Does Dirichlet Prior Smoothing Solve the Shannon Entropy Estimation Problem?

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

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—The Dirichlet prior is widely used in estimating discrete distributions and functionals of discrete distributions. In terms of Shannon entropy estimation, one approach is to plug-in the Dirichlet prior smoothed distribution into the entropy functional, while the other one is to calculate the Bayes estimator for entropy under the Dirichlet prior for squared error, which is the conditional expectation. We show that in general they do *not* improve over the maximum likelihood estimator, which plugs-in the empirical distribution into the entropy functional. No matter how we tune the parameters in the Dirichlet prior, this approach cannot achieve the minimax rates in entropy estimation, as recently characterized by Jiao, Venkat, Han, and Weissman [1], and Wu and Yang [2]. The performance of the minimax rate-optimal estimator with n samples is essentially *at least* as good as that of the Dirichlet smoothed entropy estimators with $n \ln n$ samples.**

We harness the theory of approximation using positive linear operators for analyzing the bias of plug-in estimators for general functionals under arbitrary statistical models, thereby further consolidating the interplay between these two fields, which was thoroughly exploited by Jiao, Venkat, Han, and Weissman [3] in estimating various functionals of discrete distributions. We establish new results in approximation theory, and apply them to analyze the bias of the Dirichlet prior smoothed plug-in entropy estimator. This interplay between bias analysis and approximation theory is of relevance and consequence far beyond the specific problem setting in this paper.

I. Introduction

One of the key tasks of information theory is to characterize fundamental limits of operational problems by means of *information measures*, namely, functionals of probability distributions or conditional distributions (channels). Among the most fundamental of such functionals are the *Shannon entropy* [4],

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

and the *mutual information*, which emerged in Shannon's 1948 masterpiece [4] as the answers to the most fundamental questions of compression and communication.

In addition to their prominent operational roles in the traditional realms of information theory, information measures have found numerous applications in fields ranging from statistics, machine learning, to biology, ecology, to name a few. In most real-world inferential applications, the true underlying distribution that generates the data is unknown. Thus the applications rest upon data-driven procedures for accurately estimating information measures.

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* [5, Thm. 8.11, Lemma 8.14]. It is therefore not surprising to encounter the following quote from the introduction of Wyner and Foster [6] 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 do 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, e.g., more than half of the words in the collected works of Shakespeare appeared only once [7].

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 [1] for a detailed review and apologize for omitting a large body of references due to space constraint. One recent breakthrough in this direction came from Valiant and Valiant [8], who constructed the first *explicit* entropy estimator with sample complexity $n \asymp \frac{S}{\ln S}$, which they also proved to be necessary. It was also shown in [3], [9] that the MLE requires $n \asymp S$ samples.

Later, Jiao et al. [1], and Wu and Yang [2] independently developed schemes based on approximation theory, and obtained the minimax L_2 convergence rates for the entropy. Furthermore, Jiao et al. [1] proposed a general methodology for estimating functionals, and showed that for a wide class of functionals (including entropy, mutual information, and Rényi entropy), this methodology can construct minimax rate-optimal estimators whose performance with n samples are nearly that of the MLE with $n \ln n$ samples. It was argued in [1] that the "only" approach that can achieve the minimax rates for entropy must either implicitly or explicitly conduct best polynomial approximation as [1] did. A question that

arises naturally then is whether modifications of the plug-in approach, such as the Dirichlet prior smoothing ideas, can at least improve over the plug-in idea in terms of maximum risk. This paper answers this question negatively.

Dirichlet smoothing may have two connotations in the context of entropy estimation:

- [10], [11] One first obtains a Bayes estimate for the discrete distribution P, which we denote by $\hat{P}_B$, and then plugs it in the entropy functional to obtain the entropy estimate $H(\hat{P}_B)$.
- [12], [13] One calculates the Bayes estimate for entropy $H(P)$ under Dirichlet prior for squared error. The estimator is the conditional expectation $\mathbb{E}[H(P)|\mathbf{X}]$, where $\mathbf{X}$ represents the samples.

We show in the present paper that neither approach results in improvements over the MLE in the large alphabet regime. Specifically, these approaches require at least $n \gg S$ to be consistent, while the minimax rate-optimal estimators such as the ones in [1] [2] only need $n \gg \frac{S}{\ln S}$ to achieve consistency.

A main motivation for the present paper, beyond that discussed above, is to demonstrate the power of *approximation theory using positive linear operators* for bounding the bias of plug-in estimators for functionals of parameters under arbitrary statistical models. It was explicitly pointed out in Jiao et al. [3] that under mild conditions, the problem of bias analysis of plug-in estimators for functionals from arbitrary finite dimensional statistical models is equivalent to approximation theory using positive linear operators, a subfield of approximation theory which has been developing for more than a century. Applying advanced tools from positive linear operator theory [14], Jiao et al. [3] obtained tight non-asymptotic characterizations of maximum L_2 risks for MLE in estimating a variety of functionals of probability distributions. In this paper, we contribute to the general positive linear operator theory [14], and use the Dirichlet smoothing prior plug-in estimator as an example to demonstrate the efficacy of this general theory in dealing with analysis of the bias in estimation problems. We believe this connection has far reaching implications beyond analyzing bias in statistical estimation, which itself is an important problem.

The remainder of this paper is organized as follows. In Section II, we introduce the Dirichlet smoothing, L_2 risk analysis and approximation theory using positive linaer operators. In Section III, we investigate the L_2 risk of the Dirichlet smoothed entropy estimator and state our main results. Section IV applies the approximation theory using positive linear operators to analyze the bias. We refer the readers to the journal version [15] for complete details of the proofs.

II. Preliminaries: Dirichlet Smoothing, L_2 Risk Analysis, and Positive Linear Operators

A. Dirichlet Smoothing

The Dirichlet smoothing is widely used in practice to overcome the undersampling problem, i.e., one observes too few samples from a distribution P. The probability density function of Dirichlet distribution with order $S \geq 2$ and parameters $\alpha_1, \ldots, \alpha_S > 0$ is given by

$$f(x_1, \cdots, x_S; \alpha_1, \cdots, \alpha_S) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{S} x_i^{\alpha_i - 1} \tag{2}$$

on the open $(S-1)$-dimensional simplex $\{x_1, x_2, \cdots, x_S \in \mathbb{R}_+, \sum_{i=1}^{S} x_i = 1\}$ and zero elsewhere, and the normalizing constant is the multinomial Beta function.

Assuming the unknown distribution P follows prior distribution $P \sim \mathrm{Dir}(\boldsymbol{\alpha})$, and we observe a vector $\mathbf{X} = (X_1, X_2, \ldots, X_S)$ with Multinomial distribution $\mathsf{Multi}(n; p_1, p_2, \ldots, p_S)$, then the posterior mean (conditional expectation) of p_i given $\mathbf{X}$ is given by [16, Example 5.4.4]

$$\delta_i(\mathbf{X}) \triangleq \mathbb{E}[p_i|\mathbf{X}] = \frac{\alpha_i + X_i}{n + \sum_{i=1}^{S} \alpha_i}. \tag{3}$$

The estimator $\delta_i(\mathbf{X})$ is widely used in practice for various choices of α. For example, if $\alpha_i = \sqrt{n}/S$, then the corresponding $(\delta_1(\mathbf{X}), \delta_2(\mathbf{X}), \ldots, \delta_S(\mathbf{X}))$ is the minimax estimator for P under squared loss [16, Example 5.4.5]. Note that the estimator $\delta_i(\mathbf{X})$ subsumes the MLE $\hat{p}_i = \frac{X_i}{n}$ as a special case by taking the limit $\boldsymbol{\alpha} \to \mathbf{0}$.

The Dirichlet prior smoothed distribution estimate is denoted as $\hat{P}_B$, where

$$\hat{P}_B = \frac{n}{n + \sum_{i=1}^{S} \alpha_i} P_n + \frac{\sum_{i=1}^{S} \alpha_i}{n + \sum_{i=1}^{S} \alpha_i} \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{S} \alpha_i}. \tag{4}$$

Note that the *smoothed* distribution $\hat{P}_B$ can be viewed as a convex combination of the empirical distribution P_n and the *prior* distribution $\frac{\boldsymbol{\alpha}}{\sum_{i=1}^{S} \alpha_i}$. We call the estimator $H(\hat{P}_B)$ the *Dirichlet prior smoothed plug-in estimator*.

Another way to apply Dirichlet prior in entropy estimation is to compute the Bayes estimator for $H(P)$ under squared error, given that P follows Dirichlet prior. It is well known that the Bayes estimator under squared error is the conditional expectation. It was shown in Wolpert and Wolf [12] that

$$\hat{H}^{\mathsf{Bayes}} \tag{5}$$

$$\triangleq \mathbb{E}[H(P)|\mathbf{X}] \tag{6}$$

$$= \psi\left(\sum_{i=1}^{S}(\alpha_i + X_i) + 1\right) - \sum_{i=1}^{S}\left(\frac{\alpha_i + X_i}{\sum_{i=1}^{S}(\alpha_i + X_i)}\right)\psi(\alpha_i + X_i + 1), \tag{7}$$

where $\psi(z) \triangleq \frac{\Gamma'(z)}{\Gamma(z)}$ is the digamma function. We call the estimator $\hat{H}^{\mathsf{Bayes}}$ the *Bayes estimator under Dirichlet prior*.

B. Non-asymptotic analysis of L_2 risk

We adopt the conventional statistical decision theoretic framework in analyzing the performance of $H(\hat{P}_B)$. Denote by $\mathcal{M}_S$ all discrete distributions with support size S. We adopt

the minimax criterion, and evaluate the *maximum L_2 risk*

$$\begin{aligned}&\sup_{P\in\mathcal{M}_S}\mathbb{E}_P\left(H(P)-H(\hat{P}_B)\right)^2\\&=\sup_{P\in\mathcal{M}_S}\left[\left(\mathsf{Bias}(H(\hat{P}_B))\right)^2+\mathsf{Var}(H(\hat{P}_B))\right]\end{aligned}\tag{8}$$

where we have decomposed the L_2 risk into the bias part and the variance part:

$$\mathsf{Bias}(H(\hat{P}_B))\triangleq\mathbb{E}H(\hat{P}_B)-H(P)\tag{9}$$

$$\mathsf{Var}(H(\hat{P}_B))\triangleq\mathbb{E}\left(H(\hat{P}_B)-\mathbb{E}H(\hat{P}_B)\right)^2.\tag{10}$$

Hence, it suffices to analyze the bias and variance term, respectively, for the non-asymptotic analysis of L_2 risk. The literature on concentration inequalities [17] provide us with excellent techniques in controlling the variance non-asymptotically. However, the focus of this paper is on bias analysis, rather than variance. It may read surprising: how hard can the bias analysis be? It turns out that there is a subfield of mathematics with more than a century's history, called *approximation theory using positive linear operators*, that is largely devoted to analyzing the bias of statistical estimators.

An operator L defined on a linear space of functions, V, is called *linear* if

$$L(\alpha f+\beta g)=\alpha L(f)+\beta L(g),\quad\forall f,g\in V,\alpha,\beta\in\mathbb{R},\tag{11}$$

and is called *positive*, if

$$L(f)\geq 0\quad\text{for all } f\in V, f\geq 0.\tag{12}$$

For example, the classical Bernstein operator $B_n(f)$ maps a continuous function $f\in C[0,1]$ to another continuous function $B_n(f)\in C[0,1]$ such that

$$B_n(f)(x)=\sum_{j=0}^{n}f\left(\frac{j}{n}\right)\binom{n}{j}x^j(1-x)^{n-j}.\tag{13}$$

One can easily verify that this operator is positive and linear. More generally, as argued in Jiao et al. [3], for any estimator $\hat{\theta}$ of a parametric model indexed by θ, the expectation $\varphi\mapsto\mathbb{E}_\theta\varphi(\hat{\theta})$ is a positive linear operator for φ, and analyzing the bias $\mathbb{E}_\theta\varphi(\hat{\theta})-\varphi(\theta)$ is equivalent to analyzing the approximation properties of the positive linear operator $\mathbb{E}_\theta\varphi(\hat{\theta})$ in approximating $\varphi(\theta)$.

Hence, we conclude that the most general problems of bias analysis about plug-in methods in functional estimation are a subset of the general theory of approximation using positive linear operators. Surprising as it may sound, the *converse* is also true. Firstly, it is obvious that the study of positive linear operators in this context can be reduced to that of positive linear functionals. Then it was shown in Paltanea [14, Remark 1.1.2.] that if I is an interval of $\mathbb{R}$, then for any positive linear functional $F:C(I)\to\mathbb{R}$, there exists a positive regular Borel measure μ, such that we have the integral representation[1]:

$$F(f)=\int_I f d\mu,\quad f\in C(I),\tag{14}$$

where $C(I)$ denotes the space of continuous functions on I.

Denote by $e_j, j\in\mathbb{N}_+\cup\{0\}$, the monomial functions $e_j(x)=x^j, x\in I$. If we naturally require $F(e_0)=1$, then the measure μ in (14) is a probability measure, implying that $F(f)$ can be written as an expectation

$$F(f)=\mathbb{E}_\mu f(Z),\quad Z\sim\mu, Z\in I.\tag{15}$$

Thus, there is an equivalence between bias analysis of continuous plug-in estimators and approximation theory using positive linear operators, with extensive literature left for us to explore. Paltanea [14] provides a comprehensive account of the state-of-the-art theory in this subject. We remark that the current theory is "highly developed for certain problems, but far from complete": we now have highly non-trivial tools for positive linear operators of functions on one dimensional compact sets, but the theory is incomplete for vector valued multivariate functions on non-compact sets [14].

III. Main Results

For simplicity, we restrict attention to the case where the parameter $\boldsymbol{\alpha}$ in the Dirichlet distribution takes the form $(a,a,\ldots,a)$. We remark that for general $\boldsymbol{\alpha}$ the analysis goes through smoothly. In comparison to MLE $H(P_n)$, where P_n is the empirical distribution, the Dirichlet smoothing scheme $H(\hat{P}_B)$ has a disadvantage: it requires the knowledge of the alphabet size S in general. We define $\hat{p}_{B,i}=\frac{n\hat{p}_i+a}{n+Sa}$, and $p_{B,i}=\mathbb{E}[\hat{p}_{B,i}]=\frac{np_i+a}{n+Sa}$.

Throughout we adopt the following notations: $a_n\lesssim b_n$ means $\sup_n a_n/b_n<\infty$, $a_n\gtrsim b_n$ means $b_n\lesssim a_n$, $a_n\asymp b_n$ means $a_n\lesssim b_n$ and $a_n\gtrsim b_n$, or equivalently, there exist two universal constants c,C such that

$$0<c<\liminf_{n\to\infty}\frac{a_n}{b_n}\leq\limsup_{n\to\infty}\frac{a_n}{b_n}<C<\infty.\tag{16}$$

The main results of this paper are the following theorems.

Theorem 1. *If $n\geq Sa$, then the maximum L_2 risk of $H(\hat{P}_B)$ in estimating $H(P)$ is upper bounded as*

$$\begin{aligned}&\sup_{P\in\mathcal{M}_S}\mathbb{E}_P\left(H(\hat{P}_B)-H(P)\right)^2\leq\\&\left(\ln\left(1+\frac{S-1}{n+Sa}\right)+\frac{2Sa}{n+Sa}\ln\left(\frac{n+Sa}{2a}\right)\right)^2\\&+\frac{2n}{(n+Sa)^2}\left[3+\ln\left(\frac{n+Sa}{a+1}\wedge S\right)\right]^2,\end{aligned}\tag{17}$$

where $a\wedge b=\min\{a,b\}$. Here the first term bounds the squared bias, and the second term bounds the variance.

The following corollary is immediate.

[1]We remark that not every positive linear functional can be formulated in the form (14). See [14, Remark 1.1.3.].

Corollary 1. *If $n \gg S$ and a is upper bounded by a constant, then the maximum L_2 risk of $H(\hat{P}_B)$ vanishes.*

Theorem 2. *The maximum L_2 risk of $H(\hat{P}_B)$ in estimating $H(P)$ is lower bounded as*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P\left(H(\hat{P}_B) - H(P)\right)^2 \geq$$
$$\begin{cases} \frac{1}{2}\left[\frac{(S-3)a}{4(n+Sa)}\ln\left(\frac{n+Sa}{a}\right) + \frac{S-1}{8n} + \frac{S^2}{80n^2} - \frac{1}{48n^2}\right]^2 + c\frac{\ln^2 S}{n} \\ \qquad\qquad \text{if } n \geq \max\{15S, Sa\}, \\ \left(\frac{(S-3)a}{4(n+Sa)}\ln\left(\frac{n+Sa}{a}\right) + \frac{\lfloor n/15\rfloor}{8n} - \frac{1}{16n}\right)_+^2 \quad \text{if } n < 15S, \\ \frac{\ln^2 S}{16} \qquad\qquad \text{if } n < Sa. \end{cases}$$

where $c > 0$ is a universal constant, and $\lfloor x\rfloor$ is the largest integer that does not exceed x, and $(x)_+ = \max\{x, 0\}$ represents the positive part of x.

We have the following corollary.

Corollary 2. *If $n \lesssim S$ or $n < Sa$, then the maximum L_2 risk of $H(\hat{P}_B)$ is bounded away from zero.*

The next theorem presents a lower bound on the maximum risk of the Bayes estimator under Dirichlet prior. Since we have assumed that all $\alpha_i = a, 1 \leq i \leq S$, the Bayes estimator under Dirichlet prior is

$$\hat{H}^{\mathsf{Bayes}} = \psi(Sa + n + 1) - \sum_{i=1}^{S} \frac{a + X_i}{Sa + n}\psi(a + X_i + 1). \quad (18)$$

Theorem 3. *If $S \geq e(2n + 1)$ and $n \geq Sa$, then*

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P\left(\hat{H}^{\mathsf{Bayes}} - H(P)\right)^2 \geq \left(\ln\left(\frac{S}{e(2n+1)}\right)\right)^2. \quad (19)$$

If $n < Sa$, then

$$\sup_{P\in\mathcal{M}_S} \mathbb{E}_P\left(\hat{H}^{\mathsf{Bayes}} - H(P)\right)^2 \geq \left[\left(\ln\left(\frac{Sa+n}{e(a+n+1)}\right)\right)_+\right]^2. \quad (20)$$

Evident from Theorem 1, 2, and 3 is the fact that in the best situation (i.e. a not too large), both the Dirichlet prior smoothed plug-in estimator and the Bayes estimator under Dirichlet prior still require at least $n \gg S$ samples to be consistent, which is the same as MLE. In contrast, the minimax rate-optimal estimator in Jiao et al. [1] is consistent if $n \gg \frac{S}{\ln S}$, which is the best possible rate for consistency. Thus, we can conclude that the Dirichlet smoothing technique does *not* solve the entropy estimation problem. From an intuitive point of view it is also clear: both the Dirichlet prior smoothed plug-in estimator and the Bayes estimator under Dirichlet prior do not exploit the special properties of the entropy functional $p\ln(1/p)$, i.e. the functional has a nondifferentiable point at $p = 0$. The analysis in [1] demonstrates that the minimax rate-optimal estimator has to exploit the special structure of the entropy function.

IV. Approximation theory for bias analysis

For a linear positive functional F, we adopt the following notation

$$B_F(x) = |F(e_1) - xF(e_0)|, \quad V_F = F\left((e_1 - F(e_1)e_0)^2\right),$$

which represent the "bias" and "variance" of a positive linear functional F. Define the first order and second order Ditzian–Totik modulus of smoothness [20] by

$$\omega_1^{\varphi}(f, 2h) \triangleq \sup\{|f(u) - f(v)| : u, v \in [0,1], |u - v| \leq 2h\varphi\left(\frac{u+v}{2}\right)\}$$
$$\omega_2^{\varphi}(f, h) \triangleq \sup\left\{\left|f(u) - 2f\left(\frac{u+v}{2}\right) + f(v)\right| : u, v \in [0,1], |u - v| \leq 2h\varphi\left(\frac{u+v}{2}\right)\right\}.$$

A. A contribution to approximation theory

First we recall the following result, which is a direct corollary of [14, Thm. 2.5.1].

Lemma 1. *If $F : C[0,1] \to \mathbb{R}$ is a linear positive functional and $F(e_0) = 1$. Then we have*

$$|F(f) - f(x)| \leq \frac{B_F(x)}{2h_1\varphi(x)} \cdot \omega_1^{\varphi}(f, 2h_1) + \frac{5}{2}\omega_2^{\varphi}(f, h_1), \quad (21)$$

for all $f \in C[0,1]$ and $h_1 \in (0, \frac{1}{2}]$, where $\varphi(x) = \sqrt{x(1-x)}$ and $h_1 = \sqrt{F\left((e_1 - xe_0)^2\right)}/\varphi(x) = \sqrt{V_F + (B_F(x))^2}/\varphi(x)$.

We remark that Lemma 1 cannot yield the desired result for $f(p) = -p\ln p$ and

$$F(f) = \sum_{k=0}^{n} f\left(\frac{k+a}{n+Sa}\right) \cdot \binom{n}{k} p^k (1-p)^{n-k}. \quad (22)$$

Specifically, it is easy to show that

$$B_F(p) = \left|\frac{np + a}{n + Sa} - p\right| = \frac{|1 - pS|a}{n + Sa}, \quad V_F = \frac{np(1-p)}{(n+Sa)^2},$$

and for $f(x) = -x\ln x$, $\omega_1^{\varphi}(f, 2h) \asymp h$, and

$$\omega_2^{\varphi}(f, h) = \frac{h^2 \ln 4}{1 + h^2}, \quad h \leq 1. \quad (23)$$

Hence, when $x \to 0$, we conclude that

$$h_1 = \frac{\sqrt{V_F + (B_F(x))^2}}{\varphi(x)} \geq \frac{a}{n + Sa} \cdot \frac{|1 - xS|}{\sqrt{x(1-x)}} \quad (24)$$

is unbounded as $x \to 0$, thus does not satisfy the condition $h_1 \leq 1/2$. Thus, we cannot directly use the result of Lemma 1.

It turns out that the general result in Lemma 1 can be strictly improved in a general fashion. The result is given by the following lemma.

Lemma 2. *If $F : C[0,1] \to \mathbb{R}$ is a linear positive functional and $F(e_0) = 1$. Then*

$$|F(f) - f(x)| \le \omega_1(f, B_F(x); x) + \frac{5}{2}\omega_2^{\varphi}(f, h_2) \qquad (25)$$

for all $f \in C[0,1]$ and $0 < h_2 \le \frac{1}{2}$, where $\varphi(x) = \sqrt{x(1-x)}$ and $h_2 = \sqrt{V_F}/\varphi(x)$, and

$$\omega_1(f, h; x) \triangleq \sup\{|f(u) - f(x)| : u \in [0,1], |u - x| \le h\}.$$

Proof. Applying Lemma 1 to $x = F(e_1)$ we have

$$|F(f) - f(F(e_1))| \le \frac{5}{2}\omega_2^{\varphi}(f, h_2) \qquad (26)$$

and then (25) is the direct result of the triangle inequality $|F(f) - f(x)| \le |F(f) - f(F(e_1))| + |f(F(e_1)) - f(x)|$. □

We show that Lemma 2 is indeed stronger than Lemma 1. First, due to $h_1 \ge h_2$, we have $\omega_2^{\varphi}(f, h_2) \le \omega_2^{\varphi}(f, h_1)$. Second, for $x \le 1/2$, we have

$$\begin{aligned}\frac{B_F(x)}{2h_1\varphi(x)} \cdot \omega_1^{\varphi}(f, 2h_1) &\approx \frac{B_F(x)}{2h_1\varphi(x)} \cdot \sup_{0 \le s \le 1} 2h_1\varphi(s)f'(s) \\ &\ge B_F(x) \cdot \sup_{x \le s \le 1-x} f'(s) \\ &\approx \sup_{x \le s \le 1-x} \omega_1(f, B_F(x); s)\end{aligned}$$

which is no less than the pointwise result $\omega_1(f, |F(e_1 - xe_0)|; x)$, and here we have used the inequality $\varphi(s) \ge \varphi(x)$ for $x \le s \le 1-x$. A similar argument also holds for $x > 1/2$. Hence, Lemma 2 transforms the first order term from the norm result in Lemma 1 to a pointwise result, which may exhibit great advantages when we applied it to specific problems.

B. Application of the improved general bound to our problem

Theorem 4. *If $n \ge \max\{Sa, 4\}$,*

$$\begin{aligned}&\sup_{P \in \mathcal{M}_S} \mathbb{E}_P|H(\hat{P}_B) - H(P)| \\ &\qquad \le \frac{5nS\ln 2}{(n+Sa)^2} + \frac{2Sa}{n+Sa}\ln\left(\frac{n+Sa}{2a}\right).\end{aligned} \qquad (27)$$

Note that Theorem 4 implies a slightly weaker bias bound than Theorem 1, but it is only sub-optimal up to a constant. The bias bound in Theorem 1 is obtained using another technique which is tailored for the entropy function, for which we refer the readers to the journal version [15] for details.

Now we give the proof of Theorem 4 using the approximation theoretic machinery we just established. Note that $h_2 = \frac{\sqrt{n}}{n+Sa}$. where $n \ge 4$ ensures $h_2 \le 1/2$. In light of Lemma 2, we have

$$\begin{aligned}&\mathbb{E}_P|H(\hat{P}_B) - H(P)| \\ &\le \sum_{i=1}^{S}\left(\omega_1\left(f, \frac{|1 - p_iS|a}{n+Sa}; p_i\right) + \frac{5n\ln 2}{(n+Sa)^2}\right) \\ &\le -\left(\sum_{i=1}^{S}\frac{|1-p_iS|a}{n+Sa}\right)\ln\left(\frac{1}{S}\sum_{i=1}^{S}\frac{|1-p_iS|a}{n+Sa}\right) + \frac{5nS\ln 2}{(n+Sa)^2} \\ &\le \frac{2Sa}{n+Sa}\ln\left(\frac{n+Sa}{2a}\right) + \frac{5nS\ln 2}{(n+Sa)^2}\end{aligned}$$

where we have used the fact that if $|x-y| \le 1/2, x, y \in [0,1]$, then $|x\ln x - y\ln y| \le -|x-y|\ln|x-y|$. The readers are referred to the proof of Cover and Thomas [21, Thm. 17.3.3] for details. We also utilized the fact that if $n \ge Sa$, then

$$\frac{|1 - p_iS|a}{n+Sa} \le \frac{Sa}{n+Sa} \le \frac{1}{2}, \quad 1 \le i \le S. \qquad (28)$$

REFERENCES

[1] 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.

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

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

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

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

[6] A. J. Wyner and D. Foster, "On the lower limits of entropy estimation," *IEEE Transactions on Information Theory, submitted for publication*, 2003.

[7] 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

[8] 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.

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

[10] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 6, no. 3, pp. 414–427, 1996.

[11] S. Schober, "Some worst-case bounds for bayesian estimators of discrete distributions," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2194–2198.

[12] D. H. Wolpert and D. R. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Physical Review E*, vol. 52, no. 6, p. 6841, 1995.

[13] D. Holste, I. Grosse, and H. Herzel, "Bayes' estimators of generalized entropies," *Journal of Physics A: Mathematical and General*, vol. 31, no. 11, p. 2551, 1998.

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

[15] Y. Han, J. Jiao, and T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" *submitted to IEEE Transactions on Information Theory*.

[16] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998, vol. 31.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

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

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.