Adaptive Estimation of Shannon Entropy

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

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—"To be considered for an 2015 IEEE Jack Keil Wolf ISIT Student Paper Award." We consider estimating the Shannon entropy of a discrete distribution P from n i.i.d. samples. Recently, Jiao, Venkat, Han, and Weissman, and Wu and Yang independently constructed approximation theoretic estimators that achieve the minimax L_2 rates in estimating entropy. Their estimators are consistent given $n \gg \frac{S}{\ln S}$ samples, where S is the alphabet size, and it is the best possible sample complexity. On the contrary, the Maximum Likelihood Estimator (MLE), which is the empirical entropy, requires $n \gg S$ samples.**

We aim to significantly refine the minimax results of existing work in this paper. To alleviate the pessimism of minimaxity, we adopt the adaptive estimation framework, and show that the minimax rate-optimal estimator in Jiao, Venkat, Han, and Weissman is an adaptive estimator, i.e., it achieves the minimax rates simultaneously over a nested sequence of subsets of distributions P, without knowing the alphabet size S or which subset P lies in. We also characterize the maximum risk of the MLE over this nested sequence, and show for every subset in the sequence, the performance of the minimax rate-optimal estimator with n samples is essentially that of the MLE with $n \ln n$ samples, thereby contributing another example to a general phenomenon discovered by Jiao, Venkat, Han, and Weissman.

I. Introduction

Shannon entropy $H(P)$, defined as

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

is one of the most fundamental quantities of information theory and statistics, which emerged in Shannon's 1948 masterpiece [1] as the answers to foundational questions of compression and communication.

Consider the problem of estimating Shannon entropy $H(P)$ from n i.i.d. samples. Classical theory is mainly concerned with the case where the number of samples $n \to \infty$ while the alphabet size S is fixed. In that scenario, the maximum likelihood estimator (MLE), $H(P_n)$, which plugs in the empirical distribution into the definition of entropy, is *asymptotically efficient* [2, Thm. 8.11, Lemma 8.14] in the sense of the Hájek convolution theorem [3] and the Hájek–Le Cam local asymptotic minimax theorem [4]. It is therefore not surprising to encounter the following quote from the introduction of Wyner and Foster [5] who considered entropy estimation:

> *"The plug-in estimate is universal and optimal not only for finite alphabet i.i.d. sources but also for finite alphabet, finite memory sources. On the other hand, practically as well as theoretically, these problems are of little interest. "*

In contrast, various modern data-analytic applications deal with datasets which does not fall into the regime of $n \to \infty$. In fact, in many applications the alphabet size S is comparable to, or even larger than the number of samples n. For example:

- Corpus linguistics: about half of the words in the Shakespearean canon appeared only once [6].
- Network traffic analysis: many customers or website users are seen a small number of times [7].
- Analyzing neural spike trains: natural stimuli generate neural responses of high timing precision resulting in a massive space of meaningful responses [8]–[10].

A. Existing literature

The problem of entropy estimation in the large alphabet regime (or non-asymptotic analysis) has been investigated extensively in various disciplines, which we refer to [11] for a detailed review. One recent breakthrough in this direction came from Valiant and Valiant [12], who constructed the first explicit entropy estimator whose sample complexity is $n \asymp \frac{S}{\ln S}$ samples, which they also proved to be necessary. It was also shown in [13] [14] that the MLE requires $n \asymp S$ samples, implying that MLE is strictly sub-optimal in terms of sample complexity.

However, the aforementioned estimators have not been shown to achieve the minimax L_2 rates. In light of this, Jiao et al. [11], and Wu and Yang in [15] independently developed schemes based on approximation theory, and obtained the minimax L_2 convergence rates for the entropy. Further more, Jiao et al. [11] proposed a general methodology for estimating functionals, and showed that for a wide class of functionals (including entropy, mutual information, and Renyi entropy), their methodology can construct minimax rate-optimal estimators whose performance with n samples are nearly that of the MLE with $n \ln n$ samples. They also obtained minimax L_2 rates for estimating a large class of functionals. On the practical side, Jiao et al. [16] showed that the minimax rate-optimal estimators introduced in [11] can lead to consistent and substantial performance boosts in various machine learning algorithms.

Recall that the minimax risk of estimating functional $F(P)$ is defined via $\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F} - F(P)\right)^2$, where $\mathcal{M}_S$ denotes all distributions with alphabet size S, and the infimum is taken with respect to all estimators $\hat{F}$. Correspondingly, the maximum risk of MLE $F(P_n)$, which evaluates the functional $F(\cdot)$ at the empirical distribution P_n, is defined via

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2}+\frac{\ln^2 S}{n}$ $\left(n \succeq \frac{S}{\ln S}\right)$ ([11], [15])	$\frac{S^2}{n^2}+\frac{\ln^2 S}{n}$ $(n \succeq S)$ [14]
$F_\alpha(P), 0<\alpha\leq\frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}}$ $\left(n \succeq S^{1/\alpha}/\ln S, \ln n \preceq \ln S\right)$ ([11])	$\frac{S^2}{n^{2\alpha}}$ $\left(n \succeq S^{1/\alpha}, \ln n \preceq \ln S\right)$ [14]
$F_\alpha(P), \frac{1}{2}<\alpha<1$	$\frac{S^2}{(n\ln n)^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$ $\left(n \succeq S^{1/\alpha}/\ln S\right)$ ([11])	$\frac{S^2}{n^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$ $\left(n \succeq S^{1/\alpha}\right)$ [14]
$F_\alpha(P), 1<\alpha<\frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)}$ $(S \succeq n\ln n)$ ([11])	$n^{-2(\alpha-1)}$ $(S \succeq n)$ [14]
$F_\alpha(P), \alpha\geq\frac{3}{2}$	n^{-1} [14]	n^{-1}

TABLE I: Comparison of the minimax L_2 rates and the L_2 rates of MLE in estimating $H(P)$ and $F_\alpha(P) \triangleq \sum_{i=1}^S p_i^\alpha$. Whenever there are two terms, the first term corresponds to squared bias, and the second term corresponds to variance. It is evident that one can obtain the minimax rates from the L_2 rates of MLE via replacing n with $n\ln n$ in the dominating (bias) terms.

$\sup_{P\in\mathcal{M}_S}\mathbb{E}_P\left(F(P_n)-F(P)\right)^2$. Table I in Jiao et al. [11] summaries the minimax L_2 rates and the L_2 rates of MLE in estimating $H(P)$ and $F_\alpha(P) \triangleq \sum_{i=1}^S p_i^\alpha$. Whenever there are two terms, the first term corresponds to squared bias, and the second term corresponds to variance. It is evident that one can obtain the minimax rates from the L_2 rates of MLE via replacing n with $n\ln n$ in the dominating (bias) terms. We adopt the following notation: $a_n \preceq b_n$ means $\sup_n a_n/b_n < \infty$, $a_n \succeq b_n$ means $b_n \preceq a_n$, $a_n \asymp b_n$ means $a_n \preceq b_n$ and $a_n \succeq b_n$, or equivalently, there exists two universal constants c, C such that

$$0<c<\liminf_{n\to\infty}\frac{a_n}{b_n}\leq\limsup_{n\to\infty}\frac{a_n}{b_n}<C<\infty. \tag{2}$$

B. Refined minimaxity: adaptive estimation

One concern the readers may have about results on minimax rates is that they are too pessimistic. Indeed, in the definition $\inf_{\hat{F}}\sup_{P\in\mathcal{M}_S}\mathbb{E}_P\left(\hat{F}-F(P)\right)^2$, we have considered the worst case distribution P over all possible distributions supported on S elements, and it would be disappointing if the estimator in Jiao et al. [11] fail to behave optimally when we consider distributions lying in subsets of $\mathcal{M}_S$. A usual approach to alleviate this concern is the adaptive estimation framework, which we briefly review below.

The primary approach to alleviate the pessimism of minimaxity in statistics is the construction of adaptive procedures, which was particularly emphasized in nonparametric statistics [17]. The goal of adaptive inference is to construct a single procedure that achieves optimality simultaneously over a collection of parameter spaces. Informally, an adaptive procedure automatically adjusts to the *unknown* parameter, and acts as if it knows the parameter lies in a specific subset of the whole parameter space. A common way to evaluate such a procedure is to compare its maximum risk over each subset of the parameter space in the collection with the corresponding minimax risk. If they are nearly equal, then we say such a procedure is *adaptive* with respect to that collection of subsets of the parameter space.

The primary results of this paper are twofold.

1) First, we show that the minimax rate-optimal entropy estimator in Jiao et al. [11] is adaptive with respect to the collection of parameter space $\mathcal{M}_S(H)$, where $\mathcal{M}_S(H) \triangleq \{P : H(P)\leq H, P\in\mathcal{M}_S\}$. Moreover, the estimator does not need to know S nor H, which is an advantage in practice since usually the alphabet size S nor an *a priori* upper bound on the true entropy $H(P)$ are known.
2) Second, we show that the sample size *enlargement* effect still holds in this adaptive estimation scenario. Table I demonstrates that in estimating various functionals, the performance of the minimax rate-optimal estimator with n samples is nearly that of the MLE with $n\ln n$ samples, which the authors termed "sample size enlargement" in [11]. We compute the maximum risk of the MLE over each $\mathcal{M}_S(H)$, and show that for every H, the performance of the estimator in [11] with n samples is still nearly that of the MLE with $n\ln n$ samples.

The facts listed above in this paper suggest that the estimator in Jiao et al. [11] is *optimal* in a very strong sense, for which we refer the readers to [11] for a detailed discussion on methodology behind their estimator, literature survey, and experimental results.

This paper is organized as follows. In Section II, we introduce our mathematical framework and recall the approximation theoretic estimator in [11]. In Section III, we investigate the MLE and the minimax estimator in the adaptive estimation regime, and state the main results. The Appendices outline proofs of some theorems, and we refer the readers to the journal version [18] for complete proofs. All logarithms in this paper are assumed to be in natural base.

II. Mathematical Framework and Estimator Construction

Before we discuss the main results, we would like to recall the construction of the entropy estimator in [11]. The approach is to tackle the estimation problem separately for the cases of "small p" and "large p" in $H(P)$ estimation, corresponding to treating regions where the functional is "nonsmooth" and "smooth" in different ways. Specifically, after we obtain the empirical distribution P_n, for each coordinate $P_n(i)$, if $P_n(i)\ll\ln n/n$, we (i) compute the best polynomial approximation for $-p_i\ln p_i$ in the regime $0\leq p_i\ll\ln n/n$, (ii) use the unbiased estimators for integer powers p_i^k to estimate

the corresponding terms in the polynomial approximation for $-p_i \ln p_i$ up to order $K_n \sim \ln n$, and (iii) use that polynomial as an estimate for $-p_i \ln p_i$. If $P_n(i) \gg \ln n/n$, we use the estimator $-P_n(i) \ln P_n(i) + \frac{1}{2n}$ to estimate $-p_i \ln p_i$. Then, we add the estimators corresponding to each coordinate.

To simplify the analysis, we utilize the Poisson sampling model rather than the Multinomial model, i.e., instead of setting $(Z_1, Z_2, \cdots, Z_S) \sim \mathsf{Multi}(n; p_1, \cdots, p_S)$, we first draw a random variable $N \sim \mathsf{Poi}(n)$, and then conduct N samples from the distribution P. It is equivalent to having a S-dimensional random vector $\mathbf{Z}$ such that each component Z_i in $\mathbf{Z}$ has distribution $\mathsf{Poi}(np_i)$, and all coordinates of $\mathbf{Z}$ are independent. It follows from [11], [15] that the minimax risks under the Multinomial model and the Poissonized model are essentially equivalent.

Moreover, for simplicity of analysis, we conduct the classical "splitting" operation [19] on the Poisson random vector $\mathbf{Z}$, and obtain two independent identically distributed random vectors $\mathbf{X} = [X_1, X_2, \ldots, X_S]^T, \mathbf{Y} = [Y_1, Y_2, \ldots, Y_S]^T$, such that each component X_i in $\mathbf{X}$ has distribution $\mathsf{Poi}(np_i/2)$, and all coordinates in $\mathbf{X}$ are independent. For each coordinate i, the splitting process generates a random variable T_i such that $T_i|\mathbf{Z} \sim \mathrm{B}(Z_i, 1/2)$, and assign $X_i = T_i, Y_i = Z_i - T_i$. All the random variables $\{T_i : 1 \le i \le S\}$ are conditionally independent given our observation $\mathbf{Z}$. We also note that for random variable X such that $nX \sim \mathsf{Poi}(np)$,

$$\mathbb{E} \prod_{r=0}^{k-1} \left(X - \frac{r}{n} \right) = p^k, \tag{3}$$

for any $k \in \mathbb{N}_+$.

For simplicity, we re-define $n/2$ as n, and denote

$$\hat{p}_{i,1} = \frac{X_i}{n}, \hat{p}_{i,2} = \frac{Y_i}{n}, \Delta = \frac{c_1 \ln n}{n}, K = c_2 \ln n, t = \frac{\Delta}{4}, \tag{4}$$

where c_1, c_2 are positive parameters to be specified later. Note that Δ, K, t are functions of n, where we omit the subscript n for brevity.

The estimator $\hat{H}$ in Jiao et al. [11] is constructed as follows.

$$\hat{H} \triangleq \sum_{i=1}^{S} \left[L_H(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} \le 2\Delta) + U_H(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} > 2\Delta) \right], \tag{5}$$

where

$$S_{K,H}(x) \triangleq \sum_{k=1}^{K} g_{k,H} (4\Delta)^{-k+1} \prod_{r=0}^{k-1} \left(x - \frac{r}{n} \right) \tag{6}$$

$$L_H(x) \triangleq \min \{ S_{K,H}(x), 1 \} \tag{7}$$

$$U_H(x) \triangleq I_n(x) \left(-x \ln x + \frac{1}{2n} \right). \tag{8}$$

We explain each equation in detail as follows.

1) Equation (5): Note that $\hat{p}_{i,1}$ and $\hat{p}_{i,2}$ are i.i.d. random variables such that $n\hat{p}_{i,1} \sim \mathsf{Poi}(np_i)$. We use $\hat{p}_{i,2}$ to determine whether we are operating in the "nonsmooth" regime or not. If $\hat{p}_{i,2} \le 2\Delta$, we declare we are in the "nonsmooth" regime, and plug in $\hat{p}_{i,1}$ into function $L_\alpha(\cdot)$. If $\hat{p}_{i,2} > 2\Delta$, we declare we are in the "smooth" regime, and plug in $\hat{p}_{i,1}$ into $U_\alpha(\cdot)$.
2) Equation (6):
The coefficients $r_{k,H}, 0 \le k \le K$ are coefficients of the best polynomial approximation of $-x \ln x$ over $[0, 1]$ up to degree K, i.e.,

$$\sum_{k=0}^{K} r_{k,H} x^k = \arg \min_{y(x) \in \mathsf{poly}_K} \sup_{x \in [0,1]} |y(x) - (-x \ln x)|, \tag{9}$$

where poly_K denotes the set of algebraic polynomials up to order K. Note that in general $g_{k,\alpha}$ depends on K, which we do not make explicit for brevity.
Then we define $\{g_{k,H}\}_{1 \le k \le K}$ as

$$g_{k,H} = r_{k,H}, 2 \le k \le K, g_{1,H} = r_{1,H} - \ln(4\Delta). \tag{10}$$

It can be shown that for $nX \sim \mathsf{Poi}(np)$,

$$\mathbb{E} S_{K,H}(X) = \sum_{k=1}^{K} g_{k,H} (4\Delta)^{-k+1} p^k \tag{11}$$

is a near-best polynomial approximation for $-p \ln p$ on $[0, 4\Delta]$. Thus, we can understand $S_{K,H}(X), nX \sim \mathsf{Poi}(np)$ as a random variable whose expectation is nearly [1] the best approximation of function $-x \ln x$ over $[0, 4\Delta]$.
3) Equation (7):
Any reasonable estimator for $-p \ln p$ should not exceed one. We cutoff $S_{K,H}(x)$ by upper bound 1, and define the function $L_H(x)$, which means "lower part".
4) Equation (8):
The function $U_H(x)$ (means "upper part") is nothing but a product of an interpolation function $I_n(x)$ and the bias-corrected MLE. The interpolation function $I_n(x)$ is defined as follows:

$$I_n(x) = \begin{cases} 0 & x \le t \\ g(x - t; t) & t < x < 2t \\ 1 & x \ge 2t \end{cases} \tag{12}$$

The following lemma characterizes the properties of the function $g(x; a)$ appearing in the definition of $I_n(x)$. In particular, it shows that $I_n(x) \in C^4[0, 1]$.

Lemma 1. *For the function $g(x; a)$ on $[0, a]$ defined as follows,*

$$\begin{aligned} g(x; a) \triangleq 126 \left(\frac{x}{a} \right)^5 - 420 \left(\frac{x}{a} \right)^6 + 540 \left(\frac{x}{a} \right)^7 \\ - 315 \left(\frac{x}{a} \right)^8 + 70 \left(\frac{x}{a} \right)^9, \end{aligned} \tag{13}$$

we have the following properties:

$$g(0; a) = 0, \quad g^{(i)}(0; a) = 0, 1 \le i \le 4 \tag{14}$$

$$g(a; a) = 1, \quad g^{(i)}(a; a) = 0, 1 \le i \le 4 \tag{15}$$

[1] Note that we have removed the constant term from the best polynomial approximation. It is to ensure that we assign zero to symbols we do not see.

The function $g(x;1)$ is depicted in Figure 1.

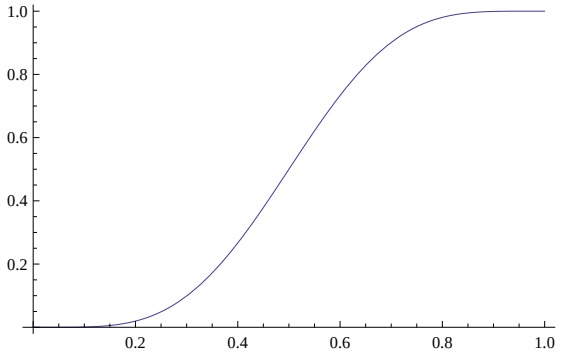

Fig. 1: The function $g(x;1)$ over interval $[0,1]$.

III. Main Results

Since $\sup_{P\in\mathcal{M}_S} H(P) = \ln S$, we will assume throughout this paper that $0 < H \leq \ln S$. Denote by $\mathcal{M}_S(H)$ the set of all discrete probability distributions P with support size $|\mathsf{supp}(P)| = S$ and entropy $H(P) \leq H$. We call an estimator $\hat{H} \equiv \hat{H}(\mathbf{Z})$ is within accuracy $\epsilon > 0$, if and only if

$$\sup_{P\in\mathcal{M}_S(H)} \left(\mathbb{E}_P|\hat{H} - H(P)|^2\right)^{\frac{1}{2}} \leq \epsilon. \tag{16}$$

For the plug-in estimator $H(P_n)$, the following theorem presents the non-asymptotic upper and lower bounds for the L_2 risk.

Theorem 1. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, then for the plug-in estimator $H(P_n)$, we have*

$$\sup_{P\in\mathcal{M}_S(H)} \mathbb{E}_P|H(P_n) - H(P)|^2 \asymp \begin{cases} \left(\frac{S}{n}\right)^2 + \frac{H\ln S}{n} & \text{if } S\ln S \leq enH, \\ \left[\frac{H}{\ln S}\ln\left(\frac{S\ln S}{nH}\right)\right]^2 & \text{otherwise.} \end{cases} \tag{17}$$

Note that the only assumption in Theorem 1 is that the upper bound H should be no smaller than a constant, which is a reasonable assumption to avoid the subtle case where the naive zero estimator $\hat{H} \equiv 0$ has a satisfactory performance. The minimum sample complexity of the plug-in approach can be immediately obtained from Theorem 1.

Corollary 1. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, the plug-in estimator $H(P_n)$ is within accuracy ϵ if and only if $n \succeq \left(S^{1-\frac{\epsilon}{H}} \cdot \frac{\ln S}{H}\right)$.*

Recall that it requires $n \succeq \left(\frac{S}{\epsilon}\right)$ samples for the MLE to achieve accuracy ϵ when there is no constraint on the entropy [14]. Hence, when the upper bound is loose, i.e., $H \asymp \ln S$, the minimum sample complexity in the bounded entropy case exactly reduces to [14], i.e., we cannot essentially improve the estimation performance. On the contrary, when the upper bound is tight, i.e., $H \ll \ln S$, the required sample complexity enjoyed a significant reduction, i.e., we only need sublinear samples for an accurate entropy estimation.

When it comes to the maximum L_2 risk, we conclude from Theorem 1 that the bounded entropy property helps only at the boundary, i.e., when n is close to S and H is small. Moreover, this help vanishes quickly as S increases: when $n = S^{1-\delta}$, the maximum L_2 risk will be at the order $(\delta H)^2$, which is the same risk achieved by the naive zero estimator when δ is not close to zero.

Is the plug-in estimator $H(P_n)$ optimal in the minimax sense? It has been shown in [11], [12], [15] that when there is no constraint on $H(P)$, i.e., $H = \ln S$, the answer is *negative*. What about subsets of $\mathcal{M}_S$, such as $\mathcal{M}_S(H)$? The following theorem characterizes the minimax L_2 rates over $\mathcal{M}_S(H)$.

Theorem 2. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, then*

$$\inf_{\hat{H}} \sup_{P\in\mathcal{M}_S(H)} \mathbb{E}_P|\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n\ln n)^2} + \frac{H\ln S}{n} & \text{if } S\ln S \leq enH\ln n, \\ \left[\frac{H}{\ln S}\ln\left(\frac{S\ln S}{nH\ln n}\right)\right]^2 & \text{otherwise.} \end{cases} \tag{18}$$

where the infimum is taken over all possible estimators. Moreover, the upper bound is adaptively achieved by the estimator in [11] under the Poissonized model without the knowledge of H nor S.

An immediate result on the sample complexity is as follows.

Corollary 2. *If $H \geq H_0 > 0$, where H_0 is a universal positive constant, the minimax rate-optimal estimator in [11] is within accuracy ϵ if and only if $n \succeq \left(\frac{1}{H}S^{1-\frac{\epsilon}{H}}\right)$.*

For the minimum sample complexity, we still distinguish H into two cases. Firstly, when $H \asymp \ln S$, the required sample complexity is $n \asymp \frac{S}{\epsilon\ln S}$, which exactly reduces to the minimax results with no constraint on entropy in [11]. Secondly, when $H \ll \ln S$, there is a significant improvement.

We also conclude from Theorem 2 that the bounded entropy constraint again helps only at the boundary, and this help vanishes quickly as S increases: when $n = S^{1-\delta}$, we do not have sufficient information to make inference, and the naive zero estimator is near-minimax.

To sum up, we have obtained the following conclusions.

1) The minimax rate-optimal entropy estimator in Jiao et al. [11] is adaptive with respect to the collection of parameter space $\mathcal{M}_S(H)$, where $\mathcal{M}_S(H) \triangleq \{P : H(P) \leq H, P \in \mathcal{M}_S\}$. Moreover, the estimator does not need to know S nor H, which is an advantage in practice since usually the alphabet size S nor an *a priori* upper bound on the true entropy $H(P)$ are known.
2) Second, the sample size *enlargement* effect still holds in this adaptive estimation scenario. Table I demonstrates that in estimating various functionals, the performance of the minimax rate-optimal estimator with n samples is nearly that of the MLE with $n\ln n$ samples, which the authors termed "sample size enlargement" in [11]. Theorems 1 and 2 show that over every $\mathcal{M}_S(H)$, the performance of the estimator in [11] with n samples is still nearly that of the MLE with $n\ln n$ samples.

IV. Future work

This paper applies adaptive estimation framework to strengthen the optimality properties of the approximation theoretic entropy estimator proposed in Jiao et al. [11]. We remark that the techniques in this paper are by no means constrained to entropy, and we believe similar statements are also true for estimators of $F_\alpha(P) = \sum_{i=1}^S p_i^\alpha$ in [11]. Furthermore, the fact that the sample size enlargement effect still holds in the adaptive estimation setting is very intriguing to the authors, and we believe there is a larger picture surrounding this theme to be explored.

V. Acknowledgments

The authors would like to express their most sincere gratitude to Dany Leviatan for valuable advice on the literature of approximation theory, in particular, for introducing one of the key steps in the proof of Lemma 2.

Appendix A
Outlines of Proofs of Main Theorems

A. Analysis of the MLE

1) Upper bound: We first consider the bias. For $p < \frac{1}{n}$ and $n\hat{p} \sim \mathsf{Poi}(np)$, the Poisson tail bounds show that $\hat{p} < \frac{\ln n}{n}$ with overwhelming probability, which leads to $-\hat{p}\ln\hat{p} \asymp \hat{p}\ln n$ and

$$\mathbb{E}[-\hat{p}\ln\hat{p}] \asymp \mathbb{E}[\hat{p}\ln n] = p\ln n. \tag{19}$$

Thus, the bias for $p_i < \frac{1}{n}$ will be $-p_i\ln(np_i)$, and one can show that summing up these terms subject to $H(P) \le H$ cannot exceed $A\ln\frac{S}{nA}$ with $A \asymp \frac{H}{\ln S}$. For $p_i \ge \frac{1}{n}$, the norm bound $\frac{1}{n}$ in [14] works.

As for the variance, the Efron-Stein inequality in [20] can be applied to yield

$$\mathsf{Var}(H(P_n)) \preceq \sum_{i=1}^S p_i(\ln p_i)^2 \preceq H\ln S. \tag{20}$$

2) Lower bound: As for the bias, it suffices to consider the distribution $(p, \cdots, p, 1-(S-1)p)$ with entropy H and then use the lower bounds in [14]. The minimax lower bound for variance follows from Le Cam's two-point method [21, Sec. 2.4.2].

B. Analysis of the minimax estimator

1) Upper bound: Most of the analysis goes through smoothly with the help of [11, Lem. 3, Lem. 4], and the remaining thing is to compute the bias

$$|\mathbb{E}S_{K,H}(\hat{p}) + p\ln p| = \left|4\Delta\sum_{k=1}^K g_{k,H}\left(\frac{p}{4\Delta}\right)^k + p\ln p\right| \tag{21}$$

$$= 4\Delta\left|\sum_{k=1}^K r_{k,H}\left(\frac{p}{4\Delta}\right)^k + \frac{p}{4\Delta}\ln\frac{p}{4\Delta}\right| \tag{22}$$

for $n\hat{p} \sim \mathsf{Poi}(np)$ and $p < \frac{1}{n\ln n}$. Let $p_K(x) = \sum_{k=1}^K r_{k,H}x^k$, we have the following lemma in approximation theory.

Lemma 2. *There exists a universal constant $D_p > 0$ such that for any $C \ge 1$, we have*

$$|p_K(x) - 2\ln K\cdot x| \le D_pCx, \quad \forall x \in \left[0, \frac{C}{K^2}\right]. \tag{23}$$

In light of Lemma 2, one can show that $|\mathbb{E}S_{K,H}(\hat{p}_i) + p_i\ln p_i| \preceq -p_i\ln(p_in\ln n)$ for all $p_i < \frac{1}{n\ln n}$, whose sum subject to $H(P) \le H$ cannot exceed $A\ln\frac{S}{An\ln n}$ with $A \asymp \frac{H}{\ln S}$.

2) Lower bound: The minimax lower bound is based on the so-called fuzzy hypothesis testing [21], and this idea is inspired by [11], [15].

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] J. Hájek, "A characterization of limiting distributions of regular estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 4, pp. 323–330, 1970.
[4] ——, "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.
[5] A. J. Wyner and D. Foster, "On the lower limits of entropy estimation," *IEEE Transactions on Information Theory, submitted for publication*, 2003.
[6] B. Efron and R. Thisted, "Estimating the number of unsen species: How many words did shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. pp. 435–447, 1976. [Online]. Available: http://www.jstor.org/stable/2335721
[7] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009, pp. 49–62.
[8] M. J. Berry, D. K. Warland, and M. Meister, "The structure and precision of retinal spike trains," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5411–5416, 1997.
[9] Z. F. Mainen and T. J. Sejnowski, "Reliability of spike timing in neocortical neurons," *Science*, vol. 268, no. 5216, pp. 1503–1506, 1995.
[10] R. R. d. R. van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, "Reproducibility and variability in neural spike trains," *Science*, vol. 275, no. 5307, pp. 1805–1808, 1997.
[11] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distribution," *submitted to IEEE Trans. Inf. Theory*, 2014.
[12] 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 annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.
[13] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
[14] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Non-asymptotic theory for the plug-in rule in functional estimation," *arXiv preprint arXiv:1406.6959*, 2014.
[15] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *available on arXiv*, 2014.
[16] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.
[17] T. T. Cai *et al.*, "Minimax and adaptive inference in nonparametric function estimation," *Statistical Science*, vol. 27, no. 1, pp. 31–50, 2012.
[18] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of shannon entropy," *submitted to the Annals of Statistics*, 2014.
[19] A. B. Tsybakov, "Aggregation and high-dimensional statistics," *Lecture notes for the course given at the École dété de Probabilités in Saint-Flour, URL http://www. crest. fr/ckfinder/userfiles/files/Pageperso/ATsybakov/Lecture_notes_SFlour. pdf*, vol. 16, p. 20, 2013.
[20] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
[21] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.