{\text{minimax}}(\mathcal{P})$ for various choices of $\mathcal{P}$. There are good reasons for focusing on the ℓ_1 loss, as we do in this paper. Other loss functions in distribution estimation, such as the ℓ_2 loss, have been extensively studied, while fewer results are known for the ℓ_1 loss. For the ℓ_2 loss, the minimax estimator is unique and depends on the alphabet size S [3, Pg. 349], which is highly

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

impractical in cases S is unknown. This fact partially motivates our focus on the ℓ_1 loss, which turns out to bridge our understanding of both parametric and nonparametric models. The ℓ_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 [4].

This paper is organized as follows. In Section II, we investigate the MLE and the minimax estimator in the preceding three regimes and state our main results. Various connections between our results and the literature are drawn in Section III. The Appendices outline proofs of some theorems, and we refer the readers to the journal version [5] for complete proofs. All logarithms in this paper are assumed to be in natural base.

II. Main results

We investigate the maximum ℓ_1 risk of the MLE $R_{\text{maximum}}(\mathcal{P}; P_n)$ and the minimax ℓ_1 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 $\mathbb{E}|X/n - p|$, where $X \sim \mathsf{B}(n, p)$ follows a Binomial distribution. Berend and Kontorovich [6] provided tight upper and lower bounds on the Binomial mean absolute deviation, and we summarize some key results in Lemma 1 of the Appendix. It is well-known that

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{\frac{S}{n}}. \tag{6}$$

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

A. Classical asymptotics

It follows easily from (6) that

$$\limsup_{n \to \infty} \sqrt{n} \cdot \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{S} < \infty. \tag{7}$$

Regarding the lower bound, the well-known Hájek-Le Cam local asymptotic minimax theorem (Theorem 4 in Appendix A) and corresponding achievability theorems [1, Lemma 8.14] show that the MLE is optimal (even in constants) in classical asymptotics. Concretely, one corollary of Theorem 4 in the Appendix shows that for a fixed $S \geq 2$,

$$\liminf_{n \to \infty} \sqrt{n} \cdot \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \sqrt{\frac{2}{\pi}} \cdot \frac{(S-1)^{\frac{3}{2}}}{S} > 0, \tag{8}$$

where the infimum is taken over all possible estimators.

B. High dimensional asymptotics

It follows easily from (6) that for $S = n/c$,

$$\limsup_{c \to \infty} \sqrt{c} \cdot \limsup_{n \to \infty} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq 1 < \infty. \tag{9}$$

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

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

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \begin{Bmatrix} \exp\left(-\frac{2(1+\zeta)n}{S}\right), & \frac{(1+\zeta)n}{S} \leq \frac{e}{16} \\ \frac{1}{8}\sqrt{\frac{eS}{(1+\zeta)n}}, & \frac{(1+\zeta)n}{S} > \frac{e}{16} \end{Bmatrix} - \exp(-\frac{\zeta^2 n}{24}) - 12 \exp\left(-\frac{\zeta^2 S}{32(\ln S)^2}\right), \tag{10}$$

where the infimum is taken over all possible estimators.

Theorem 1 implies the following minimax lower bound in high dimensional asymptotics, if we take $\zeta \to 0^+$.

Corollary 1 *For any constant $c > 0$, if $S = n/c$, the convergence rate of the maximum ℓ_1 risk is at least $c^{-\frac{1}{2}}$, i.e.,*

$$\liminf_{c \to \infty} \sqrt{c} \cdot \liminf_{n \to \infty} \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{\sqrt{e}}{8}, \tag{11}$$

where the infimum in both formulas is taken over all possible estimators.

Hence, we show that $S = n/c$ is the *critical scaling* in high dimensional asymptotics. In other words, if $n \ll S$, then no estimator for the distribution P can be consistent under ℓ_1 loss. This phenomenon has been observed in several papers, such as [6].

C. Infinite dimensional asymptotics

The non-asymptotic performance of MLE in the regime of bounded entropy is characterized in the following theorem.

Theorem 2 *The empirical distribution P_n satisfies that, for any $H > 0$ and $\eta > 1$,*

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 \leq \frac{2H}{\ln n - 2\eta \ln \ln n} + \frac{1}{(\ln n)^\eta}. \tag{12}$$

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

$$\sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 \geq \frac{2cH}{\ln n} \left(1 - ((1-c)n)^{-\frac{1}{c}}\right)^n. \tag{13}$$

It follows from Theorem 2 after taking $c \to 1^-$ to obtain the following Corollary.

Corollary 2 *For any $H > 0$, the MLE P_n satisfies*

$$\lim_{n \to \infty} \frac{\ln n}{H} \cdot \sup_{P: H(P) \leq H} \mathbb{E}_P \|P_n - P\|_1 = 2. \tag{14}$$

It implies that we not only have obtained the $(\ln n)^{-1}$ convergence rate of the asymptotic ℓ_1 risk of MLE, but also shown that the corresponding coefficient is exactly $2H$. We note that this logarithmic convergence rate is really slow, since the sample size needs to be squared to reduce the maximum ℓ_1 risk by a half. Also note that the maximum ℓ_1 risk is proportional to the entropy H, thus the smaller the entropy of a distribution, the easier it is to estimate.

Now we consider the minimax lower bound. Corollary 2 shows that the convergence rate for MLE is $(\ln n)^{-1}$. Given this slow rate, it is of utmost importance to obtain estimators such that the corresponding constant is small. We show that MLE cannot achieve the optimal constant. In the following theorem, an asymptotically minimax estimator is explicitly constructed.

Theorem 3 *For any $\eta > 1$, for the estimator defined as $\hat{P}(\mathbf{X}) = (g_n(X_1), g_n(X_2), \cdots, g_n(X_S))$ with*

$$g_n(X_i) = \frac{X_i}{n}\mathbb{1}\left(\frac{X_i}{n} > \frac{e^2(\ln n)^{2\eta}}{n}\right), \tag{15}$$

we have

$$\sup_{P:H(P)\le H} \mathbb{E}_P\|\hat{P}-P\|_1 \le \frac{H}{\ln n - \ln(2e^2) - 2\eta\ln\ln n} + \frac{1}{(\ln n)^{\eta}} + \frac{1}{n^{e^2-1}(\ln n)^{2\eta}} + \frac{1}{2e^2 n^{\frac{e^2}{4}-1}(\ln n)^{2\eta}}. \tag{16}$$

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

$$\begin{aligned}&\inf_{\hat{P}}\sup_{P:H(P)\le H} \mathbb{E}_P\|\hat{P}-P\|_1\\ &\ge \frac{cH}{\ln(2n)}\exp\left(-e(1-c)^{-\frac{1}{c}}(2n)^{1-\frac{1}{c}}\right) - 6\left(1+\frac{cH}{\ln n}\right)\\ &\cdot\left[\exp\left(-\frac{((1-c)n)^{\frac{1}{c}}}{16ceH(\ln n)^3}\right) + \exp\left(-\frac{c^3H((1-c)n)^{\frac{1}{c}}}{e(\ln n)^3}\right)\right]\\ &- \exp(-\frac{n}{24}) - \frac{1}{4(\ln n)^2}\end{aligned} \tag{17}$$

where the infimum is taken over all possible estimators.

Theorem 3 presents both a non-asymptotic achievable maximum ℓ_1 risk and a non-asymptotic lower bound of the minimax ℓ_1 risk, and it is straightforward to verify that the upper bound and lower bound coincide asymptotically by choosing $c \to 1^-$. As a result, the asymptotic minimax ℓ_1 risk is characterized in the following corollary.

Corollary 3 *For any $H > 0$, the asymptotic minimax risk is $\frac{H}{\ln n}$:*

$$\liminf_{n\to\infty} \frac{\ln n}{H} \cdot \inf_{\hat{P}} \sup_{P:H(P)\le H} \mathbb{E}_P\|\hat{P}-P\|_1 = 1, \tag{18}$$

and the estimator $\hat{P}$ in Theorem 3 is asymptotically minimax.

In light of Corollaries 2 and 3, the asymptotic minimax ℓ_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 $(\ln n)^{-1}$, 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.

III. Discussions

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 Gyorfi [4] for an excellent overview. The conclusion that it is necessary and sufficient to use $n \gg S$ samples to consistently estimate an arbitrary discrete distribution with alphabet size S has appeared in the literature [7], 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 ℓ_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 ℓ_1 norm [8, Thm. 17.3.3], i.e., if $\|P-Q\|_1 \le 1/2$, then

$$|H(P)-H(Q)| \le -\|P-Q\|_1 \ln\frac{\|P-Q\|_1}{S}, \tag{19}$$

where S is the alphabet size. At first glance, it seems to suggest that the estimation of entropy can be reduced to estimation of discrete distributions under ℓ_1 loss. However, this question is far more complicated than it appears.

First, people have already noticed that this near-Lipschitz continuity result only holds on finite alphabets [9]. An evident proof of this fact is the following result by Antos and Kontoyiannis [10] on entropy estimation over countably infinite alphabet sizes.

Remark 1 *Among all discrete sources with finite entropy and $\mathsf{Var}(-\ln p(X)) < \infty$, 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*

$$\limsup_{n\to\infty} \frac{\mathbb{E}_P|H_n - H|}{a_n} = \infty. \tag{20}$$

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 ℓ_1 distance in the infinite alphabet setting, then Corollary 2 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 $n \gg \frac{S}{\ln S}$ samples, a result first discovered by Valiant and Valiant [11], then recovered by the present authors using a different approach in [12]. Since it is shown in Corollary 1 that we need $n \gg S$ 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. 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 [13] that this estimator again requires $n \gg S$ samples.

3) ℓ_1 divergence estimation between two distributions: Now we turn to the estimation problem for the ℓ_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 distribution to be deterministic, the problem of ℓ_1 divergence estimation seems a perfect dual to the distribution estimation problem under ℓ_1 loss. However, compared to the required sample complexity $n \gg S$ in the distribution estimation problem under ℓ_1 loss, the minimax sample complexity for estimating the ℓ_1 divergence between two arbitrary distributions is $n \gg \frac{S}{\ln S}$ samples, a result first discovered by Valiant and Valiant [11], then recovered and extended using a different approach by the present authors in [14] who also obtained the minimax rates. Hence, this result shows that it is easier to estimate the ℓ_1 divergence than to estimate the distribution with a vanishing ℓ_1 risk. Note that for distribution estimation, for each symbol we need to obtain a good estimate for p_i in terms of the ℓ_1 risk, while for ℓ_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: Marton and Shields [2] 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 and only if $d \leq \frac{(1-\epsilon)\ln n}{H}$ for any $\epsilon > 0$. An intuitive explanation of this result is that when d is large enough, the d-tuples have effective alphabet size e^{dH}, hence it requires $n \sim e^{dH}$ samples to consistently estimate the joint d-tuple distribution. It remains any interesting question to explore the connection between the infinite dimensional asymptotics regime considered in this paper and the results by Marton and Shields [2].

5) Hard-thresholding estimator is asymptotically minimax: Corollary 3 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 ℓ_p ball under ℓ_q error, $p > 0, p < q, q \in [1, \infty)$, was first noticed by Donoho and Johnstone [15]. Following the rationale of the James-Stein shrinkage estimator, they proposed the soft- and hard-thresholding estimators for the normal mean given that we know *a priori* that the mean θ lies in a ℓ_p ball, $p \in (0, \infty)$. Note that the set $\{P : H(P) \leq H\}$ forms a ball similar to the ℓ_p ball, and the loss function is ℓ_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 3 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 [16].

Appendix A
Auxiliary Lemmas

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 asymptotic efficiency.

Theorem 4 *[1, Thm. 8.11] Let the experiment $(P_\theta, \theta \in \Theta)$ be differentiable in quadratic mean at θ with nonsingular Fisher information matrix I_θ. Let $\psi(\cdot)$ be differentiable at θ. 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,*

$$\begin{aligned}\sup_I \liminf_{n\to\infty} \sup_{h\in I} \mathbb{E}_{\theta+\frac{h}{\sqrt{n}}} l\left(\sqrt{n}\left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}}\right)\right)\right) \\ \geq \int l d\mathcal{N}(0, \psi'(\theta) I_\theta^{-1} \psi'(\theta)^T),\end{aligned} \tag{21}$$

here the first supremum is taken over all finite subsets $I \subset \mathbb{R}^k$.

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

Lemma 1 *[6] For $X \sim \mathsf{B}(n, p)$, we have*

$$\mathbb{E}\left|\frac{X}{n} - p\right| \leq \min\left\{\sqrt{\frac{p}{n}}, 2p\right\}. \tag{22}$$

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

$$\mathbb{E}\left|\frac{X}{n} - p\right| = 2p(1-p)^n. \tag{23}$$

Appendix B
Outlines of Proofs of Main Theorems

Due to space constraints, we will not discuss the proof of the lower bounds. We refer the readers to the journal version [5] for detailed proofs.

A. Analysis of MLE

Lemma 1 yields

$$\mathbb{E}_P\|P - P_n\|_1 = \sum_{i=1}^S \mathbb{E}_P\left|\frac{X_i}{n} - p_i\right| \leq \sum_{i=1}^S \sqrt{\frac{p_i}{n}} \leq \sqrt{\frac{S}{n}}.$$

For the upper bound in Theorem 2, we use Lemma 1 again and obtain

$$\mathbb{E}_P\|P-P_n\|_1 \le 2\sum_{p_i\le\frac{(\ln n)^{2\eta}}{n}} p_i + \frac{1}{\sqrt{n}}\sum_{p_i>\frac{(\ln n)^{2\eta}}{n}}\sqrt{p_i} \quad (24)$$

$$\le \frac{2H}{\ln n - 2\eta\ln\ln n} + \frac{1}{(\ln n)^\eta}. \quad (25)$$

For the lower bound, we consider the distribution $P = (\delta/S', \cdots, \delta/S', 1-\delta)$ with entropy H, then it is easy to obtain that for $c \in (0,1), \delta \le c$,

$$S' \ge \delta \exp\left(\frac{H}{\delta}\right)\cdot(1-c)^{\frac{1}{c}}. \quad (26)$$

We assume without loss of generality that S' is an integer. For $c \in (0,1)$, by choosing $\delta = cH/\ln n \le c$, we can check that $\frac{\delta}{S'} < \frac{1}{n}$, and the identity in Lemma 1 can be applied to obtain

$$\mathbb{E}_P\|P-P_n\|_1 \ge \sum_{i=1}^{S'} 2p_i(1-p_i)^n = 2\delta\left(1-\frac{\delta}{S'}\right)^n \quad (27)$$

$$\ge \frac{2cH}{\ln n}\left(1-((1-c)n)^{-\frac{1}{c}}\right)^n. \quad (28)$$

B. Analysis of the Estimator in Theorem 3

We first establish a lemma on the properties of $g_n(X)$ while omitting its proof. Define $\Delta_n = (\ln n)^{2\eta}/n$.

Lemma 2 *If $X \sim \mathsf{B}(n,p)$, we have*

$$\mathbb{E}\,|g_n(X)-p| \le \begin{cases} p + \left(\frac{np}{e(\ln n)^{2\eta}}\right)^{e^2\ln n} & p \le \Delta_n, \\ \sqrt{\frac{p}{n}} + p & \Delta_n < p < 2e^2\Delta_n, \\ \sqrt{\frac{p}{n}} + n^{-\frac{e^2}{4}} & p \ge 2e^2\Delta_n. \end{cases} \quad (29)$$

Combining these results, we conclude that

$$\begin{aligned} \mathbb{E}_P\|P-\hat{P}\|_1 \le & \sum_{p_i<2e^2\Delta_n} p_i + \sum_{p_i>\Delta_n}\sqrt{\frac{p_i}{n}} \\ & + \sum_{p_i\le\Delta_n}\left(\frac{np_i}{e(\ln n)^{2\eta}}\right)^{e^2\ln n} + \sum_{p_i\ge 2e^2\Delta_n} n^{-\frac{e^2}{4}} \quad (30) \\ \le & \left(\ln\frac{1}{2e^2\Delta_n}\right)^{-1}\cdot H + \frac{1}{\sqrt{n\Delta_n}} \\ & + e^{-e^2\ln n}\cdot\frac{n}{(\ln n)^{2\eta}} + n^{-\frac{e^2}{4}}\cdot\frac{1}{2e^2\Delta_n}. \end{aligned}$$

which is exactly the upper bound of Theorem 3.

References

[1] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[2] 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.

[3] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY: Springer-Verlag, 1998.

[4] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. New York, NY: John Wiley, 1985.

[5] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *submitted to IEEE Transactions on Information Theory*, 2014.

[6] 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.

[7] I. Diakonikolas, "Beyond histograms: Structure and distribution estimation," in *STOC'14 workshop*, 2014.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[9] 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.

[10] 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.

[11] G. Valiant and P. Valiant, "The power of linear estimators," in *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, Oct 2011, pp. 403–412.

[12] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[13] ——, "Maximum likelihood estimation of functionals of discrete distributions," *arXiv preprint arXiv:1406.6959*, 2014.

[14] J. Jiao, Y. Han, and T. Weissman, "Minimax estimation of divergence functions," *in preparation*, 2014.

[15] D. L. Donoho and J. M. Johnstone, "Minimax risk over ℓ_p-balls for ℓ_q-error," *Probability Theory and Related Fields*, vol. 99, pp. 277–303, 1994.

[16] T. T. Cai, "Minimax and adaptive inference in nonparametric function estimation," *Statistical Science*, vol. 27, no. 1, pp. 31–50, 2012.