Maximum Likelihood Estimation of Information Measures

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Kartik Venkat
Stanford University
kvenkat@stanford.edu

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

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—The Maximum Likelihood Estimator (MLE) is widely used in estimating information measures, and involves "plugging-in" the empirical distribution of the data to estimate a given functional of the unknown distribution. In this work we propose a general framework and procedure to analyze the non-asymptotic performance of the MLE in estimating functionals of discrete distributions, under the worst-case mean squared error criterion.**

We show that existing theory is insufficient for analyzing the bias of the MLE, and propose to apply the theory of approximation using positive linear operators to study this bias. The variance is controlled using the well-known tools from the literature on concentration inequalities. Our techniques completely characterize the maximum L_2 risk incurred by the MLE in estimating the Shannon entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$, and $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha$ up to a multiplicative constant. As a corollary, for Shannon entropy estimation, we show that it is necessary and sufficient to have $n \gg S$ observations for the MLE to be consistent, where S represents the support size. In addition, we obtain that it is necessary and sufficient to consider $n \gg S^{1/\alpha}$ samples for the MLE to consistently estimate $F_\alpha(P), 0 < \alpha < 1$. The minimax rate-optimal estimators for both problems require $S/\ln S$ and $S^{1/\alpha}/\ln S$ samples, which implies that the MLE is strictly sub-optimal. When $1 < \alpha < 3/2$, we show that the maximum L_2 rate of convergence for the MLE is $n^{-2(\alpha-1)}$ for infinite support size, while the minimax L_2 rate is $(n \ln n)^{-2(\alpha-1)}$. When $\alpha \geq 3/2$, the MLE achieves the minimax optimal L_2 convergence rate n^{-1} regardless of the support size.

I. Introduction

The entropy, and related information measures have found numerous applications in information theory, statistics, machine learning, biology, neuroscience, image processing, linguistics, secrecy, ecology, physics, and finance, among others. Various inferential applications rely on data driven procedures to estimate these quantities (see, e.g. [1]–[6]). Consider the problem of estimating the Shannon entropy of an unknown discrete distribution P based on n i.i.d. samples. This problem has a rich history, which we refer to [7] for a review. One of the most widely used estimators for this purpose is the Maximum Likelihood Estimator (MLE), which is simply the empirical entropy, i.e. the entropy evaluated on the empirical distribution of the data. This is known as the plug-in principle in functional estimation, where a good point estimate of the parameter (distribution P) is used to construct an estimator for a *functional* of the parameter. The idea of using the MLE for estimating information measures of interest (in this case entropy), is not just intuitive, but has strong mathematical justification: *asymptotic efficiency*.

The beautiful theory of Hájek and Le Cam [8]–[10] showed that, as the number of observed samples grows without bound while the parameter dimension (support size) remains fixed, the MLE performs optimally in estimating any differentiable functionals under the benign LAN condition [10]. Thus, for finite dimensional problems, the problems of parameter and functional estimation are well understood in an asymptotic sense, and the MLE appears to be a very natural answer. But does it make sense to employ the MLE to estimate the entropy in most practical applications?

Unfortunately, while asymptotically optimal for entropy estimation, the MLE is by no means sacrosanct in real applications, especially in regimes where the support size is comparable to, or even larger than the number of observations. It was shown that the MLE for entropy is strictly sub-optimal in the large support regime [11]–[13]. Therefore, classical asymptotic theory does not satisfactorily address high dimensional settings, which are becoming increasingly important in the modern era of "big data". A satisfactory non-asymptotic theory should have two key components:

- Analysis: one should be able to analyze the non-asymptotic performance of estimators that are known to be asymptotically efficient, such as the MLE;
- Estimators: one should be able to construct estimators that are (near) optimal in a non-asymptotic sense.

The main contribution of this paper is to provide novel tools for analysis of plug-in estimators like the MLE, leaving the companion paper [7] to present new estimators that are shown to be *minimax rate-optimal* for a family of problems. The papers [7], [14] also demonstrate that employing the MLE in functional estimation can result in highly sub-optimal estimators in inferential applications.

A. Problem formulation

To illustrate our methodology, we shall focus on the following general problem. Suppose we observe n independent samples from an unknown discrete probability distribution $P = (p_1, p_2, \ldots, p_S)$, with *unknown* support size S, and would like to estimate a functional of the distribution of the form:

$$F(P) = \sum_{i=1}^{S} f(p_i), \tag{1}$$

where $f : (0, 1] \to \mathbb{R}$ is a continuous function. We shall focus on two concrete and well-motivated examples of functionals of this form. In particular, the Shannon entropy,

$$H(P) \triangleq \sum_{i=1}^{S} -p_i \ln p_i, \tag{2}$$

plays significant roles in information theory and serves as fundamental limits to various operational problems [15]. Another interesting class of functionals is $F_\alpha(P), \alpha > 0$:

$$F_\alpha(P) \triangleq \sum_{i=1}^{S} p_i^\alpha, \alpha > 0. \tag{3}$$

The α-moments of a distribution often emerge in various operational problems. In particular, the significance of $F_\alpha(P)$ can also be seen via the connection $H_\alpha(P) = \frac{\ln F_\alpha(P)}{1-\alpha}$, where $H_\alpha(P)$ is the Rényi entropy [16]. Like Shannon entropy, Rényi entropy is an important information measure emerging in an increasing variety of disciplines such as ecology, quantum information, information theory, and statistics, to name a few. Recently [17] has studied the complexity of estimating the Rényi entropy.

In this work, we analyze, for the above two classes of functionals, the non-asymptotic performance of the most natural plug-in rule for functionals of discrete distributions, i.e., the maximum likelihood estimator (MLE).

Definition 1. *The Maximum Likelihood Estimator (MLE) for the functional $F(P)$ in (1) is defined as:*

$$F(P_n) \triangleq \sum_{i=1}^{S} f(P_n(i)), \tag{4}$$

where P_n denotes the empirical distribution.

We use the conventional decision theoretic framework, and present matching upper and lower bounds (up to a constant) on the maximum L_2 risk of the MLE, i.e.

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P (F(P) - F(P_n))^2, \tag{5}$$

where $\mathcal{M}_S$ denotes the set of discrete distributions of support size S. Understanding the performance of the MLE serves two key purposes. First, the approach is a natural benchmark for comparing other more nuanced procedures for estimation of functionals. Second, performance analysis for the MLE reveals regimes where the problem is difficult, and motivates the development of improvements over the same, cf. [7], [14].

B. Bias of the MLE: an approximation theoretic perspective

In the entropy estimation literature, considerable effort has been devoted to understanding the non-asymptotic performance of the MLE $H(P_n)$ in estimating $H(P)$. One of the earliest investigations in this direction is due to Miller [18] in 1955, who showed that, for any fixed distribution P,

$$\mathbb{E}H(P_n) = H(P) - \frac{S-1}{2n} + O\left(\frac{1}{n^2}\right). \tag{6}$$

Equation (6) was later refined by Harris [19] using higher order Taylor series expansions to yield

$$\mathbb{E}H(P_n) = H(P) - \frac{S-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^{S} \frac{1}{p_i}\right) + O\left(\frac{1}{n^3}\right). \tag{7}$$

Harris's result reveals an undesirable consequence of the Taylor expansion method: one cannot obtain uniform bounds on the bias of the MLE. Indeed, the term $\sum_{i=1}^{S} \frac{1}{p_i}$ can be arbitrarily large for some distribution P. However, it is evident that both $H(P_n)$ and $H(P)$ are bounded above by $\ln S$, since the maximum entropy of any distribution supported on S elements is $\ln S$. Conceivably, for such a distribution P that would make $\sum_{i=1}^{S} \frac{1}{p_i}$ very large, we need to compute even higher order Taylor expansions to obtain more accuracy, but even with such efforts we can never obtain a uniform bias bound for all P.

We gain one of our key insights into the bias of the MLE by relating it to the approximation error induced by the *Bernstein polynomial approximation* of the function f. To see this, we first compute the bias of $F(P_n)$.

Lemma 1. *The bias of the estimator $F(P_n)$ is given by*

$$\begin{aligned}\mathsf{Bias}(F(P_n)) &\triangleq \mathbb{E}F(P_n) - F(P) \\ &= \sum_{i=1}^{S} \left(\sum_{j=0}^{n} f\left(\frac{j}{n}\right) \binom{n}{j} p_i^j (1-p_i)^{n-j} - f(p_i) \right).\end{aligned} \tag{8}$$

The bias term in (8) can be equivalently expressed as

$$\mathsf{Bias}(F(P_n)) = \sum_{i=1}^{S} \left(\sum_{j=0}^{n} f\left(\frac{j}{n}\right) B_{j,n}(p_i) - f(p_i) \right), \tag{9}$$

where $B_{j,n}(x) \triangleq \binom{n}{j} x^j (1-x)^{n-j}$ is the well-known Bernstein polynomial basis. Bernstein in 1912 [20] provided an insightful constructive proof of the Weierstrass theorem on approximation of continuous functions using polynomials, by showing that the Bernstein polynomial of any continuous function converges uniformly to that function.

From a functional analysis viewpoint, the Bernstein polynomial is an operator that maps a continuous function $f \in C[0,1]$ to another continuous function $B_n[f] \in C[0,1]$. This operator is linear in f, and is *positive* because $B_n[f]$ is also pointwise non-negative if f is pointwise non-negative. Apparently, bounding the approximation error incurred by the Bernstein polynomial is equivalent to bounding the bias of the MLE $f(X/n)$, where $X \sim \mathsf{Binomial}(n, x)$. As is discussed above, Taylor series expansions are not sufficient to satisfactorily analyze this bias. Fortunately, the theory of *approximation using positive linear operators* [21] provides us with sophisticated tools that serve admirably to this effect. A century ago, probability theory served Bernstein in breaking new ground in function approximation. It is therefore very satisfying that advancements in the latter have come full circle to help us understand probability theory better.

II. Main Results

Below, we present bounds on the maximum L_2 risk incurred by the MLE for estimating $H(P)$ and $F_\alpha(P)$.

Notation: $a \wedge b$ denotes $\min\{a, b\}$, and $\mathcal{M}_S$ denotes the collection of discrete distributions with support size S. For two non-negative series $\{a_n\}, \{b_n\}$, notation $a_n \lesssim b_n$ means that $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$. Notation $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Notation $a_n \gg b_n$ means that a_n asymptotically dominates b_n, i.e. $\forall C > 0, \liminf_{n\to\infty} \frac{a_n}{b_n} > C$.

Ditzian and Totik [22] introduced a class of modulus of smoothness, which proves to be extremely useful in characterizing the incurred approximation errors. For simplicity, for functions defined on $[0, 1]$, $\varphi(x) = \sqrt{x(1-x)}$, the second-order Ditzian–Totik modulus of smoothness is defined as

$$\omega_\varphi^2(f, t) \triangleq \sup\Big\{ \left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right|, \\ u, v \in [0, 1], |u - v| \le 2t\varphi\left(\frac{u+v}{2}\right) \Big\}. \tag{10}$$

Theorem 1 (Upper bounds on the risk of the MLE)**.** *For the functional $F_\alpha(P)$,*

1) $\alpha \ge 2$*:*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \le \left(\frac{\alpha(\alpha-1)}{n}\right)^2 + \frac{\alpha^2}{4n}. \tag{11}$$

2) $1 < \alpha < 2$*:*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \le \left(\frac{4}{n^{\alpha-1}} \wedge \frac{3S^{1-\alpha/2}}{n^{\alpha/2}} \wedge C_{\alpha,n}\frac{5S}{2n}\right)^2 + \frac{\alpha^2}{4n}, \tag{12}$$

where $C_{\alpha,n} \triangleq n\omega_\varphi^2(x^\alpha, n^{-1/2}) > 0$ satisfies $\limsup_{n\to\infty} C_{\alpha,n} < \infty$ for $1 < \alpha < 2$, and ω_φ^2 is the second-order Ditzian–Totik modulus of smoothness.

3) $1/2 \le \alpha < 1$*:*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \le \left(\frac{3S^{1-\alpha/2}}{2n^{\alpha/2}} \wedge \frac{5S}{2n^\alpha}\right)^2 + \left(\frac{10S^{2-2\alpha}}{n} + \frac{120}{\alpha^2}\left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}}\right)\right). \tag{13}$$

4) $0 < \alpha < 1/2$*:*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \le \left(\frac{3S^{1-\alpha/2}}{2n^{\alpha/2}} \wedge \frac{5S}{2n^\alpha}\right)^2 + \left(\frac{10S}{n^{2\alpha}} + \frac{120}{\alpha^2}\left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}}\right)\right). \tag{14}$$

5) *For the entropy $H(P)$,*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P)\right)^2 \le \left(\ln\left(1 + \frac{S-1}{n}\right)\right)^2 + \left(\frac{(\ln n)^2}{n} \wedge \frac{2(\ln S + 2)^2}{n}\right). \tag{15}$$

Moreover, in all the bounds presented above, the first term bounds the square of the bias, and the second term bounds the variance.

Theorem 1 has several interesting implications highlighted in the following corollaries.

Corollary 1. *For the functional $F_\alpha(P), \alpha > 1$, if $n \gg 1$, MLE is consistent.*

In words, if the functional $F_\alpha(P)$ is differentiable everywhere, then the number of samples required to make the maximum L_2 risk vanish has no dependence on the support size S. Results of this form have appeared in the literature, for example, Antos and Kontoyiannis [23] showed that it suffices to take $n \gg 1$ samples to consistently estimate $F_\alpha(P), \alpha \ge 2, \alpha \in \mathbb{Z}$. To our knowledge, Theorem 1 gives the first non-asymptotic result for estimation of $F_\alpha(P)$, α non-integer.

Corollary 2. *There exist universal convergence rates for $F_\alpha(P), \alpha > 1$. For any support size S (possibly infinite), we have,*

$$\sup_S \sup_{P\in\mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \lesssim \begin{cases} n^{-2(\alpha-1)} & 1 < \alpha < 3/2 \\ n^{-1} & \alpha \ge 3/2 \end{cases} \tag{16}$$

Corollary 2 implies that, when $\alpha \ge 3/2$, the MLE achieves the best possible rate $1/n$. However, when $1 < \alpha < 3/2$, the rate $n^{-2(\alpha-1)}$ is considerably slower. It turns out that the convergence rate $n^{-2(\alpha-1)}$ is in fact tight for the MLE when $1 < \alpha < 3/2$. Precisely, we have the following matching lower bound.

Theorem 2. *If $S = cn, 1 < \alpha < 3/2$, $c > 0$ is a positive constant, then*

$$\liminf_{n\to\infty} n^{2(\alpha-1)} \cdot \sup_{P\in\mathcal{M}_S} \mathbb{E}_P(F_\alpha(P_n) - F_\alpha(P))^2 \ge c_\alpha > 0, \tag{17}$$

where c_α only depends on α and c.

Interestingly, there exist estimators that demonstrate better convergence rates for estimating $F_\alpha(P), 1 < \alpha < 3/2$. [7] showed that the minimax L_2 rate in estimating $F_\alpha(P), 1 < \alpha < 3/2$ is in fact $(n \ln n)^{-2(\alpha-1)}$.

Let us now examine the case where $0 < \alpha < 1$, which is also another interesting regime that has not been characterized before. In this case, we observe significant gradation in the difficulty of the estimation problem. In particular, the relative scaling between the number of observations n and the support

size S for consistent estimation of $F_\alpha(P)$ exhibits a phase transition, encapsulated in the following.

Corollary 3. *For the functional $F_\alpha(P), 0 < \alpha < 1$, if $n \gg S^{\frac{1}{\alpha}}$, the MLE is consistent.*

For the region $0 < \alpha < 1$, the above corollary shows that $S^{\frac{1}{\alpha}}$ samples are sufficient for estimating $F_\alpha(P)$. Again, it turns out that this tightly characterizes the performance of the MLE. Precisely, we have the following lower bound for the performance of the MLE in this region.

Theorem 3. *If $n \geq S$, then*

1) $1/2 \leq \alpha < 1$*:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \gtrsim \frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}. \tag{18}$$

2) $0 < \alpha < 1/2$*:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \geq \frac{\alpha^2(1-\alpha)^2}{36n^{2\alpha}}(S-1)^2 \left(1 - \frac{1}{n}\right)^2. \tag{19}$$

Corollary 4. *The maximum L_2 risk of the MLE $F_\alpha(P_n)$ in estimating $F_\alpha(P)$ can be characterizes as follows when $n \geq S$:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P)\right)^2 \asymp \begin{cases} \frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n} & 1/2 < \alpha < 1 \\ \frac{S^2}{n^{2\alpha}} & 0 < \alpha \leq 1/2 \end{cases} \tag{20}$$

Corollary 4 follows directly from Theorem 1 and Theorem 3. In particular, it implies that it is necessary and sufficient to take $n \gg S^{1/\alpha}$ samples to consistently estimate $F_\alpha(P), 0 < \alpha < 1$ using MLE. Thus, as one might expect, the scale of the number of measurements required for consistent estimation increases as α decreases. When $\alpha \to 0$, the number of samples required for the MLE grows super-polynomially in S. On the other hand, for $\alpha > 1$, it suffices to take $n \gg 1$ samples independent of S for the MLE to be consistent.

We observe a sharp phase transition at $\alpha = 1$, as the sample size requirement shifts from $n \gg S^{\frac{1}{\alpha}}$ to $n \gg 1$, depending on whether α is in the left or right neighborhood of 1, respectively. Hence, $\alpha = 1$ is a critical point in that consistent estimation requires a number of measurements super-linear or constant in the size of the support according to whether $\alpha < 1$ or $\alpha > 1$.

A natural question arising in light of Corollary 3 and Theorem 3, respectively, is whether one can construct estimators that are better than the MLE in terms of required sample complexity for consistent estimation. The answer turns out to be affirmative, as we show in the companion paper [7]. In particular, the scheme we introduce therein is a consistent estimator of $F_\alpha(P)$ in the regime $0 < \alpha < 1$ when $n \gg \frac{S^{\frac{1}{\alpha}}}{\ln S}$, which is a logarithmic improvement in the sample complexity over the MLE. In fact, it is also shown that the scheme of [7] is minimax rate-optimal.

Let us now shift our focus to the case of entropy $H(P)$, which may be roughly intuitively viewed as the functional $F_\alpha(P)$ when $\alpha \to 1^-$. Theorem 1 implies the following corollary.

Corollary 5. *Case $H(P)$: The maximum L_2 risk of the MLE vanishes provided $n \gg S$.*

As it turns out, $n \asymp S$ is the optimal scaling for consistency of $H(P_n)$ in estimating $H(P)$. Further, we show that the same is true of the Miller–Madow bias-corrected estimator [18] defined as

$$H^{\text{MM}}(P_n) = H(P_n) + \frac{S-1}{2n}. \tag{21}$$

Theorem 4. *For the entropy $H(P)$, if $n \geq 15S$, then*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P)\right)^2 \geq \frac{1}{2}\left(\frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2}\right)^2 + c\frac{\ln^2 S}{n}. \tag{22}$$

Moreover, if $n \geq 15S$, the Miller–Madow bias-corrected estimator satisfies

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H^{\text{MM}}(P_n) - H(P)\right)^2 \geq \frac{1}{2}\left(\frac{S^2}{20n^2} - \frac{1}{12n^2}\right)^2 + c\frac{\ln^2 S}{n}, \tag{23}$$

where the positive constant $c > 0$ in both expressions do not depend on S or n.

Theorem 1 and Theorem 4 together imply the following corollary.

Corollary 6. *The maximum L_2 risk of the MLE $H(P_n)$ in estimating $H(P)$ is characterized as follows when $n \geq 15S$:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P)\right)^2 \asymp \frac{S^2}{n^2} + \frac{\ln^2 S}{n}. \tag{24}$$

We can raise an analogous question for the estimation of $H(P)$ as we did for $F_\alpha(P), 0 < \alpha < 1$: does there exist an entropy estimator with vanishing maximum L_2 risk with sublinear $n \ll S$ number of samples? The answer is affirmative, as was shown in [24], and more recently via a different scheme by the present authors in [7], and independently by Wu and Yang in [25], that one can construct estimators for which $n \gg S/\ln S$ samples suffice for consistent estimation. It was further shown in [24], [25] that this is the optimal order as the number of samples needed for any estimator to achieve vanishing maximum L_2 risk is at least of order $S/\ln S$. As in the case of $F_\alpha(P), 0 < \alpha < 1$, the MLE serves as a very good benchmark when we compare it with the minimax rate-optimal estimator.

To sum up our results regarding achievability and lower bounds on estimation of $F_\alpha(P)$ and $H(P)$, we have Table I.

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2} + \frac{\ln^2 S}{n} \quad \left(n \gtrsim \frac{S}{\ln S}\right)$ ([7], [25])	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n} \quad (n \gtrsim S)$ [13]
$F_\alpha(P), 0 < \alpha \leq \frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}} \quad \left(n \gtrsim S^{1/\alpha}/\ln S, \ln n \lesssim \ln S\right)$ [7]	$\frac{S^2}{n^{2\alpha}} \quad \left(n \gtrsim S^{1/\alpha}\right)$ [13]
$F_\alpha(P), \frac{1}{2} < \alpha < 1$	$\frac{S^2}{(n\ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n} \quad \left(n \gtrsim S^{1/\alpha}/\ln S\right)$ [7]	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n} \quad \left(n \gtrsim S^{1/\alpha}\right)$ [13]
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)} \quad (S \gtrsim n\ln n)$ [7]	$n^{-2(\alpha-1)} \quad (S \gtrsim n)$ [13]
$F_\alpha(P), \alpha \geq \frac{3}{2}$	n^{-1} [13]	n^{-1}

TABLE I: Summary of results in this paper and the companion [7]. When the L_2 rates have two terms, the first and second terms represent respectively the contributions of the bias and the variance. When there is a single term, only the dominant term is retained. Conditions for these results are presented in parentheses.

Table I demonstrates that the MLE cannot achieve the minimax risk for estimation of $H(P)$, and $F_\alpha(P)$ when $0 < \alpha < 3/2$. In these cases, there exist strictly better estimators whose performance with n samples is roughly the same as that of the MLE with $n \ln n$ samples. In other words, the optimal estimators *enlarge* the effective sample size by a logarithmic factor.

To our knowledge, Paninski [11] was the first to have realized the connection between Bernstein polynomials and bias of MLE. In the same paper [11], Paninski demonstrated that if $n = cS$, where $c > 0$ is a constant, then the maximum squared bias of $H(P_n)$, and of the Miller–Madow bias-corrected estimator $H^{\mathrm{MM}}(P_n)$ would be bounded from zero. Theorem 4 is the first to provide precise non-asymptotic constants in lower bounding the maximum L_2 risk of $H(P_n)$ and $H^{\mathrm{MM}}(P_n)$.

In summary, our focus in this paper is on estimating functionals of discrete distributions. We reiterate that our techniques are equally applicable to the general plug-in rule of functional estimation in general statistical experiments. We implore the reader to refer to the full version of this manuscript [13] for detailed proofs of all stated results, as well as additional discussions pertaining to the rich history and context of plug-in rules in functional estimation, as well as a more comprehensive treatment of approximation theoretic tools essential to analyze the bias incurred in functional estimation.

REFERENCES

[1] C. Olsen, P. E. Meyer, and G. Bontempi, "On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, no. 1, p. 308959, 2009.

[2] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 8, pp. 986–1004, 2003.

[3] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[4] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: a comprehensive study," *Journal of Cryptology*, vol. 24, no. 2, pp. 269–291, 2011.

[5] M. O. Hill, "Diversity and evenness: a unifying notation and its consequences," *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.

[6] F. Franchini, A. Its, and V. Korepin, "Rényi entropy of the XY spin chain," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 2, p. 025302, 2008.

[7] 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.

[8] J. Hájek, "A characterization of limiting distributions of regular estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 4, pp. 323–330, 1970.

[9] ——, "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.

[10] L. Le Cam, *Asymptotic methods in statistical decision theory*. Springer, 1986.

[11] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[12] ——, "Estimating entropy on m bins given fewer than m samples," *Information Theory, IEEE Transactions on*, vol. 50, no. 9, pp. 2200–2203, 2004.

[13] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *arXiv preprint arXiv:1406.6959*, 2014.

[14] ——, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.

[15] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[16] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[17] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy," in *SODA*, 2015.

[18] G. A. Miller, "Note on the bias of information estimates," *Information Theory in Psychology: Problems and Methods*, vol. 2, pp. 95–100, 1955.

[19] B. Harris, "The statistical estimation of entropy in the non-parametric case," DTIC Document, Tech. Rep., 1975.

[20] S. Bernstein, "Collected works: Vol 1. constructive theory of functions (1905-1930), English translation," *Atomic Energy Commission, Springfield, Va*, 1958.

[21] R. Paltanea, *Approximation theory using positive linear operators*. Springer, 2004.

[22] Z. Ditzian and V. Totik, *Moduli of smoothness*. Springer, 1987.

[23] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[24] 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.

[25] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *arXiv preprint arXiv:1407.0381*, 2014.