Maximum Likelihood Estimation of Functionals of Discrete Distributions

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

Abstract—We consider the problem of estimating functionals of discrete distributions, and focus on tight (up to universal multiplicative constants for each specific functional) nonasymptotic analysis of the worst case squared error risk of widely used estimators. We apply concentration inequalities to analyze the random fluctuation of these estimators around their expectations, and the theory of approximation using positive linear operators to analyze the deviation of their expectations from the true functional, namely their *bias*.

We explicitly characterize the worst case squared error risk incurred by the Maximum Likelihood Estimator (MLE) in estimating the Shannon entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$, and the power sum $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \alpha > 0$, up to universal multiplicative constants for each fixed functional, for any alphabet size $S \leq \infty$ and sample size n for which the risk may vanish. 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. In addition, we establish 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 has a strictly sub-optimal sample complexity. When $1 < \alpha < 3/2$, we show that the worst-case squared error rate of convergence for the MLE is $n^{-2(\alpha-1)}$ for infinite alphabet size, while the minimax squared error rate is $(n \ln n)^{-2(\alpha-1)}$. When $\alpha \geq 3/2$, the MLE achieves the minimax optimal rate n^{-1} regardless of the alphabet size.

As an application of the general theory, we analyze the Dirichlet prior smoothing techniques for Shannon entropy estimation. In this context, 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 such estimators do *not* improve over the maximum likelihood estimator. No matter how we tune the parameters in the Dirichlet prior, this approach cannot achieve the minimax rates in entropy estimation. The performance of the minimax rate-optimal estimator with n samples is essentially *at least* as good as that of Dirichlet smoothed entropy estimators with $n \ln n$ samples.

I. INTRODUCTION

Entropy and related information measures arise in information theory, statistics, machine learning, biology, neuroscience, image processing, linguistics, secrecy, ecology, physics, and finance, among other fields. Numerous inferential tasks rely on data driven procedures to estimate these quantities (see, e.g. [1]–[6]). We focus on two concrete and well-motivated examples of information measures, namely the Shannon entropy [7]

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

and the power sum $F_\alpha(P), \alpha > 0$:

$$F_\alpha(P) \triangleq \sum_{i=1}^{S} p_i^\alpha, \alpha > 0. \tag{2}$$

The power sum $F_\alpha(P)$ functional often emerges in various operational problems [8]. It also has connections to the Rényi entropy [9] $H_\alpha(P)$ via the formula $H_\alpha(P) = \frac{\ln F_\alpha(P)}{1-\alpha}$.

Consider estimating the Shannon entropy $H(P)$ based on n i.i.d. samples following unknown discrete distribution P with *unknown* alphabet size S. This problem has a rich history with extensive study in various fields ranging from information theory, statistics, neuroscience, physics, psychology, medicine, etc. We refer the reader to [10] 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. The empirical entropy is an instantiation of the plug-in principle in functional estimation, where a point estimate of the parameter (distribution P in this case) is used to construct an estimator for a *functional* of the parameter via the plug-in approach. The idea of using the MLE for estimating information measures of interest (in this case entropy), is not only intuitive, but has sound justification: *asymptotic efficiency*.

The beautiful theory of Hájek and Le Cam [11]–[13] shows that, as the number of observed samples grows without bound while the finite parameter dimension (e.g., alphabet size) remains fixed, the MLE performs optimally in estimating any differentiable functional when the statistical model complies with the benign LAN (Local Asymptotic Normality) condition [13]. 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 not only natural but also theoretically justified. But does it make sense to employ the MLE to estimate the entropy in most practical applications?

As it turns out, while asymptotically optimal in entropy estimation, the MLE is by no means sacrosanct in many real applications, especially in regimes where the alphabet 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 alphabet regime [14], [15]. Therefore, classical asymptotic theory does not satisfactorily address high dimensional settings, which are becoming increasingly important in the modern era of high dimensional statistics.

There has been a wave of recent research activities focusing on analyzing existing approaches of functional estimation, as well as proposing new estimators that are provably near optimal in the large alphabet regime. Paninski [14] showed that the MLE needs $n \gg S$ samples to consistently estimate the Shannon entropy, and Paninski [15] established the existence of a (non-explicit) estimator that only required $n \ll S$ samples. It implies that the MLE is strictly sub-optimal in terms of sample complexity. It was Valiant and Valiant [16] who first explicitly constructed a linear programming based estimator (later modified in [17]) that achieves consistency in entropy estimation with $n \gg S/\ln S$ samples, which they also proved to be necessary. Valiant and Valiant [18] constructed another approximation based estimator that achieved better theoretical properties than the linear programming ones, which was not yet shown to be minimax rate-optimal for all ranges of S and n. The authors [10] constructed the first minimax rate-optimal estimators for $H(P)$ and $F_\alpha(P), \alpha > 0$ based on best polynomial approximation, which are agnostic to the alphabet size S. Utilizing the released MATLAB and Python packages of the estimators in [10], [19], [20] demonstrated that these minimax rate-optimal estimators can lead to significant performance boosts in various machine learning tasks. Wu and Yang [21] independently applied the best polynomial approximation idea to entropy estimation and obtained the minimax rates. However, their estimator requires the knowledge of the alphabet size S. The approximation ideas proved to be very fruitful in Acharya *et al.* [22], Wu and Yang [23], Han, Jiao, and Weissman [24], Jiao, Han, and Weissman [25], Bu *et al.* [26], Orlitsky, Suresh, and Wu [27], Wu and Yang [28].

The main contribution of this paper is an explicit characterization of the worst case squared error risk of estimating $H(P)$ and $F_\alpha(P)$ using the MLE up to a universal multiplicative constant for each specific functional, for all ranges of S and n in which the risk may vanish. Understanding the benefits and limitations of the MLE in a nonasymptotic setting 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, which have been validated in [10], [14]–[18], [21], [22]. As a byproduct of the analysis, we explicitly point out an equivalence between bias analysis of functional estimators using plug-in rules and approximation

theory using positive linear operators. We believe these powerful tools introduced from approximation theory may have far reaching impacts in various applications in the information theory community.

We mention that there exist numerous other approaches proposed in various disciplines to estimate entropy, many among which are difficult to analyze theoretically. Among them we mention the Miller–Madow bias-corrected estimator and its variants [29]–[31], the jackknife estimator [32], the shrinkage estimator [33], the coverage adjusted estimator [34], the Best Upper Bound (BUB) estimator [14], the B-Splines estimator [35], and [36], [37] etc. For a Bayesian statistician, a natural approach is to first impose a prior on the unknown discrete distribution before considering estimating entropy. The Dirichlet prior, being the conjugate prior to the multinomial distribution, appears to be particularly popular in the Bayesian approach to entropy estimation. Dirichlet smoothing may have two connotations in the context of entropy estimation:

- [38], [39] 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 $H(\hat{P}_B)$.
- [40] [41] 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.

Nemenman, Shafee, and Bialek [42] argued in an intuitive way why Dirichlet prior is bad for entropy estimation and proposed to use mixtures of Dirichlet priors. Archer, Park, and Pillow [43] have come up with priors that perform better than the Dirichlet prior. Also see [44], [45].

Another contribution of this paper is an explicit characterization of the worst case squared error risk of estimating $H(P)$ using the Dirichlet prior plug-in approach up to a universal multiplicative constant, for all ranges of S and n in which the risk may vanish. We show rigorously that neither of the two approaches utilizing the Dirichlet prior result 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 [10] [21] only need $n \gg \frac{S}{\ln S}$ to achieve consistency.

The rest of the paper is organized as follows. We present the main results in Section III, discuss the fundamental ideas behind the proofs in Section IV, and detail the proofs in Section V and VI. Proofs of auxiliary lemmas are deferred to the appendices.

II. PRELIMINARIES

The Dirichlet distribution with order $S \geq 2$ with parameters $\alpha_1, \ldots, \alpha_S > 0$ has a probability density function with respect to Lebesgue measure on the Euclidean space $\mathbb{R}^{S-1}$ given by

$$f(x_1, \cdots, x_S; \alpha_1, \cdots, \alpha_S) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{S} x_i^{\alpha_i - 1} \quad (3)$$

on the open $S-1$-dimensional simplex defined by:

$$x_1, \cdots, x_{S-1} > 0 \qquad (4)$$
$$x_1 + \cdots + x_{S-1} < 1 \qquad (5)$$
$$x_S = 1 - x_1 - \cdots - x_{S-1} \qquad (6)$$

and zero elsewhere. The normalizing constant is the multinomial Beta function, which can be expressed in terms of the Gamma function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{S} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{S} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_S). \qquad (7)$$

Assuming the unknown discrete 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 $\mathrm{multi}(n; p_1, p_2, \ldots, p_S)$, then one can show that the posterior distribution $P_{P|\mathbf{X}}$ is also a Dirichlet distribution with parameters

$$\boldsymbol{\alpha} + \mathbf{X} = (\alpha_1 + X_1, \alpha_2 + X_2, \ldots, \alpha_S + X_S). \qquad (8)$$

Furthermore, the posterior mean (conditional expectation) of p_i given $\mathbf{X}$ is given by [46, Example 5.4.4]

$$\delta_i(\mathbf{X}) \triangleq \mathbb{E}[p_i | \mathbf{X}] = \frac{\alpha_i + X_i}{n + \sum_{i=1}^{S} \alpha_i}. \qquad (9)$$

The estimator $\delta_i(\mathbf{X})$ is widely used in practice for various choices of α. For example, if $\alpha_i = \frac{\sqrt{n}}{S}$, 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 [46, Example 5.4.5]. However, it is no longer minimax under other loss functions such as ℓ_1 loss, which was investigated in [47].

Note that the estimator $\delta_i(\mathbf{X})$ subsumes the MLE $\hat{p}_i = \frac{X_i}{n}$ as a special case, since we can take the limit $\boldsymbol{\alpha} \to 0$ for $\delta_i(\mathbf{X})$ to obtain MLE. We denote the empirical distribution by $P_n = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_S)$. The Dirichlet prior smoothed distribution estimate is denoted as $\hat{P}_B$, where

$$\hat{P}_B = \frac{n}{n + \sum_{i=1}^{S} \alpha_i} P_n + \frac{\sum_{i=1}^{S} \alpha_i}{n + \sum_{i=1}^{S} \alpha_i} \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{S} \alpha_i}. \qquad (10)$$

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 [40] that

$$\hat{H}^{\mathsf{Bayes}} \triangleq \mathbb{E}[H(P) | \mathbf{X}]$$
$$= \psi\left(1 + \sum_{i=1}^{S} (\alpha_i + X_i)\right)$$
$$- \sum_{i=1}^{S} \left(\frac{\alpha_i + X_i}{\sum_{i=1}^{S}(\alpha_i + X_i)}\right) \psi(\alpha_i + X_i + 1), \qquad (11)$$

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.*

Throughout this paper, we observe n i.i.d. samples from an unknown discrete distribution $P = (p_1, p_2, \ldots, p_S)$. We denote the n samples as n i.i.d. random variables $\{Z_i\}_{1 \le i \le n}$ taking values in $\mathcal{Z} = \{1, 2, \ldots, S\}$ with probability $(p_1, p_2, \ldots, p_S)$. Defining

$$X_i \triangleq \sum_{j=1}^{n} \mathbb{1}(Z_j = i), \quad 1 \le i \le S, \qquad (12)$$

we know that $(X_1, X_2, \ldots, X_S)$ follows a multinomial distribution with parameter $(n; p_1, p_2, \ldots, p_S)$. Denote $h_j \triangleq \sum_{i=1}^{S} \mathbb{1}(X_i = j)$, $0 \le j \le n$. The Maximum Likelihood Estimator (MLE) for $H(P)$ and $F_\alpha(P)$ are defined, respectively, as $H(P_n)$ and $F_\alpha(P_n)$, with P_n being the empirical distribution. We assume the functional $F(P)$ takes the form

$$F(P) = \sum_{i=1}^{S} f(p_i). \qquad (13)$$

Then it is evident that the MLE $F(P_n)$ for estimating functional $F(P)$ in (13) can be alternatively represented as the following linear function of $(h_0, h_1, \ldots, h_n)$:

$$F(P_n) = \sum_{j=0}^{n} f\left(\frac{j}{n}\right) h_j. \qquad (14)$$

Recall that the risk function under squared error for any estimator $\hat{F}$ in estimating functional $F(P)$ may be decomposed as

$$\mathbb{E}_P(F(P) - \hat{F})^2 = (\mathbb{E}_P \hat{F} - F(P))^2 + \mathbb{E}_P\left(\hat{F} - \mathbb{E}_P \hat{F}\right)^2, \qquad (15)$$

where $(\mathbb{E}_P \hat{F} - F(P))^2$ represents the squared bias, and $\mathbb{E}_P\left(\hat{F} - \mathbb{E}_P \hat{F}\right)^2$ represents the variance. The subscript P means that the expectation is taken with respect to the distribution P that generates the i.i.d. observations. We omit the subscript for the expectation operator $\mathbb{E}$ if the meaning of the expectation is clear from the context.

Notation: $a \wedge b$ denotes $\min\{a, b\}$, $a \vee b$ denotes $\max\{a, b\}$. For two non-negative series $\{a_n\}, \{b_n\}$, notation $a_n \lesssim b_n$ means that there exists a positive universal constant $C < \infty$ such that $\frac{a_n}{b_n} \le C$, for all n. The 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 $\liminf_{n \to \infty} \frac{a_n}{b_n} = \infty$. Throughout this paper, the notations $\lesssim, \gtrsim, \ll, \gg$ involve absolute constants that may only depend on α but not S or n. We denote by $\mathcal{M}_S$ the space of discrete distributions with alphabet size S.

III. MAIN RESULTS

A. Estimating $F_\alpha(P)$

We split the upper bounds and the lower bounds into two theorems, and present their succinct summaries in Corollary 1 and 2.

Theorem 1 (Upper bounds)**.** *We have the following upper bounds on the worst case squared error risk of MLE in estimating $F_\alpha(P)$:*

1) $\alpha \geq 2$:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2 \leq \left(\frac{\alpha(\alpha-1)}{2n} \right)^2 + \frac{\alpha^2}{4n}. \tag{16}$$

2) $1 < \alpha < 2$:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\leq \left(\frac{4}{n^{\alpha-1}} \wedge \frac{3S^{1-\alpha/2}}{n^{\alpha/2}} \wedge C_{\alpha,n} \frac{5S}{2n} \right)^2 + \frac{\alpha^2}{4n}, \tag{17}$$

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 introduced in Section IV-B.*

3) $1/2 \leq \alpha < 1$:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\leq \left(\frac{3S^{1-\alpha/2}}{2n^{\alpha/2}} \wedge \frac{5S}{2n^\alpha} \right)^2$$
$$+ \left(\frac{10S^{2-2\alpha}}{n} + \frac{120}{\alpha^2} \left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}} \right) \right). \tag{18}$$

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

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\leq \left(\frac{3S^{1-\alpha/2}}{2n^{\alpha/2}} \wedge \frac{5S}{2n^\alpha} \right)^2$$
$$+ \left(\frac{10S}{n^{2\alpha}} + \frac{120}{\alpha^2} \left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}} \right) \right). \tag{19}$$

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

Theorem 2 (Lower bounds). *We have the following lower bounds on the worst case squared error risk of MLE in estimating* $F_\alpha(P)$:

1) $\alpha \geq 3/2$: *there exists a constant* $C_\alpha > 0$ *such that for all* n,

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2 \geq \frac{C_\alpha}{n}. \tag{20}$$

2) $1 < \alpha < 3/2$: *if* $S = cn$, *for any* $c > 0$, *then*

$$\liminf_{n \to \infty} n^{2(\alpha-1)} \cdot \sup_{P \in \mathcal{M}_S} \mathbb{E}_P (F_\alpha(P_n) - F_\alpha(P))^2 > 0. \tag{21}$$

3) $1/2 \leq \alpha < 1$: *if* $n \geq S$, *then*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\geq \frac{\alpha^2(1-\alpha)^2}{72n^{2\alpha}} (S-1)^2 \left(1 - \frac{1}{n} \right)^2$$
$$+ \left(\frac{\alpha^2}{64en} \left[(2(S-1))^{1-\alpha} - 2^{-\alpha} \right. \right.$$
$$\left. - \frac{1-\alpha}{4n} \left((2(S-1))^{1-\alpha} + 2^{-\alpha} \right) \right]^2$$
$$\left. - \frac{1}{2} e^{-n/4} S^{2(1-\alpha)} \right), \tag{22}$$

4) $0 < \alpha < 1/2$: *if* $n \geq S$, *then*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\geq \frac{\alpha^2(1-\alpha)^2}{36n^{2\alpha}} (S-1)^2 \left(1 - \frac{1}{n} \right)^2. \tag{23}$$

There are several interesting implications of this result, highlighted in the following corollaries.

Corollary 1. *For any fixed* $\alpha > 1$, *there exist universal convergence rates for* $F_\alpha(P)$:

$$\sup_{S \in \mathbb{N}_+} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$
$$\asymp \begin{cases} n^{-2(\alpha-1)} & 1 < \alpha < 3/2 \\ n^{-1} & \alpha \geq 3/2 \end{cases} \tag{24}$$

Corollary 1 implies that, when $\alpha \geq 3/2$, estimation of $F_\alpha(P)$ is extremely simple in terms of convergence rate: plug-in estimation achieves the best possible rate n^{-1} (as shown in the theory of regular statistical experiments of classical asymptotic theory, see [48, Chap. 1.7.]). Results of this form have appeared in the literature, for example, Antos and Kontoyiannis [49] showed that it suffices to take $n \gg 1$ samples to consistently estimate $F_\alpha(P), \alpha \geq 2, \alpha \in \mathbb{Z}$. However, when $1 < \alpha < 3/2$, the rate $n^{-2(\alpha-1)}$ is considerably slower. Interestingly, there exist estimators that demonstrate better convergence rates for estimating $F_\alpha(P), 1 < \alpha < 3/2$. Jiao *et al.* [10] showed that the minimax rate in estimating $F_\alpha(P), 1 < \alpha < 3/2$, is $(n \ln n)^{-2(\alpha-1)}$ as long as $S \gtrsim n \ln n$, which is achieved using the general methodology developed therein for constructing minimax rate-optimal estimators for nonsmooth functionals.

Let us now examine the case $0 < \alpha < 1$, another interesting regime that has not been characterized before. In this regime, we observe significant increase in the difficulty of the estimation problem. In particular, the relative scaling between the number of observations n and the alphabet size S for consistent estimation of $F_\alpha(P)$ exhibits a phase transition, encapsulated in the following.

Corollary 2. *Fix* $\alpha \in (0, 1)$. *The worst case squared error risk of the MLE* $F_\alpha(P_n)$ *in estimating* $F_\alpha(P)$ *is characterized*

as follows when $n \geq S$:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$

$$\asymp \begin{cases} \frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n} & 1/2 < \alpha < 1 \\ \frac{S^2}{n^{2\alpha}} & 0 < \alpha \leq 1/2 \end{cases} \tag{25}$$

Corollary 2 follows directly from Theorem 1 and Theorem 2. 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, which is consistent with the intuition that $F_\alpha(P), \alpha \to 0$ is essentially equivalent to the alphabet size of a distribution, whose estimation is known to be very hard when there may exist symbols with very small probabilities [50].

We exhibit some of our findings by plotting the value required of $\ln n / \ln S$ for consistent estimation of $F_\alpha(P)$ using the MLE $F_\alpha(P_n)$, as a function of α, in Figure 1.
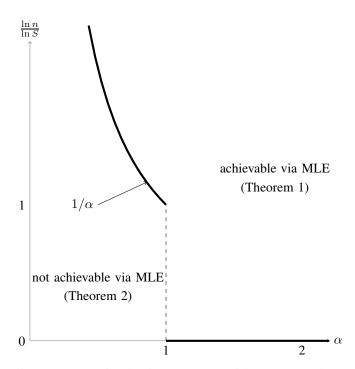

Fig. 1: For any fixed point above the thick curve, consistent estimation of $F_\alpha(P)$ is achieved using MLE $F_\alpha(P_n)$ as shown in Theorem 1. For any fixed point below the thick curve in the regime $0 < \alpha < 1$, Theorem 2 shows that the MLE does not have vanishing maximum squared error risk.

It turns out that one can construct estimators that are better than the MLE in terms of required sample complexity for consistent estimation for the regime $0 < \alpha < 1$. Indeed, Jiao *et al.* [10] showed that the minimax rate-optimal estimator requires $n \gg \frac{S^{\frac{1}{\alpha}}}{\ln S}$ samples to achieve consistency, which attains a logarithmic improvement in the sample complexity over the MLE.

B. Estimating $H(P)$

We not only consider $H(P_n)$, but also the so-called Miller–Madow bias-corrected estimator [29] defined as

$$H^{\mathrm{MM}}(P_n) = H(P_n) + \frac{S-1}{2n}. \tag{26}$$

Theorem 3. *The worst case squared error risk of $H(P_n)$ admits the following upper bound for all S, n:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P) \right)^2$$

$$\leq \left(\ln \left(1 + \frac{S-1}{n} \right) \right)^2 + \left(\frac{(\ln n)^2}{n} \wedge \frac{2(\ln S + 3)^2}{n} \right). \tag{27}$$

If $n \geq 15S$, then

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P) \right)^2$$

$$\geq \frac{1}{2} \left(\frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2} \right)^2 + c \frac{\ln^2 S}{n}. \tag{28}$$

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

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H^{\mathrm{MM}}(P_n) - H(P) \right)^2$$

$$\geq \frac{1}{2} \left(\frac{S^2}{20n^2} - \frac{1}{12n^2} \right)^2 + c \frac{\ln^2 S}{n}, \tag{29}$$

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

Theorem 3 implies the following corollary.

Corollary 3. *The worst case squared error risk of the MLE $H(P_n)$ in estimating $H(P)$ is characterized as follows when $n \geq 15S$:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P_n) - H(P) \right)^2 \asymp \frac{S^2}{n^2} + \frac{\ln^2 S}{n}. \tag{30}$$

Here the first term corresponds to the squared bias, and the second term corresponds to the variance.

Paninski [14] showed that if $n = cS$, where $c > 0$ is a constant, the maximum squared error risk of $H(P_n)$, and the Miller–Madow bias-corrected estimator $H^{\mathrm{MM}}(P_n)$, would be bounded from zero. Paninski [14] also showed that when $n \gg S, n \to \infty$, the MLE is consistent for estimating entropy. Corollary 3 implies that it is necessary and sufficient to take $n \gg S$ samples for the MLE to be consistent for estimating entropy. Comparing the results for $H(P)$ with those for $F_\alpha(P)$, we see that the intuition that $H(P)$ being viewed close to $F_\alpha(P)$ when $\alpha \to 1^{-1}$ is indeed approximately correct as $H(P)$ coincides with $\alpha \to 1^-$ on the phase transition curve shown in Figure 1.

Table I summarizes the minimax squared error rates and the worst case squared error rates of the MLE in estimating $H(P)$ and $F_\alpha(P), \alpha > 0$. It is clear that the MLE cannot achieve the minimax rates 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

	Minimax squared error rates	Maximum squared error rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2} + \frac{\ln^2 S}{n}$ $(n \gtrsim S/\ln S)$ ([10], [16], [18], [21])	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$ $(n \gtrsim S)$ (Corollary 3)
$F_\alpha(P), 0 < \alpha \le \frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}}$ $\left(n \gtrsim S^{1/\alpha}/\ln S, \ln n \lesssim \ln S\right)$ ([10])	$\frac{S^2}{n^{2\alpha}}$ $\left(n \gtrsim S^{1/\alpha}\right)$ (Corollary 2)
$F_\alpha(P), \frac{1}{2} < \alpha < 1$	$\frac{S^2}{(n\ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$ $\left(n \gtrsim S^{1/\alpha}/\ln S\right)$ ([10])	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$ $\left(n \gtrsim S^{1/\alpha}\right)$ (Corollary 2)
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)}$ $(S \gtrsim n\ln n)$ ([10])	$n^{-2(\alpha-1)}$ $(S \gtrsim n)$ (Corollary 1)
$F_\alpha(P), \alpha \ge \frac{3}{2}$	n^{-1} (Theorem 1)	n^{-1}

TABLE I: Summary of results in this paper and the companion [10]

that of the MLE with $n \ln n$ samples. This phenomenon was termed *effective sample size enlargement* in [10].

C. Dirichlet prior techniques applying to entropy estimation

For symmetry, we restrict attention to the case where the parameter α in the Dirichlet distribution takes the form $(a, a, \ldots, a)$.

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

$$\hat{p}_{B,i} = \frac{n\hat{p}_i + a}{n + Sa}, \tag{31}$$

and

$$p_{B,i} = \mathbb{E}[\hat{p}_{B,i}] = \frac{np_i + a}{n + Sa}. \tag{32}$$

It is clear that

$$\hat{P}_B = \frac{n}{n + Sa}P_n + \frac{Sa}{n + Sa}U_S \tag{33}$$

$$P_B = \frac{n}{n + Sa}P + \frac{Sa}{n + Sa}U_S, \tag{34}$$

where P_n stands for the empirical distribution, P is the true distribution, and U_S denotes the uniform distribution on the same alphabet with size S.

Theorem 4. *If $n \ge \max\{Sa, 2ea\}$, then the maximum squared error risk of $H(\hat{P}_B)$ in estimating $H(P)$ is upper bounded as*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(\hat{P}_B) - H(P) \right)^2$$

$$\le \left(\ln\left(1 + \frac{S-1}{n + Sa}\right) \vee \frac{2Sa}{n + Sa}\ln\left(\frac{n + Sa}{2a}\right) \right)^2$$

$$+ \frac{2n}{(n + Sa)^2}\left[3 + \ln\left(\frac{n + Sa}{a + 1} \wedge S\right)\right]^2. \tag{35}$$

Here the first term bounds the squared bias, and the second term bounds the variance.

Theorem 5. *If $n \ge \max\{15S, Sa, 2ea\}$, then the maximum*

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

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(\hat{P}_B) - H(P) \right)^2$$

$$\ge \frac{1}{2}\left[\frac{(S-1)a}{4(n + Sa)}\ln\left(\frac{n + Sa}{a}\right) + \frac{S-1}{8n} + \frac{S^2}{80n^2} - \frac{1}{48n^2}\right]^2$$

$$+ c\frac{\ln^2 S}{n}, \tag{36}$$

where $c > 0$ is a universal constant that does not depend on $a, S,$ or n.

If $n < Sa$, then we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(\hat{P}_B) - H(P) \right)^2 \ge \left(\frac{S-1}{2S}\right)^2 \ln^2 S. \tag{37}$$

If $n < 2ea$, then we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(\hat{P}_B) - H(P) \right)^2 \ge \left(\frac{S-1}{S + 2e}\right)^2 \ln^2 S. \tag{38}$$

If $n < 15S, n \ge 2ea$, then we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(\hat{P}_B) - H(P) \right)^2$$

$$\ge \left[\left(\frac{(S-1)a}{4(n + Sa)}\ln\left(\frac{n + Sa}{a}\right) + \frac{\lfloor n/15 \rfloor}{8n} - \frac{1}{16n}\right)_+\right]^2, \tag{39}$$

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

The following corollary immediately follows from Theorem 4 and Theorem 5.

Corollary 4. *If $n \gg S$ and a is upper bounded by a constant, then the maximum squared error risk of $H(\hat{P}_B)$ vanishes. Conversely, if $n \lesssim S$, then the maximum squared error 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 $\alpha_i = a, 1 \le i \le S$, the Bayes estimator under Dirichlet prior is

$$\hat{H}^{\mathsf{Bayes}} = \psi(Sa + n + 1) - \sum_{i=1}^{S} \frac{a + X_i}{Sa + n}\psi(a + X_i + 1). \tag{40}$$

Theorem 6. *If $S \geq 2(n+1)$, then*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H}^{\text{Bayes}} - H(P) \right)^2 \geq \left(\ln \left(\frac{Sa + S/2}{Sa + n + e^{-\gamma}} \right) \right)^2, \tag{41}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant.

Evident from Theorem 4, 5, and 6 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 estimators in Valiant and Valiant [16]–[18], Jiao *et al.* [10], Wu and Yang [21] are consistent if $n \gg \frac{S}{\ln S}$, which is the optimal sample complexity. Thus, we can conclude that the Dirichlet smoothing technique does *not* solve the entropy estimation problem.

IV. FUNDAMENTAL IDEAS OF OUR ANALYSIS

In this section, we discuss the fundamental tools we employed to obtain the results in Section III, as well as general recipes we suggest for analyzing performances of functional estimators.

A. Variance

The variance characterizes the degree to which the random variable $F(\hat{P})$ is fluctuating around its expectation, and the field of concentration inequalities perfectly fits our glove to give the desired results. For all the functionals we consider, it turns out that the Efron–Stein inequality [51] and the bounded differences inequality give very tight bounds. For completeness we state them below.

Lemma 1. *[52, Efron–Stein inequality, Theorem 3.1] Let $Z_1, \ldots, Z_n$ be independent random variables and let $f(Z_1, Z_2, \ldots, Z_n)$ be a square integrable function. Moreover, if $Z'_1, Z'_2, \ldots, Z'_n$ are independent copies of $Z_1, Z_2, \ldots, Z_n$ and if we define, for every $i = 1, 2, \ldots, n$,*

$$f'_i = f(Z_1, Z_2, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n), \tag{42}$$

then

$$\text{Var}(f) \leq \frac{1}{2} \sum_{i=1}^n \mathbb{E} \left[(f - f'_i)^2 \right]. \tag{43}$$

The following inequality, which is called the bounded differences inequality, is a useful corollary of the Efron–Stein inequality.

Lemma 2. *[52, Bounded differences inequality, Corollary 3.2] If function $f : \mathcal{Z}^n \rightarrow \mathbb{R}$ has the* bounded differences property*, i.e., for some nonnegative constants $c_1, c_2, \ldots, c_n$,*

$$\sup_{z_1, \ldots, z_n, z'_i \in \mathcal{Z}} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)|$$

$$\leq c_i, \tag{44}$$

for every $1 \leq i \leq n$, then

$$\text{Var}(f(Z_1, Z_2, \ldots, Z_n)) \leq \frac{1}{4} \sum_{i=1}^n c_i^2, \tag{45}$$

given that $Z_1, Z_2, \ldots, Z_n$ are independent random variables.

We refer the readers to Boucheron *et al.* [52] for a modern exposition of the concentration inequality toolbox.

B. Bias

It turns out that the bias analysis in estimation, albeit widely studied in statistics, seems to still largely bear an asymptotic and expansion nature in the mainstream statistical literature [53], [54]. In particular, the bootstrap [55] as a method for estimating functionals was essentially only analyzed in an asymptotic setting [56]. Among asymptotic analysis techniques, probably the most popular one is the Taylor expansion. We will show that the Taylor expansion may encounter great difficulties in analyzing the bias of MLE in information measure estimation. Then, we will introduce the field of *approximation theory using positive linear operators* and demonstrate that it is essentially equivalent to *nonasymptotic* bias analysis for plug-in functional estimators. In doing so, we present the readers with abundant handy tools from approximation theory, which could be readily applicable to many problems that may seem highly intractable with standard expansion methods.

We start from entropy estimation. In the 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 [29] in 1955, who showed that, for any fixed distribution P,

$$\mathbb{E}H(P_n) = H(P) - \frac{S-1}{2n} + O\left(\frac{1}{n^2}\right). \tag{46}$$

Equation (46) was later refined by Harris [57] using higher order Taylor series expansions to yield

$$\mathbb{E}H(P_n) = H(P) - \frac{S-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^S \frac{1}{p_i}\right) + O\left(\frac{1}{n^3}\right). \tag{47}$$

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 cannot 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, which was first observed in Paninski [14]. To see this, we first compute the bias of $F(P_n)$ in estimating the functional $F(P)$ in (13).

Lemma 3. *The bias of the estimator $F(P_n)$ is given by*

$$\mathsf{Bias}(F(P_n)) \triangleq \mathbb{E}F(P_n) - F(P)$$

$$= \sum_{i=1}^{S} \left(\sum_{j=0}^{n} f\left(\frac{j}{n}\right) \binom{n}{j} p_i^j (1-p_i)^{n-j} - f(p_i) \right) \tag{48}$$

The bias term in (48) can be equivalently expressed as[1]

$$\mathsf{Bias}(F(P_n)) = \sum_{i=1}^{S} \left(\sum_{j=0}^{n} f\left(\frac{j}{n}\right) B_{j,n}(p_i) - f(p_i) \right) \tag{49}$$

$$= \sum_{i=1}^{S} \left(B_n[f](p_i) - f(p_i) \right), \tag{50}$$

where $B_{j,n}(x) \triangleq \binom{n}{j} x^j (1-x)^{n-j}$ is the well-known Bernstein polynomial basis, and $B_n[f](x)$ is the so-called Bernstein polynomial for function $f(x)$. Bernstein in 1912 [61] 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 analytic 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. Evidently, bounding the approximation error incurred by the Bernstein polynomial is equivalent to bounding the bias of the MLE $f(X/n)$, where $X \sim \mathsf{B}(n,x)$. Fortunately, the theory of *approximation using positive linear operators* [62] provides us with advanced tools that are very effective for the bias analysis our problem calls for. 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 better understand probability theory and statistics. We briefly review the general theory of approximation using positive linear operators below.

1) Approximation theory using positive linear operators: Generally speaking, for any estimator $\hat\theta$ of a parametric model indexed by θ, the expectation $f \mapsto \mathbb{E}_\theta f(\hat\theta)$ is a positive linear operator for f, and analyzing the bias $\mathbb{E}_\theta f(\hat\theta) - f(\theta)$ is equivalent to analyzing the approximation properties of the positive linear operator $\mathbb{E}_\theta f(\hat\theta)$ in approximating $f(\theta)$. Hence, analyzing the bias of *any* plug-in estimator for functionals of parameters from *any* parametric families can be recast as a problem of approximation theory using positive linear operators [62].

Conversely, given a positive linear operator $L(f)(x)$ that operates on the space of continuous functions, the Riesz–Markov–Kakutani theorem implies that under mild conditions the operator may be written as

$$L(f)(x) = \int_I f d\mu_x = \mathbb{E}_{\mu_x} f(Z), Z \sim \mu_x, \tag{51}$$

[1]In the literature of combinatorics, the sum $\sum_{j=0}^{n} a_{j,n} B_{j,n}(x)$ is called the Bernoulli sum, and various approaches have been proposed to evaluate its asymptotics [58], [59], [60].

where $\{\mu_x\}$ is a set of probability measures parametrized by x, which may be viewed as a parameter. If we view the random variable Z as a summary statistics to plug-in the functional $f(\cdot)$, the positive linear operator $L(f)(x)$ is nothing but the expectation of the plug-in estimator $f(Z)$. In this sense, there exists a one-to-one correspondence between essentially the most general bias analysis problem in statistics, and the most general positive linear operator approximation problem in approximation theory.

After more than a century's active research on approximation using positive linear operators, we now have highly non-trivial tools for positive linear operators of functions on one dimensional compact sets, but the general theory for vector valued multivariate functions on non-compact sets is still far from complete [62]. In the next subsection, we present a sample of existing results in approximation using positive linear operators, corollaries of which will be used to analyze the bias of the MLE for two examples: $F_\alpha(P)$ and $H(P)$.

2) Some general results in bias analysis: First, some elementary approximation theoretic concepts need to be introduced in order to characterize the degree of *smoothness* of functions. For $I \subset \mathbb{R}$ an interval, the first-order modulus of smoothness $\omega^1(f,t), t \geq 0$ is defined as [62]

$$\omega^1(f,t) \triangleq \sup\{|f(u) - f(v)| : u, v \in I, |u-v| \leq t\}. \tag{52}$$

The second-order modulus of smoothness $\omega^2(f,t), t \geq 0$ [62] is defined as

$$\omega^2(f,t) \triangleq \sup\left\{ \left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right| : \right.$$
$$\left. u, v \in I, |u-v| \leq 2t \right\}. \tag{53}$$

Ditzian and Totik [63] introduced a class of moduli 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 first-order Ditzian–Totik modulus of smoothness is defined as

$$\omega_\varphi^1(f,t) \triangleq \sup\left\{ |f(u) - f(v)| : \right.$$
$$\left. u, v \in [0,1], |u-v| \leq t\varphi\left(\frac{u+v}{2}\right) \right\}, \tag{54}$$

and the second-order Ditzian–Totik modulus of smoothness is defined as

$$\omega_\varphi^2(f,t) \triangleq \sup\left\{ \left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right| : \right.$$
$$\left. u, v \in [0,1], |u-v| \leq 2t\varphi\left(\frac{u+v}{2}\right) \right\}. \tag{55}$$

Recall that we denote by $e_j, j \in \mathbb{N}_+ \cup \{0\}$, the monomial functions $e_j(y) = y^j, y \in I$. The first estimate for general positive linear operators, using modulus ω^2 and with precise constants, was given by Gonska [64]. We rephrase Paltanea [62,

Cor. 2.2.1.] as follows. Note that notation $e_1 - xe_0$ denotes a continuous function on I which is the difference of a linear function y and a constant function with constant value x over I. In other words, it is an abbreviation of $e_1(y) - xe_0(y), y \in I$, which is a function of y rather than x.

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

$$B_F(x) = |F(e_1) - xF(e_0)|, \quad V_F = F\left((e_1 - F(e_1)e_0)^2\right), \tag{56}$$

which represent the "bias" and "variance" of a positive linear functional F.

Lemma 4. *[62, Cor. 2.2.1.] Let $F \colon C(I) \to \mathbb{R}$ be a positive linear functional, where $I \subset \mathbb{R}$ is an interval. Suppose that $F(e_0) = 1, t > 0, \text{length}(I) \geq 2t, s \geq 2$. Then,*

$$|F(f) - f(x)| \leq B_F(x)\frac{\omega^1(f,t)}{t} \\ + \left(1 + \frac{F(|e_1 - xe_0|^s)}{2t^s}\right)\omega^2(f,t). \tag{57}$$

We remark that Lemma 4 can be applied to bound the bias of plug-in estimators in very general models. For example, consider an arbitrary statistical experiment $\{P_\theta, \theta \in I\}$, from which we obtain n i.i.d. samples $X_1, X_2, \ldots, X_n \sim P_\theta$. For any estimator $\hat{\theta}_n$, we would like to analyze the bias of the plug-in estimator $f(\hat{\theta}_n)$ for functional $f(\theta)$.

Suppose $\text{length}(I) \geq 2t, s \geq 2$, then Lemma 4 implies that

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq |\mathbb{E}_\theta \hat{\theta}_n - \theta|\frac{\omega^1(f,t)}{t} \\ + \left(1 + \frac{\mathbb{E}|\hat{\theta}_n - \theta|^s}{2t^s}\right)\omega^2(f,t). \tag{58}$$

If we further assume that $\hat{\theta}_n$ is an unbiased estimator for θ, i.e., $\mathbb{E}_\theta \hat{\theta}_n = \theta$ holds for all $\theta \in I$, then we have

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq \left(1 + \frac{\mathbb{E}|\hat{\theta}_n - \theta|^s}{2t^s}\right)\omega^2(f,t). \tag{59}$$

Taking $s = 2$ and assuming $\mathsf{Var}(\hat{\theta}_n) \leq \text{length}(I)/2$, we have

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq \frac{3}{2}\omega^2(f, \sqrt{\mathsf{Var}(\hat{\theta}_n)}), \tag{60}$$

after we take $t = \sqrt{\mathbb{E}|\hat{\theta}_n - \theta|^2}$.

We remark that Lemma 4 is only one way to analyze the bias, which is by no means always tight. For example, the following estimate using Ditzian–Totik modulus is significantly better than Lemma 4 for certain functions such as the entropy.

Lemma 5. *[62, Thm. 2.5.1.] If $F \colon C[0,1] \to \mathbb{R}$ is a linear positive functional and $F(e_0) = 1$, then we have*

$$|F(f) - f(x)| \leq \frac{B_F(x)}{2h_1\varphi(x)} \cdot \omega_\varphi^1(f, 2h_1) + \frac{5}{2}\omega_\varphi^2(f, h_1), \tag{61}$$

for all $f \in C[0,1]$ and $0 < h_1 \leq \frac{1}{2}$, where $\varphi(x) = \sqrt{x(1-x)}$ and $h_1 = \sqrt{F\left((e_1 - xe_0)^2\right)}/\varphi(x) = \sqrt{V_F + (B_F(x))^2}/\varphi(x)$. The "bias" $B_F(x)$ and "variance" $V_F(x)$ are defined in (56).

Considering the same statistical experiment $\{P_\theta, \theta \in I\}$, and the plug-in estimator $f(\hat{\theta}_n)$ for $f(\theta)$, if $\hat{\theta}_n$ is unbiased for θ and $\mathsf{Var}(\hat{\theta}_n) \leq \frac{\varphi(\theta)^2}{4}$, then it follows from Lemma 5 that

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq \frac{5}{2}\omega_\varphi^2\left(f, \frac{\sqrt{\mathsf{Var}(\hat{\theta}_n)}}{\varphi(\theta)}\right), \tag{62}$$

after we take $t = \frac{\sqrt{\mathsf{Var}(\hat{\theta}_n)}}{\varphi(\theta)}$.

For certain functions $f(\theta)$ and statistical models Lemma 5 is stronger than Lemma 4. For example, if $f(\theta) = -\theta \ln \theta, \theta \in [0,1]$, and we have $n \cdot \hat{\theta}_n \sim \mathsf{B}(n, \theta)$. We will show in Lemma 8 that $\omega_\varphi^2(f,t) = \frac{t^2 \ln 4}{1+t^2}$, and $\omega^2(f,t) = t \ln 4$. We also have $\mathsf{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n}$. Hence, Lemma 4 gives the upper bound

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq \frac{3\ln 4}{2}\sqrt{\frac{\theta(1-\theta)}{n}}, \tag{63}$$

whereas Lemma 5 gives

$$|\mathbb{E}_\theta f(\hat{\theta}_n) - f(\theta)| \leq \frac{5\ln 4}{2n} \cdot \frac{1}{1 + 1/n}, \tag{64}$$

which is much stronger when n is large and θ not too close to the endpoints of $[0,1]$.

There also exist various estimates for the bias when the parameter lies in sets other than an interval in $\mathbb{R}$. However, the bounds we presented are in general *not* optimal for specific functionals, thereby leaving ample room for future development. For example, note that (63) is stronger than (64) when $\theta \leq 1/n$, but Han, Jiao, and Weissman [65] showed that when $\theta \leq 1/n$ the pointwise bound in (63) is still strictly suboptimal for the entropy functional. Unsurprisingly, to obtain the results in Section III, we need to go beyond the general results in approximation theory, and incorporate the structure of specific functions.

Note: In approximation theory literature, researchers have explored the interactions between general positive linear operator approximation and its probabilistic counterpart decades ago [66]–[68]. However, in statistics literature related to positive linear approximation, usually only specific operators are used, such as the Bernstein operator [69], and the focus may not be on obtaining the tightest bound on bias [70], [71].

C. Lower bounds

To lower bound the worst case performance of a specific estimator, we have essentially two approaches: first, to analyze the bias or the variance of the specific estimator carefully; second, to prove a lower bound that is satisfied by all the estimators, which naturally include the specific estimator we need to analyze. These two approaches have different relative advantages and disadvantages, so we utilize them together in the lower bound construction.

We refer the readers to Tsybakov [72] for a nice collection of techniques to prove minimax lower bounds. One specific approach we use is the van Trees inequality, which we quote below.

Let $(\mathcal{X}, \mathcal{F}, P_\theta; \theta \in \Theta)$ be a dominated family of distributions on some sample space $\mathcal{X}$; denote the dominating measure

by μ. Assume Θ is a closed interval on the real line. Let $f(x|\theta)$ denote the density of P_θ with respect to μ. Let π be some probability distribution on Θ with a density $\lambda(\theta)$ with respect to Lebesgue measure. Suppose that λ and $f(x|\cdot)$ are both absolutely continuous (μ-almost surely), and that λ converges to zero at the endpoints of the interval Θ. We define

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \qquad (65)$$

$$\mathcal{I}(\lambda) = \mathbb{E} \left(\frac{\mathrm{d} \log \lambda(\theta)}{\mathrm{d}\theta} \right)^2 \qquad (66)$$

the Fisher information for θ and for a location parameter in λ, respectively. We assume $\mathcal{I}(\theta)$ is continuous in θ. We have the following inequality.

Lemma 6 (van Trees inequality). *[73] Under assumptions above, the average risk of an arbitrary estimator $\hat{\psi}(X)$ in estimating an absolutely continuous functional $\psi(\theta)$ under squared error loss satisfies the following inequality:*

$$\mathbb{E} \left(\hat{\psi}(X) - \psi(\theta) \right)^2 \geq \frac{(\mathbb{E}\psi'(\theta))^2}{\mathbb{E}[\mathcal{I}(\theta)] + \mathcal{I}(\lambda)} \qquad (67)$$

V. PROOFS OF THE UPPER BOUNDS

In order to upper bound the maximum squared error risk of any estimator, a natural approach would be to analyze the squared bias term and the variance term separately. Then, it suffices to find proper tools to give *nonasymptotic* analysis of the bias and variance.

A. Bounding the bias

We first work to bound the bias. Lemma 3 shows that the bias of $F(P_n)$ could be represented as

$$\mathsf{Bias}(F(P_n)) = \sum_{i=1}^{S} \left(B_n[f](p_i) - f(p_i) \right), \qquad (68)$$

where $B_n[f](x)$ is the Bernstein polynomial corresponding to $f(x)$. The following lemma summarizes some state-of-the-art bounds for approximation error of Bernstein polynomials. Lemma 7 can be derived easily from the general theory we presented in Section IV-B2. We emphasize that one cannot expect the bounds in Lemma 7 to be tight for any $f \in C[0,1]$, since the Bernstein approximation error itself could be a very complicated function in $C[0,1]$, and Lemma 7 is using relatively simple functions to upper bound it.

Lemma 7. *The following bounds are valid for function approximation error incurred by Bernstein polynomials:*

1) Pointwise estimate: *[62, Cor. 2.2.1] [74] for all continuous functions f on $[0,1]$,*

$$|f(x) - B_n[f](x)| \leq \frac{3}{2} \omega^2 \left(f, \sqrt{\frac{x(1-x)}{n}} \right), \qquad (69)$$

and the constant $3/2$ is shown by [74] to be the best constant;

2) Norm estimate: *[62, Cor. 4.1.10] for $\varphi(x) = \sqrt{x(1-x)}$ and all continuous functions f on $[0,1]$, we have*

$$\|B_n[f] - f\|_\infty \leq \frac{5}{2} \omega_\varphi^2(f, n^{-1/2}); \qquad (70)$$

3) *[75, Eqn. 10.3.4] for $f \in C^2[0,1]$, i.e., twice continuously differentiable,*

$$|f(x) - B_n[f](x)| \leq \|f''\|_\infty \frac{x(1-x)}{2n}; \qquad (71)$$

Proof. The pointwise estimate of Lemma 7 follows from Lemma 4. The norm estimate of Lemma 7 follows from Lemma 5. Regarding the third part, suppose random variable $X \sim \mathsf{B}(n,x)$. We have

$$|f(x) - B_n[f](x)|$$

$$= |\mathbb{E}_x f(X/n) - f(x)| \qquad (72)$$

$$= |\mathbb{E}_x[f'(x)(X/n - x) + \frac{1}{2}f''(\xi_X)(X/n - x)^2]| \qquad (73)$$

$$= \frac{1}{2}|\mathbb{E}_x f''(\xi_X)(X/n - x)^2| \qquad (74)$$

$$\leq \frac{\|f''\|_\infty}{2}|\mathbb{E}_x(X/n - x)^2| \qquad (75)$$

$$= \frac{\|f''\|_\infty}{2} \frac{x(1-x)}{n}, \qquad (76)$$

where we used Taylor expansion for $f(X/n)$ at point x with the Lagrange remainder. The proof is complete. $\square$

Remark 1. *Note that although (70) is in the form of an upper bound, it has been shown to be a lower bound as well. Totik [76] showed the following equivalence property on the norm estimate of Bernstein approximation errors*

$$\|B_n[f](x) - f(x)\|_\infty \asymp \omega_\varphi^2(f, n^{-1/2}).^2 \qquad (77)$$

It is easy to calculate the second-order modulus of smoothness and the Ditzian–Totik second-order modulus of smoothness for functions x^α and $-x \ln x$. The results are presented in the following lemma.

Lemma 8. *We have*

	$x^\alpha, 0 < \alpha < 1$	$x^\alpha, 1 < \alpha < 2$	$-x\ln x$
$\omega^2(f,t)$	$|2 - 2^\alpha|t^\alpha$	$|2 - 2^\alpha|t^\alpha$	$t\ln 4$
$\omega_\varphi^2(f,t)$	$|2 - 2^\alpha|\frac{t^{2\alpha}}{(1+t^2)^\alpha}$	$\asymp t^2$	$\frac{t^2\ln 4}{1+t^2}$

where the second-order modulus results hold for $0 < t \leq 1/2$, and the Ditizan–Totik second-order modulus results hold for $0 < t \leq 1$.

1) Bias of $F_\alpha(P_n)$: We first bound the bias incurred by $F_\alpha(P_n)$.

1) $\alpha \geq 2$:

---

[2]Note that it is a remarkable fact that (77) holds for any continuous function $f(x)$. The lower bound proof of (77) is considered one of the remarkable results in approximation theory, and currently there are no "short" proofs of this fact. Indeed, Ditzian [77, Section 8] mentioned that *"I still would like to see a new simple proof of (8.4) (Equation (77)) which I am sure will have implications for other operators."*

In this case, $f \in C^2[0,1]$, applying the third part of Lemma 7,

$$|f(x) - B_n[f](x)| \leq \frac{\alpha(\alpha-1)x(1-x)}{2n}. \quad (78)$$

Thus, we have

$$|\text{Bias}(F_\alpha(P_n))| \leq \sum_{i=1}^{S} \alpha(\alpha-1)\frac{p_i(1-p_i)}{2n} \leq \frac{\alpha(\alpha-1)}{2n}. \quad (79)$$

2) $1 < \alpha < 2$

The following lemma presents a bound on the bias of $F_\alpha(P_n)$, which does not depend on the alphabet size S. We note that the proof of Lemma 9 heavily utilizes the special properties of function x^α and the fact that $\sum_{i=1}^{S} p_i = 1$.

Lemma 9. *The bias of $F_\alpha(P_n)$ for estimating $F_\alpha(P), 1 < \alpha < 2$, is upper bounded by the following:*

$$|\text{Bias}(F_\alpha(P_n))| \leq \frac{4}{n^{\alpha-1}}. \quad (80)$$

We also present two additional bounds involving the alphabet size S. Using the pointwise estimate in Lemma 7, the bias term of the MLE is upper bounded as follows for all $0 < \alpha < 2, \alpha \neq 1$:

$$\sum_{i=1}^{S} \frac{3}{2}|2 - 2^\alpha| \left(\frac{p_i(1-p_i)}{n}\right)^{\alpha/2}$$

$$\leq \frac{3}{2}|2 - 2^\alpha| \frac{1}{n^{\alpha/2}} \sum_{i=1}^{S} p_i^{\alpha/2} \quad (81)$$

$$\leq \frac{3}{2}|2 - 2^\alpha| \frac{1}{n^{\alpha/2}} S \frac{1}{S^{\alpha/2}} \quad (82)$$

$$= \frac{3}{2}|2 - 2^\alpha| \frac{S^{1-\alpha/2}}{n^{\alpha/2}}. \quad (83)$$

Using the norm estimate in Lemma 7, when $1 < \alpha < 2$, the bias would be upper bounded by $C_{\alpha,n}\frac{5S}{2n}$, where $C_{\alpha,n} = n\omega_\varphi^2(x^\alpha, n^{-1/2})$ is a finite positive constant such that $\limsup_{n\to\infty} C_{\alpha,n} < \infty$ for $1 < \alpha < 2$. Combining Lemma 9, the pointwise estimate, and the norm estimate in Lemma 7, we know that the bias of $F_\alpha(P_n)$ for $1 < \alpha < 2$ is upper bounded as

$$|\text{Bias}(F_\alpha(P_n))| \leq \frac{4}{n^{\alpha-1}} \wedge \frac{3}{2}|2 - 2^\alpha|\frac{S^{1-\alpha/2}}{n^{\alpha/2}} \wedge C_{\alpha,n}\frac{5S}{2n}. \quad (84)$$

3) $0 < \alpha < 1$:

The pointwise estimate from Lemma 7 is worked out in (83). Using the norm estimate in Lemma 7, the bias would be upper bounded by $|2 - 2^\alpha|\frac{5S}{2n^\alpha}$. Combining the pointwise estimate and the norm estimate, we know that the bias of $F_\alpha(P_n)$ for $0 < \alpha < 1$ is upper bounded as

$$|\text{Bias}(F_\alpha(P_n))| \leq \frac{3}{2}|2 - 2^\alpha|\frac{S^{1-\alpha/2}}{n^{\alpha/2}} \wedge |2 - 2^\alpha|\frac{5S}{2n^\alpha}. \quad (85)$$

2) *Bias of $H(P_n)$:* We then bound the bias incurred by $H(P_n)$. Using the norm estimate in Lemma 7, we know

$$|\text{Bias}(H(P_n))| \leq \frac{5S\ln 4}{2n}. \quad (86)$$

Using the pointwise estimate in Lemma 7, we obtain

$$|\text{Bias}(H(P_n))| \leq \frac{3}{2}\sqrt{\frac{S}{n}}\ln 4. \quad (87)$$

It was shown by Paninski [14, Prop. 1] that the squared bias of MLE $H(P_n)$ is upper bounded as

$$(\text{Bias}(H(P_n)))^2 \leq \left(\ln\left(1 + \frac{S-1}{n}\right)\right)^2, \quad (88)$$

which is better than the two bounds we obtained using Bernstein polynomial results. However, we remark that (88) is obtained using special properties of the entropy function and connections between KL-divergence and χ^2-divergence [72], which cannot be applied to general functions. Strukov and Timan [66] also heavily exploited the structure of function x^α and $-x\ln x$ in order to analyze the Bernstein approximation error for these functions, and obtained tight-in-order results.

3) *Bias of $H(\hat{P}_B)$:* We apply the general theory of positive linear operator approximation. The following lemma is a strengthened version of Lemma 5.

Lemma 10. *If $F: C[0,1] \to \mathbb{R}$ is a linear positive functional and $F(e_0) = 1$, then*

$$|F(f) - f(x)| \leq \omega^1(f, B_F(x); x) + \frac{5}{2}\omega_\varphi^2(f, h_2) \quad (89)$$

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

$$\omega^1(f, h; x) \triangleq \sup\{|f(u) - f(x)| : u \in [0,1], |u - x| \leq h\}. \quad (90)$$

The "bias" $B_F(x)$ and "variance" $V_F(x)$ are defined in (56).

Proof. Applying Lemma 5 to $x = F(e_1)$ we have

$$|F(f) - f(F(e_1))| \leq \frac{5}{2}\omega_\varphi^2(f, h_2) \quad (91)$$

and then (89) is the direct result of the triangle inequality $|F(f) - f(x)| \leq |F(f) - f(F(e_1))| + |f(F(e_1)) - f(x)|$. $\square$

We show that Lemma 10 is indeed stronger than Lemma 5. Firstly, due to $h_1 \geq h_2$, we have $\omega_\varphi^2(f, h_2) \leq \omega_\varphi^2(f, h_1)$. Second, for $x \leq 1/2$, we have

$$\frac{B_F(x)}{2h_1\varphi(x)} \cdot \omega_\varphi^1(f, 2h_1) \approx \frac{B_F(x)}{2h_1\varphi(x)} \cdot \sup_{0 \leq s \leq 1} 2h_1\varphi(s)f'(s) \quad (92)$$

$$\geq B_F(x) \cdot \sup_{x \leq s \leq 1-x} f'(s) \quad (93)$$

$$\approx \sup_{x \leq s \leq 1-x} \omega^1(f, B_F(x); s) \quad (94)$$

which is almost the supremum of $\omega^1(f, |F(e_1 - xe_0)|; s)$ over $s \in [x, 1-x]$ and is no less than the pointwise result $\omega^1(f, |F(e_1 - xe_0)|; x)$, and here we have used the inequality $\varphi(s) \geq \varphi(x)$ for $x \leq s \leq 1-x$. A similar argument also holds

for $x > 1/2$. Hence, Lemma 10 transforms the first order term from the norm result in Lemma 5 to a pointwise result.

Applying Lemma 10 to the function $f(p) = -p \ln p$ and $F(f) = \mathbb{E}\left[f\left(\frac{n\hat{p}+a}{n+Sa}\right)\right]$, where $n \cdot \hat{p} \sim \mathsf{B}(n,p)$, we have the following lemma.

Lemma 11. *If $n \geq \max\{Sa, 2ea, 4\}$, then*

$$\sup_{P \in \mathcal{M}_S} |\mathbb{E}_P H(\hat{P}_B) - H(P)|$$
$$\leq \frac{5nS \ln 2}{(n+Sa)^2} + \frac{2Sa}{n+Sa} \ln\left(\frac{n+Sa}{2a}\right). \quad (95)$$

Note that Lemma 11 implies a slightly weaker bias bound than Theorem 4, but it is only sub-optimal up to a multiplicative constant. The bias bound in Theorem 4 is obtained using the following lemma, whose proof only applies to the entropy function.

Lemma 12. *If $n \geq \max\{2ea, Sa\}$,*

$$\sup_{P \in \mathcal{M}_S} |\mathbb{E}_P H(\hat{P}_B) - H(P)|$$
$$\leq \ln\left(1 + \frac{S-1}{n+Sa}\right) \vee \frac{2Sa}{n+Sa} \ln\left(\frac{n+Sa}{2a}\right). \quad (96)$$

B. Bounding the variance

The next lemma follows from an application of bounded difference inequality presented in Lemma 2.

Lemma 13. *The variance of $F(P_n)$ satisfies the following upper bound:*

$$\mathsf{Var}(F(P_n)) \leq n \cdot \max_{0 \leq j < n} \left(f((j+1)/n) - f(j/n)\right)^2. \quad (97)$$

If f is monotone, then we can strengthen the bound to be

$$\mathsf{Var}(F(P_n)) \leq \frac{n}{4} \cdot \max_{0 \leq j < n} \left(f((j+1)/n) - f(j/n)\right)^2. \quad (98)$$

We first bound the variance for $F_\alpha(P_n), \alpha > 1$. We have

$$\max_{0 \leq j < n} \left(((j+1)/n)^\alpha - (j/n)^\alpha\right)^2 \leq \left(1 - \left(1 - \frac{1}{n}\right)^\alpha\right)^2 \quad (99)$$
$$\leq \left(\frac{\alpha}{n}\right)^2, \quad (100)$$

where in the last step we used Bernoulli's inequality: $(1 + x)^r \geq 1 + rx, \forall r \geq 1, x > -1, x \in \mathbb{R}$. Using Lemma 2, we know the variance is upper bounded by

$$\mathsf{Var}(F_\alpha(P_n)) \leq \frac{\alpha^2}{4n}. \quad (101)$$

We bound the variance of $F_\alpha(P_n), 0 < \alpha < 1$ in the following lemma.

Lemma 14. *For $0 < \alpha < 1/2$, we have*

$$\sup_{P \in \mathcal{M}_S} \mathsf{Var}(F_\alpha(P_n))$$
$$\leq \frac{10S}{n^{2\alpha}}$$
$$+ \left(\frac{3\alpha \cdot 2^{3+2\alpha} + 1}{8\alpha^2}\left(\frac{8\alpha}{e}\right)^{2\alpha} + 4\right)\left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}}\right) \quad (102)$$
$$\lesssim \frac{S}{n^{2\alpha}}. \quad (103)$$

For $1/2 \leq \alpha < 1$, we have

$$\sup_{P \in \mathcal{M}_S} \mathsf{Var}(F_\alpha(P_n))$$
$$\leq \frac{10S^{2-2\alpha}}{n}$$
$$+ \left(\frac{3\alpha \cdot 2^{3+2\alpha} + 1}{8\alpha^2}\left(\frac{8\alpha}{e}\right)^{2\alpha} + 4\right)\left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}}\right) \quad (104)$$
$$\lesssim \frac{S^{2-2\alpha}}{n} + \left(\frac{S}{n^{2\alpha}} \wedge \frac{1}{n^{2\alpha-1}}\right). \quad (105)$$

Further, one can show that for all $\alpha \in (0,1)$,

$$\frac{3\alpha \cdot 2^{3+2\alpha} + 1}{8\alpha^2}\left(\frac{8\alpha}{e}\right)^{2\alpha} + 4 \leq \frac{120}{\alpha^2}, \quad (106)$$

which is used in Theorem 1.

Regarding the variance of $H(P_n)$, we have

Lemma 15.

$$\sup_{P \in \mathcal{M}_S} \mathsf{Var}(H(P_n)) \leq \frac{(\ln n)^2}{n} \wedge \frac{2(\ln S + 3)^2}{n} \quad (107)$$
$$\lesssim \frac{(\ln S)^2 \wedge (\ln n)^2}{n}. \quad (108)$$

The variance of $H(\hat{P}_B)$ is upper bounded by the following lemma.

Lemma 16. *The variance of $H(\hat{P}_B)$ is upper bounded as follows:*

$$\mathsf{Var}\left(H(\hat{P}_B)\right) \leq \frac{2n}{(n+Sa)^2}\left[3 + \ln\left(\frac{n+Sa}{a+1} \wedge S\right)\right]^2. \quad (109)$$

VI. PROOFS OF THE LOWER BOUNDS

A. Lower bounds for estimation of $F_\alpha(P)$ when $\alpha \geq 3/2$

We apply the van Trees inequality as presented in Lemma 6.

It suffices to consider the restricted case of $S = 2$ and prove the n^{-1} lower bound. Thus, the model is equivalent to observing a Binomial random variable $X \sim \mathsf{B}(n,p)$, and one aims to estimate the functional $\psi_\alpha(p) = p^\alpha + (1-p)^\alpha$. We have

$$\psi'_\alpha(p) = \alpha p^{\alpha-1} - \alpha(1-p)^{\alpha-1}. \quad (110)$$

The Fisher information for parameter p under the Binomial model is $\mathcal{I}(p) = \frac{n}{p(1-p)}$. Suppose we impose prior $\lambda(p)$ on parameter p. The van Trees inequality implies

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$

$$\geq \inf_{\hat{F}_\alpha} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F}_\alpha - F_\alpha(P) \right)^2 \qquad (111)$$

$$\geq \mathbb{E} \left(\mathbb{E}[F_\alpha(P)|X_1^S] - F_\alpha(P) \right)^2 \quad \text{(Bayes risk)} \quad (112)$$

$$\geq \frac{\left(\int \left[\alpha p^{\alpha-1} - \alpha(1-p)^{\alpha-1} \right] \lambda(p) \mathrm{d}p \right)^2}{\mathbb{E}_\lambda \left[\frac{n}{p(1-p)} \right] + \mathcal{I}(\lambda)} \qquad (113)$$

$$= \frac{\left(\int \left[\alpha p^{\alpha-1} - \alpha(1-p)^{\alpha-1} \right] \lambda(p) \mathrm{d}p \right)^2}{n \cdot \mathbb{E}_\lambda \left[\frac{1}{p(1-p)} \right] + \mathcal{I}(\lambda)} \qquad (114)$$

where the second inequality follows from the fact that the Bayes risk under any prior is upper bounded by the minimax risk [78].

Taking $\lambda(p)$ to be the Dirichlet prior with parameter (a, b), i.e.,

$$\lambda(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}, a > 2, b > 2, \qquad (115)$$

we can explicitly evaluate the integrals above. Here $B(a,b)$ is the Beta function.

Taking $a = 4, b = 3$, we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2$$

$$\geq \frac{(60\alpha(B(\alpha+3,3) - B(\alpha+2,4)))^2}{5n + 45}. \qquad (116)$$

Taking $C_\alpha = 72\alpha^2 \left(B(\alpha+3,3) - B(\alpha+2,4) \right)^2$, we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P_n) - F_\alpha(P) \right)^2 \geq \frac{C_\alpha}{n}, \quad \text{for all } n \geq 1. \qquad (117)$$

Note that $C_\alpha > 0$ for all $\alpha \geq 3/2$.

B. Lower bounds for estimation of $F_\alpha(P)$ when $1 < \alpha < 3/2$

The following lemma was proved in [69].

Lemma 17. *Let $k \geq 4$ be an even number. Suppose that the k-th derivative of f satisfies $f^{(k)} \leq 0$ in $(0,1)$, Q_{k-1} is the Taylor polynomial of order $k-1$ to f at some x_1 in $(0,1)$. Then for $x \in [0,1]$,*

$$f(x) - B_n[f](x) \geq Q_{k-1} - B_n[Q_{k-1}](x). \qquad (118)$$

Consider $f_\alpha(x) = -x^\alpha, 1 < \alpha < 2, x \in [0,1]$. Applying Lemma 17 to f_α, taking $k = 6$, we have the following result.

Lemma 18. *Suppose $f_\alpha(x) = -x^\alpha, 1 < \alpha < 2$ on $[0,1]$. For*

all $x \in (0,1)$, *we have*

$$f_\alpha(x) - B_n[f_\alpha](x)$$

$$\geq \frac{\alpha(\alpha-1)x^{\alpha-2}(1-x)}{2n} \left(x + \frac{(2-\alpha)(3\alpha-1)x}{12n} \right.$$

$$\left. + \frac{(2-\alpha)(5-3\alpha)}{12n} \right) + \frac{R_1(x)}{n^3} + \frac{R_2(x)}{n^4}, \quad (119)$$

where

$$R_1(x) = \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)x^{\alpha-3}(1-x)}{24}$$

$$\times \left(1 + 2(1-x)((5-2\alpha)x + \alpha - 4) \right), \quad (120)$$

$$R_2(x) = \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)(\alpha-4)}{120}$$

$$\times x^{\alpha-4}(1-x)(1-2x)(1-12x(1-x)). \quad (121)$$

Note that we have assumed $S = cn, c > 0$. If $c \leq 1$, we take a uniform distribution on S elements $P = (1/S, 1/S, \ldots, 1/S)$, otherwise we take distribution $P = (n^{-1} - \epsilon, n^{-1} - \epsilon, \ldots, n^{-1} - \epsilon, \frac{n\epsilon}{S-n}, \ldots, \frac{n\epsilon}{S-n})$, where ϵ will be taken to be arbitrarily small. We first analyze the $c \leq 1$ case. Applying Lemma 18, we have

$$\sum_{i=1}^S f_\alpha(1/S) - B_n[f_\alpha](1/S) \quad \text{(Note that } f_\alpha(x) = -x^\alpha\text{)}$$

$$= \mathbb{E}F_\alpha(P_n) - F_\alpha(P)$$

$$\geq S \cdot \left(\frac{\alpha(\alpha-1)}{2S^{\alpha-2}n} \left(\frac{1}{S} + \frac{(2-\alpha)(5-3\alpha)}{12n} \right) \right.$$

$$+ \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)}{24S^{\alpha-3}n^3}(1 + 2(\alpha-4))$$

$$\left. + \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)(\alpha-4)}{120S^{\alpha-4}n^4} + o(n^{-\alpha}) \right)$$

$$= \frac{\alpha(\alpha-1)}{n^{\alpha-1}} \left(\frac{1}{2c^{\alpha-3}} \left(\frac{1}{c} + \frac{(2-\alpha)(5-3\alpha)}{12} \right) \right.$$

$$+ \frac{(\alpha-2)(\alpha-3)(1+2(\alpha-4))}{24c^{\alpha-4}}$$

$$\left. + \frac{(\alpha-2)(\alpha-3)(\alpha-4)}{120c^{\alpha-5}} \right) + o(n^{-(\alpha-1)})$$

$$= \frac{\alpha(\alpha-1)c^{2-\alpha}}{n^{\alpha-1}} \left(\frac{1}{2} + \frac{(2-\alpha)(5-3\alpha)c}{24} \right.$$

$$+ \frac{(\alpha-2)(\alpha-3)(1+2(\alpha-4))c^2}{24}$$

$$\left. + \frac{(\alpha-2)(\alpha-3)(\alpha-4)c^3}{120} \right) + o(n^{-(\alpha-1)})$$

$$\geq \frac{\alpha c^{2-\alpha}(124 - 330\alpha + 285\alpha^2 - 90\alpha^3 + 11\alpha^4)}{120n^{\alpha-1}} + o(n^{-(\alpha-1)}),$$

where the first inequality follows from Lemma 18, and in the

last step we have taken $c = 1$ in the following expression

$$\frac{1}{2} + \frac{(2-\alpha)(5-3\alpha)c}{24} + \frac{(\alpha-2)(\alpha-3)(1+2(\alpha-4))c^2}{24}$$
$$+ \frac{(\alpha-2)(\alpha-3)(\alpha-4)c^3}{120}, \tag{122}$$

and considered the fact that it is a monotonically decreasing function with respect to c on $(0,1]$ for any $\alpha \in (1, 3/2)$.

For cases when $c > 1$, since we take $P = (n^{-1} - \epsilon, n^{-1} - \epsilon, \ldots, n^{-1} - \epsilon, \frac{n\epsilon}{S-n}, \ldots, \frac{n\epsilon}{S-n})$, by a continuity argument, the analysis is exactly the same as that above when we set $c = 1$ as we can take ϵ as small as possible. One can verify that the function $\alpha(124 - 330\alpha + 285\alpha^2 - 90\alpha^3 + 11\alpha^4)/120$ is positive on interval $(1, 3/2)$. Defining $\sqrt{c_\alpha} = \alpha c^{2-\alpha}(124 - 330\alpha + 285\alpha^2 - 90\alpha^3 + 11\alpha^4)/120 > 0$ when $c \leq 1$, and $\sqrt{c_\alpha} = \alpha(124 - 330\alpha + 285\alpha^2 - 90\alpha^3 + 11\alpha^4)/120 > 0$ when $c > 1$, the proof is completed.

C. Lower bounds for estimation of $F_\alpha(P)$ when $0 < \alpha < 1$

Applying Lemma 17 to function $f_\alpha(x) = x^\alpha, \alpha \in (0, 1)$, taking $k = 4$, we have the following result:

Lemma 19. *For $f_\alpha(x) = x^\alpha$ on $[0, 1]$, $\alpha \in (0, 1), x \in (0, 1)$, we have*

$$f_\alpha(x) - B_n[f_\alpha](x) \geq \frac{\alpha(1-\alpha)}{2n} x^{\alpha-2}(1-x)\left(x - \frac{2-\alpha}{3n}\right). \tag{123}$$

Suppose $n \geq S$. Define distribution $W = (w_1, w_2, \ldots, w_S) \in \mathcal{M}_S$ such that

$$1 \leq i \leq S - 1, w_i = \frac{1}{n}; \quad w_S = 1 - \frac{S-1}{n}. \tag{124}$$

Note that $w_i \geq n^{-1}, 1 \leq i \leq S$. It follows from Lemma 19 that

$$F_\alpha(W) - \mathbb{E}_W F_\alpha(P_n) \geq \sum_{i=1}^{S-1} \frac{\alpha(1-\alpha)}{6n^2}\left(\frac{1}{n}\right)^{\alpha-2}\left(1 - \frac{1}{n}\right) \tag{125}$$

$$= \frac{\alpha(1-\alpha)(S-1)}{6n^\alpha}\frac{n-1}{n}. \tag{126}$$

Thus, we know for all $0 < \alpha < 1$,

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F_\alpha(P) - F_\alpha(P_n)\right)^2$$
$$\geq \frac{\alpha^2(1-\alpha)^2(S-1)^2}{36n^{2\alpha}}\left(1 - \frac{1}{n}\right)^2. \tag{127}$$

It is shown in [10] that the following minimax lower bound holds for estimation of $F_\alpha(P), 1/2 \leq \alpha < 1$.

Lemma 20. *For $\frac{1}{2} \leq \alpha < 1$, we have*

$$\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F} - F_\alpha(P)\right)^2$$
$$\geq \frac{\alpha^2}{32en}\left[(2(S-1))^{1-\alpha} - 2^{-\alpha}\right.$$
$$\left. - \frac{1-\alpha}{4n}\left((2(S-1))^{1-\alpha} + 2^{-\alpha}\right)\right]^2$$
$$- e^{-n/4}S^{2(1-\alpha)}$$
$$\gtrsim \frac{S^{2-2\alpha}}{n}, \tag{128}$$

where the infimum is taken over all possible estimators.

Since this lower bound holds for all possible estimators, it also holds for the MLE $F_\alpha(P_n)$. Since $\max\{a, b\} \geq \frac{1}{2}(a+b)$, we have the desired lower bound.

D. Lower bounds for estimation of $H(P)$

Braess and Sauer [69] derived the following lower bound for the approximation error of Bernstein polynomials for the function $g(x) = -x \ln x$:

Lemma 21. *Define $g(x) = -x \ln x$ on $[0, 1]$. For $x \geq \frac{15}{n}, x \in [0, 1]$, we have*

$$g(x) - B_n[g](x) \geq \frac{1-x}{2n} + \frac{1}{20n^2x} - \frac{x}{12n^2}. \tag{129}$$

Applying Lemma 21 to the estimation of $H(P)$, we know that if $\forall 1 \leq i \leq S, p_i \geq \frac{15}{n}$,

$$H(P) - \mathbb{E}H(P_n) \geq \frac{S-1}{2n} + \frac{1}{20n^2}\left(\sum_{i=1}^{S}\frac{1}{p_i}\right) - \frac{1}{12n^2}. \tag{130}$$

Consider the uniform distribution P with $n \geq 15S$, which guarantees $p_i \geq \frac{15}{n}$. Since

$$\sum_{i=1}^{S}\frac{1}{p_i} \geq S^2, \tag{131}$$

we have

$$\sup_{P \in \mathcal{M}_S} (H(P) - \mathbb{E}H(P_n)) \geq \frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2}. \tag{132}$$

Thus, when $n \geq 15S$,

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P) - H(P_n)\right)^2 \geq \left(\frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2}\right)^2. \tag{133}$$

It was shown in [21, Prop. 1] that the following minimax lower bound holds.

Lemma 22. *There exists a universal constant $c > 0$ such that*

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H} - H(P)\right)^2 \geq c\frac{\ln^2 S}{n}, \tag{134}$$

where the infimum is taken over all possible estimators $\hat{H}$.

Hence, we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(H(P) - H(P_n) \right)^2$$

$$\geq \max \left\{ \left(\frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2} \right)^2, c \frac{\ln^2 S}{n} \right\} \quad (135)$$

$$\geq \frac{1}{2} \left(\frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2} \right)^2 + \frac{c}{2} \frac{\ln^2 S}{n}. \quad (136)$$

Similar arguments can be applied to the Miller–Madow estimator.

E. Lower bounds for entropy estimation using $H(\hat{P}_B)$

Since $H(\hat{P}_B)$ is a specific estimator for entropy, the following lemma is proved via considering several specific distributions.

Lemma 23. *If* $n \geq \max\{15S, Sa, 2ea\}$,

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right|$$

$$\geq \frac{(S-1)a}{4(n+Sa)} \ln \left(\frac{n+Sa}{a} \right) + \frac{S-1}{8n} + \frac{S^2}{80n^2} - \frac{1}{48n^2} \quad (137)$$

If $n < Sa$, *then*

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right| \geq \frac{S-1}{2S} \ln S. \quad (138)$$

If $n < 2ea$, *then*

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right| \geq \frac{S-1}{2e+S} \ln S. \quad (139)$$

If $n < 15S, n \geq 2ea$, *then*

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right|$$

$$\geq \frac{(S-1)a}{4(n+Sa)} \ln \left(\frac{n+Sa}{a} \right) + \frac{\lfloor n/15 \rfloor}{8n} - \frac{1}{16n}. \quad (140)$$

The corresponding results in Theorem 5 follow from Lemma 23, Lemma 22, and the inequality $\max\{a,b\} \geq \frac{a+b}{2}$.

F. Lower bounds for entropy estimation using $\hat{H}^{\mathsf{Bayes}}$

We prove Theorem 6 below. Applying Lemma 25, we have

$$\hat{H}^{\mathsf{Bayes}} \leq \psi(Sa + n + 1) - \sum_{i=1}^{S} \frac{a + X_i}{Sa + n} \psi(a+1) \quad (141)$$

$$= \psi(Sa + n + 1) - \psi(a + 1) \quad (142)$$

$$\leq \ln \left(\frac{Sa + n + e^{-\gamma}}{a + \frac{1}{2}} \right). \quad (143)$$

Since $\hat{H}^{\mathsf{Bayes}}$ is upper bounded by $\ln \left(\frac{Sa+n+e^{-\gamma}}{a+\frac{1}{2}} \right)$ for any empirical observations, the squared error it incurs in Shannon entropy estimation when the true distribution is the uniform distribution is at least

$$\left(\ln \left(\frac{Sa + S/2}{Sa + n + e^{-\gamma}} \right) \right)^2 \quad (144)$$

if $S \geq 2(n+1)$.

APPENDIX A
AUXILIARY LEMMAS

We begin with the definition of the negative association property, which allows us to upper bound the variance by treating each component of the empirical distribution $P_n(i)$ as "independent" random variables.

Definition 1. *[79, Def. 2.1] Random variables $X_1, X_2, \cdots, X_S$ are said to be negatively associated if for any pair of disjoint subsets A_1, A_2 of $\{1, 2, \cdots, S\}$, and any component-wise increasing functions f_1, f_2,*

$$\mathsf{Cov} \left(f_1(X_i, i \in A_1), f_2(X_j, j \in A_2) \right) \leq 0. \quad (145)$$

To verify whether random variables $X_1, X_2, \cdots, X_S$ are negatively associated or not, the following lemma presents a useful criterion.

Lemma 24. *[79, Thm. 2.9] Let $X_1, X_2, \cdots, X_S$ be S independent random variables with log-concave densities. Then the joint conditional distribution of $X_1, X_2, \cdots, X_S$ given $\sum_{i=1}^{S} X_i$ is negatively associated.*

In light of the preceding lemma, we can obtain the following corollary.

Corollary 5. *For any discrete probability distribution vector $P \in \mathcal{M}_S$, the random variables $\mathbf{X} = (X_1, X_2, \cdots, X_S)$ drawn from the multinomial distribution $\mathbf{X} \sim \mathsf{multi}(n; P)$ are negatively associated.*

Proof. Consider the Poissonized model $Y_i \sim \mathsf{Poi}(np_i), 1 \leq i \leq S$ with all Y_i independent, it is straightforward to verify that each Y_i possesses a log-concave distribution. Then conditioning on $\sum_{i=1}^{S} Y_i = n$, we know that $(Y_1, Y_2, \cdots, Y_S)|(\sum_{i=1}^{S} Y_i = n) \sim \mathsf{multi}(n; P)$, hence Lemma 24 yields the desired result. $\square$

The next lemma gives bounds on the digamma functions $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$.

Lemma 25. *[80, Lemma 1.7] The digamma function $\psi(z)$ is the only solution of the functional equation $F(x+1) = F(x) + \frac{1}{x}$ that is monotone, strictly concave on $\mathbb{R}_+$ and satisfies $F(1) = -\gamma$, where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant.*

Let x be a positive real number. Then,

$$\ln(x + 1/2) < \psi(x+1) \le \ln(x + e^{-\gamma}). \qquad (146)$$

If $x \ge 1$, then

$$\ln(x + 1/2) < \psi(x+1) \le \ln(x + e^{1-\gamma} - 1). \qquad (147)$$

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

Lemma 26. *[81, Exercise 4.7] If $X \sim \mathsf{Poi}(\lambda)$ or $X \sim \mathsf{B}(n, \frac{\lambda}{n})$, then for any $\delta > 0$, we have*

$$\mathbb{P}(X \ge (1+\delta)\lambda) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\lambda, \qquad (148)$$

$$\mathbb{P}(X \le (1-\delta)\lambda) \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\lambda \le e^{-\delta^2\lambda/2}. \qquad (149)$$

To establish the upper bound of the variance obtained by the plug-in estimator $F_\alpha(P_n)$, we split p into two different regimes $p \le 1/n$ or $p > 1/n$, and the following lemmas give the corresponding variance bounds.

Lemma 27. *For $nX \sim \mathsf{B}(n,p), p \le 1/n$, we have*

$$\mathsf{Var}(X^\alpha) \le \frac{2}{n^{2\alpha}} \wedge \frac{2p}{n^{2\alpha-1}} \quad 0 < \alpha < 1. \qquad (150)$$

Lemma 28. *For $nX \sim \mathsf{B}(n,p), p \ge 1/n, 0 < \alpha < 1$, we have*

$$\mathsf{Var}(X^\alpha) \le \frac{10p^{2\alpha-1}}{n} + \frac{3}{2\alpha}\left(\frac{16\alpha}{en}\right)^{2\alpha} + \frac{2}{n^{2\alpha}} + \frac{1}{8\alpha^2}\left(\frac{8\alpha}{en}\right)^{2\alpha} \qquad (151)$$

APPENDIX B
PROOFS OF MAIN LEMMAS

A. Proof of Lemma 3

We compute the first moment of $F(P_n)$.

$$\mathbb{E}F(P_n) = \sum_{j=0}^n f\left(\frac{j}{n}\right)\mathbb{E}h_j, \qquad (152)$$

and

$$\mathbb{E}h_j = \mathbb{E}\sum_{i=1}^S \mathbb{1}(X_i = j) \qquad (153)$$

$$= \sum_{i=1}^S \mathbb{P}(X_i = j) \qquad (154)$$

$$= \sum_{i=1}^S \binom{n}{j}p_i^j(1-p_i)^{n-j}. \qquad (155)$$

Thus, we have

$$\mathbb{E}F(P_n) = \sum_{j=0}^n f\left(\frac{j}{n}\right)\sum_{i=1}^S \binom{n}{j}p_i^j(1-p_i)^{n-j} \qquad (156)$$

$$= \sum_{j=0}^n \sum_{i=1}^S f\left(\frac{j}{n}\right)\binom{n}{j}p_i^j(1-p_i)^{n-j}. \qquad (157)$$

The bias of $F(P_n)$ is

$$\begin{aligned}&\mathsf{Bias}(F(P_n))\\ &= \mathbb{E}F(P_n) - F(P) \qquad (158)\\ &= \sum_{i=1}^S\left(\sum_{j=0}^n f\left(\frac{j}{n}\right)\binom{n}{j}p_i^j(1-p_i)^{n-j} - f(p_i)\right).\end{aligned} \qquad (159)$$

B. Proof of Lemma 8

We first compute the second-order modulus. Fix $t, 0 < t \le 1/2$. Defining $M \triangleq \frac{u+v}{2}$, then the computation of second-order modulus is equivalent to maximization of $|f(M-t) - 2f(M) + f(M+t)|$ over interval $M \in [t, 1-t]$, since all the functions we consider are strictly convex or concave over $[0,1]$.

For $f(x) = x^\alpha, 0 < \alpha < 1$, $f(x)$ is strictly concave on $[0,1]$. It follows from Jensen's inequality that

$$g(M) = (M-t)^\alpha - 2M^\alpha + (M+t)^\alpha \le 0, \qquad (160)$$

and it suffices to minimize this function of M in order to obtain the modulus. Taking derivative of $g(M)$, we have

$$g'(M) = \alpha\left((M-t)^{\alpha-1} - 2M^{\alpha-1} + (M+t)^{\alpha-1}\right) \ge 0, \qquad (161)$$

since $x^{\alpha-1}$ is a convex function on $[t, 1-t]$. It implies that the function $g(M)$ is non-decreasing, and the minimum of $g(M)$ over $M \in [t, 1-t]$ is attained at $M = t$, and the minimum value is $g(t) = (2^\alpha - 2)t^\alpha$. Hence, the corresponding second-order modulus is $|2 - 2^\alpha|t^\alpha$.

Analogous procedures computes the second-order modulus for $x^\alpha, 1 < \alpha < 2$ and $-x\ln x$.

Now we consider the computation of Ditzian–Totik second-order modulus. Fix $t, 0 < t \le 1$. Again denote $M \triangleq \frac{u+v}{2} \in [0,1]$. Then the optimization is over the regime $|u - v| \le 2t\varphi(M) = 2t\sqrt{M(1-M)}$. Equivalently, it is the interval $[M - t\sqrt{M(1-M)}, M + t\sqrt{M(1-M)}] \cap [0,1]$.

Since the function $f(x) = -x\ln x$ is strictly convex on $[0,1]$, the maximum of $\left|f(u) - 2f\left(\frac{u+v}{2}\right) + f(v)\right|$ is definitely attained when u and v take the boundary values of the feasible interval $[M - t\sqrt{M(1-M)}, M + t\sqrt{M(1-M)}] \cap [0,1]$.

Define $\Delta \triangleq t\sqrt{\frac{1-M}{M}}$. The feasible interval can be equivalently written as $[M - \Delta M, M + \Delta M] \cap [0,1]$. We have

$$M - t\sqrt{M(1-M)} \ge 0 \Leftrightarrow M \ge \frac{t^2}{1+t^2}, \qquad (162)$$

as well as

$$M + t\sqrt{M(1-M)} \le 1 \Leftrightarrow M \le \frac{1}{1+t^2}. \qquad (163)$$

Hence, it is equivalent to maximize over three regimes:
1) Regime A:
   $u = 0, v = 2M, 0 \le M \le \frac{t^2}{1+t^2}$.
2) Regime B:
   $u = M - \Delta M, v = M + \Delta M, M \in \left[\frac{t^2}{1+t^2}, \frac{1}{1+t^2}\right]$
3) Regime C:
   $u = 2M - 1, v = 1, 1 \ge M \ge \frac{1}{1+t^2}$.

Over regime A, we have

$$\left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right| = 2M \ln 2. \qquad (164)$$

Maximizing over $0 \le M \le \frac{t^2}{1+t^2}$, the maximum value is $\frac{t^2 \ln 4}{1+t^2}$, attained at $M = \frac{t^2}{1+t^2}$.

Over regime C, we have

$$\left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right| = |2M \ln M - (2M-1)\ln(2M-1)|. \qquad (165)$$

Maximizing over $\frac{1}{1+t^2} \le M \le 1$, the maximum is attained at $M = \frac{1}{1+t^2}$, and the maximum value is no more than $\frac{t^2 \ln 4}{1+t^2}$.

Now we consider regime B. Since $M \in \left[\frac{t^2}{1+t^2}, \frac{1}{1+t^2}\right]$ in regime B, we know $\Delta = t\sqrt{\frac{1-M}{M}} \in [t^2, 1]$. We have

$$\left| f(u) - 2f\left(\frac{u+v}{2}\right) + f(v) \right|$$
$$= M \left| (1-\Delta)\ln(1-\Delta) + (1+\Delta)\ln(1+\Delta) \right|. \qquad (166)$$

Since $\Delta = t\sqrt{\frac{1-M}{M}}$ implies $M = \frac{t^2}{t^2 + \Delta^2}$, we can recast the corresponding optimization problem as maximizing

$$\frac{t^2}{\Delta^2 + t^2} \left| (1-\Delta)\ln(1-\Delta) + (1+\Delta)\ln(1+\Delta) \right| \qquad (167)$$

subject to constraint $\Delta \in [t^2, 1]$. One can show that the maximum is always attained at $\Delta = 1$, with the maximum value $\frac{t^2 \ln 4}{1+t^2}$.

To sum up, we conclude that when $0 < t \le 1$, the maximum of the optimization problem defining $\omega_\varphi^2(-x \ln x, t)$ is always attained at $u = 0, v = \frac{2t^2}{1+t^2}$, with the resulting modulus $\frac{t^2 \ln 4}{1+t^2}$.

Analogous computation can also be done for function $x^\alpha, 0 < \alpha < 1$. For the function $x^\alpha, 1 < \alpha < 2$, it is hard to compute the modulus exactly, but it is easy to show that it is of order t^2.

C. Proof of Lemma 9

The bias of $F_\alpha(P_n), 1 < \alpha < 2$ can be expressed as follows:

$$|\text{Bias}(F_\alpha(P_n))| = |\mathbb{E}\sum_{i=1}^{S} P_n^\alpha(i) - p_i^\alpha| \qquad (168)$$

$$\le \left| \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} P_n^\alpha(i) - p_i^\alpha \right|$$

$$+ \left| \mathbb{E}\sum_{i:p_i > \frac{1}{n}} P_n^\alpha(i) - p_i^\alpha \right| \qquad (169)$$

$$\triangleq B_1 + B_2. \qquad (170)$$

Now we bound B_1 and B_2 separately. It follows from Jensen's inequality that for any i,

$$\mathbb{E}P_n^\alpha(i) \ge p_i^\alpha, \quad 1 \le i \le S. \qquad (171)$$

Hence, we have

$$B_1 = \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} P_n^\alpha(i) - p_i^\alpha \qquad (172)$$

$$= \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} P_n^\alpha(i) - \frac{(np_i)^\alpha}{n^\alpha} \qquad (173)$$

$$\le \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} P_n^\alpha(i) - \frac{(np_i)^2}{n^\alpha} \qquad (174)$$

$$= \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} \frac{(nP_n(i))^\alpha}{n^\alpha} - \frac{(np_i)^2}{n^\alpha} \qquad (175)$$

$$\le \mathbb{E}\sum_{i:p_i \le \frac{1}{n}} \frac{(nP_n(i))^2}{n^\alpha} - \frac{(np_i)^2}{n^\alpha} \qquad (176)$$

$$= \sum_{i:p_i \le \frac{1}{n}} \frac{\mathbb{E}(nP_n(i))^2}{n^\alpha} - \frac{(np_i)^2}{n^\alpha} \qquad (177)$$

$$= \sum_{i:p_i \le \frac{1}{n}} \frac{(np_i)^2 + np_i(1-p_i)}{n^\alpha} - \frac{(np_i)^2}{n^\alpha} \qquad (178)$$

$$\le \sum_{i:p_i \le \frac{1}{n}} \frac{np_i}{n^\alpha} \qquad (179)$$

$$= \frac{1}{n^{\alpha-1}}, \qquad (180)$$

where we have used the fact that $nP_n(i) \ge 1$ for any $P_n(i) \ne 0$.

Regarding B_2, we have the following bounds:

$$B_2 = \left| \mathbb{E}\sum_{i:p_i > \frac{1}{n}} P_n^\alpha(i) - p_i^\alpha \right| \qquad (181)$$

$$\le \sum_{i:p_i > \frac{1}{n}} \mathbb{E}|P_n^\alpha(i) - p_i^\alpha| \qquad (182)$$

$$\le \sum_{i:p_i > \frac{1}{n}} \frac{3}{2}|2 - 2^\alpha| \left(\frac{p_i(1-p_i)}{n}\right)^{\alpha/2} \qquad (183)$$

$$\le \frac{3|2 - 2^\alpha|}{2n^{\alpha/2}} \sum_{i:p_i > \frac{1}{n}} p_i^{\alpha/2}, \qquad (184)$$

where the second inequality follows from the pointwise estimate in Lemma 7.

Denoting $|\{i : p_i > \frac{1}{n}\}| = K \le n$, we know

$$\sum_{i:p_i > \frac{1}{n}} p_i^{\alpha/2} \le K^{1-\alpha/2} \le n^{1-\alpha/2}, \qquad (185)$$

which implies that

$$B_2 \le \frac{3|2 - 2^\alpha|}{2n^{\alpha/2}} n^{1-\alpha/2} = \frac{3|2 - 2^\alpha|}{2n^{\alpha-1}}. \qquad (186)$$

Therefore, we have

$$|\text{Bias}(F_\alpha(P_n))| \leq B_1 + B_2 \tag{187}$$

$$\leq \frac{1}{n^{\alpha-1}} + \frac{3|2-2^\alpha|}{2n^{\alpha-1}} \tag{188}$$

$$\leq \frac{3|2-2^\alpha|+2}{2n^{\alpha-1}} \tag{189}$$

$$\leq \frac{4}{n^{\alpha-1}}. \tag{190}$$

D. Proof of Lemma 11

We apply Lemma 10. Note that $h_2 = \frac{\sqrt{n}}{n+Sa}$. In order to ensure that $h_2 \leq 1/2$, it suffices to take $n \geq 4$. Also, since $n \geq Sa$, for any $i, 1 \leq i \leq S$,

$$\frac{|1-p_iS|a}{n+Sa} \leq \frac{Sa}{n+Sa} \leq \frac{1}{2}. \tag{191}$$

Meanwhile, since the function $\sum_{i=1}^S \frac{|1-p_iS|a}{n+Sa}$ is a convex function of $P = (p_1, p_2, \ldots, p_S)$, it attains its maximum at one of the corner points of the simplex. Hence,

$$\sum_{i=1}^S \frac{|1-p_iS|a}{n+Sa} \leq \frac{|1-S|a}{n+Sa} + (S-1) \cdot \frac{a}{n+Sa} \tag{192}$$

$$= \frac{2(S-1)a}{n+Sa}. \tag{193}$$

In light of Lemma 10, we have

$$|\mathbb{E}_P H(\hat{P}_B) - H(P)|$$

$$\leq \sum_{i=1}^S \left(\omega^1 \left(f, \frac{|1-p_iS|a}{n+Sa}; p_i \right) + \frac{5n\ln 2}{(n+Sa)^2} \right) \tag{194}$$

$$\overset{(a)}{\leq} -\left(\sum_{i=1}^S \frac{|1-p_iS|a}{n+Sa} \right) \ln \left(\frac{1}{S} \sum_{i=1}^S \frac{|1-p_iS|a}{n+Sa} \right)$$

$$+ \frac{5nS\ln 2}{(n+Sa)^2} \tag{195}$$

$$\overset{(b)}{\leq} \frac{2Sa}{n+Sa} \ln \left(\frac{n+Sa}{2a} \right) + \frac{5nS\ln 2}{(n+Sa)^2}, \tag{196}$$

where (a) follows from the fact that if $|x-y| \leq 1/2, x, y \in [0,1]$, then $|x\ln x - y\ln y| \leq -|x-y|\ln|x-y|$ [82, Thm. 17.3.3] and Jensen's inequality. Step (b) follows from the fact that the function $-y\ln y$ is monotonically increasing on the interval $[0, e^{-1}]$, and

$$\frac{1}{S} \sum_{i=1}^S \frac{|1-p_iS|a}{n+Sa} \leq \frac{2a}{n+Sa} \tag{197}$$

$$\leq \frac{2a}{n} \tag{198}$$

$$\leq e^{-1}, \tag{199}$$

where in the last step we used the assumption that $n \geq 2ea$.

E. Proof of Lemma 12

We have

$$H(\hat{P}_B) = \sum_{i=1}^S -\hat{p}_{B,i} \ln \hat{p}_{B,i} \tag{200}$$

$$= H(P_B) + \sum_{i=1}^S (p_{B,i} - \hat{p}_{B,i}) \ln p_{B,i} - \sum_{i=1}^S \hat{p}_{B,i} \ln \frac{\hat{p}_{B,i}}{p_{B,i}}. \tag{201}$$

Taking expectations on both sides, we have

$$\mathbb{E}H(\hat{P}_B) - H(P) = H(P_B) - H(P) - \mathbb{E}D(\hat{P}_B\|P_B), \tag{202}$$

where $D(P\|Q) = \sum_{i=1}^S p_i \ln \frac{p_i}{q_i}$ is the KL divergence between distributions P and Q. Since $H(P_B) = H\left(\frac{n}{n+Sa}P + \frac{Sa}{n+Sa}U_S \right)$, where U_S denotes the uniform distribution with alphabet size S, it follows from Jensen's inequality and the concavity of the entropy function that

$$H(P_B) \geq \frac{n}{n+Sa}H(P) + \frac{Sa}{n+Sa}H(U_S) \tag{203}$$

$$\geq H(P). \tag{204}$$

Hence,

$$\left| \mathbb{E}H(\hat{P}_B) - H(P) \right| \leq \max\{H(P_B) - H(P), \mathbb{E}D(\hat{P}_B\|P_B)\}. \tag{205}$$

In order to analyze the bias, it suffices to analyze the two terms separately. We first analyze $\mathbb{E}D(\hat{P}_B\|P_B)$.

It follows from Jensen's inequality that

$$D(P\|Q) = \sum_{i=1}^S p_i \ln \frac{p_i}{q_i} \leq \ln \left(\sum_{i=1}^S \frac{p_i^2}{q_i} \right), \tag{206}$$

whose derivation here follows from Tsybakov [72, Lemma 2.7].

By Jensen's inequality, we have

$$\mathbb{E}D(\hat{P}_B\|P_B) \leq \mathbb{E}\ln \left(\sum_{i=1}^S \frac{\hat{p}_{B,i}^2}{p_{B,i}} \right) \leq \ln \left(\sum_{i=1}^S \frac{\mathbb{E}\hat{p}_{B,i}^2}{p_{B,i}} \right). \tag{207}$$

We also have

$$\sum_{i=1}^S \frac{p_i^2}{q_i} = 1 + \sum_{i=1}^S \frac{(p_i - q_i)^2}{q_i}, \tag{208}$$

and that

$$\mathbb{E}(\hat{p}_{B,i} - p_{B,i})^2 = \frac{n^2}{(n+Sa)^2}\mathbb{E}(\hat{p}_i - p_i)^2 = \frac{np_i(1-p_i)}{(n+Sa)^2}. \tag{209}$$

Hence,

$$\mathbb{E}D(\hat{P}_B\|P_B) \leq \ln \left(1 + \sum_{i=1}^S \frac{np_i(1-p_i)}{(n+Sa)(np_i+a)} \right) \tag{210}$$

$$= \ln \left(1 + \sum_{i=1}^S \frac{np_i(1-p_i)}{(n+Sa)np_i} \frac{np_i}{np_i+a} \right) \tag{211}$$

$$\leq \ln \left(1 + \sum_{i=1}^S \frac{1-p_i}{(n+Sa)} \right), \tag{212}$$

which implies that

$$\mathbb{E}D(\hat{P}_B\|P_B) \leq \ln\left(1 + \frac{S-1}{n+Sa}\right). \qquad (213)$$

Now we consider the deterministic gap $H(P_B) - H(P)$. It follows from a refinement result of Cover and Thomas [82, Thm. 17.3.3] that when $|p_{B,i} - p_i| \leq 1/2$ for all i, we have

$$|H(P_B) - H(P)| \leq -\|P_B - P\|_1 \ln\frac{\|P_B - P\|_1}{S} \qquad (214)$$

$$= S \cdot f\left(\frac{\|P_B - P\|_1}{S}\right), \qquad (215)$$

where $f(x) = -x\ln x, x \in [0,1]$. Note that the condition $n \geq Sa$ ensures that $|p_{B,i} - p_i| \leq 1/2$.

We have

$$\frac{1}{S}\|P_B - P\|_1 = \frac{1}{S}\sum_{i=1}^{S}\frac{Sa}{n+Sa}|p_i - 1/S| \qquad (216)$$

$$= \frac{1}{S}\sum_{i=1}^{S}\frac{|1 - p_iS|a}{n+Sa} \qquad (217)$$

$$\leq \frac{2a}{n+Sa}, \qquad (218)$$

where the last step follows from (192). Since we have assumed $n \geq 2ea$, we have $\frac{2a}{n+Sa} \leq \frac{2a}{n} \leq e^{-1}$. Since the function $f(x) = -x\ln x$ is monotonically increasing on the interval $[0, e^{-1}]$, we know

$$|H(P_B) - H(P)| \leq \frac{2Sa}{n+Sa}\ln\frac{n+Sa}{2a}. \qquad (219)$$

F. Proof of Lemma 13

In our case, apparently $F(P_n)$ is a function of n independent random variables $\{Z_i\}_{1 \leq i \leq n}$ taking values in $\mathcal{Z} = \{1, 2, \ldots, S\}$. Changing one location of the sample would make some symbol with count j to have count $j+1$, and another symbol with count i to have count $i-1$. Then the total change in the functional estimator is

$$f\left(\frac{j+1}{n}\right) - f\left(\frac{j}{n}\right) - f\left(\frac{i}{n}\right) + f\left(\frac{i-1}{n}\right). \qquad (220)$$

If f is monotone, then the total change would be upper bounded by $\max_{0 \leq j < n}|f((j+1)/n) - f(j/n)|$. If f is not monotone, the total change can be upper bounded by $2 \cdot \max_{0 \leq j < n}|f((j+1)/n) - f(j/n)|$. Applying Lemma 2, we have the desired bounds.

G. Proof of Lemma 14

In light of Lemma 27 and 28, we have

$$\sum_{i=1}^{S}\mathsf{Var}(P_n(i)^\alpha)$$

$$= \sum_{i:p_i \leq 1/n}\mathsf{Var}(P_n(i)^\alpha) + \sum_{i:p_i > 1/n}\mathsf{Var}(P_n(i)^\alpha) \qquad (221)$$

$$\leq \sum_{i:p_i \leq 1/n}\frac{2}{n^{2\alpha}} \wedge \frac{2p_i}{n^{2\alpha-1}}$$

$$+ \sum_{i:p_i > 1/n}\left(\frac{10p_i^{2\alpha-1}}{n} + \frac{3}{2\alpha}\left(\frac{16\alpha}{en}\right)^{2\alpha}\right.$$

$$\left. + \frac{2}{n^{2\alpha}} + \frac{1}{8\alpha^2}\left(\frac{8\alpha}{en}\right)^{2\alpha}\right). \qquad (222)$$

We obtain the desired bounds after using the concavity of $x^{2\alpha-1}$ when $1/2 \leq \alpha < 1$.

Now we exploit the negative association property of all random variables $P_n(i), 1 \leq i \leq S$. Corollary 5 and the monotonically increasing property of x^α yield

$$\mathsf{Var}(F_\alpha(P_n)) = \sum_{i=1}^{S}\mathsf{Var}(P_n(i)^\alpha)$$

$$+ 2\sum_{1 \leq i < j \leq S}\mathsf{Cov}(P_n(i)^\alpha, P_n(j)^\alpha) \qquad (223)$$

$$\leq \sum_{i=1}^{S}\mathsf{Var}(P_n(i)^\alpha), \qquad (224)$$

which finishes the proof of Lemma 14.

H. Proof of Lemma 15

The upper bound $(\ln n)^2/n$ follows from Lemma 13. We apply the Efron–Stein inequality (Lemma 1) to obtain the other bound. Denote the n i.i.d. samples from distribution P as $Z_1, Z_2, \ldots, Z_n \in \mathcal{Z}$. Denoting the MLE $H(P_n)$ as $\hat{H}(Z_1, Z_2, \ldots, Z_n)$, since it is invariant to any permutation of $\{Z_1, Z_2, \ldots, Z_n\}$, we know that the Efron–Stein inequality implies

$$\mathsf{Var}(H(P_n)) \leq \frac{n}{2}\mathbb{E}\left(\hat{H}(Z_1, Z_2, \ldots, Z_n) - \hat{H}(Z_1', Z_2, \ldots, Z_n)\right)^2, \qquad (225)$$

where Z_1' is an i.i.d. copy of Z_1.

Recall that

$$X_i = \sum_{j=1}^{n}\mathbb{1}(Z_j = i), \quad 1 \leq i \leq S. \qquad (226)$$

For notional brevity, we denote the S-tuple $(X_1, X_2, \ldots, X_S)$ as X_1^S, and the n-tuple $(Z_1, Z_2, \ldots, Z_n)$ as Z_1^n. A specific realization of $(X_1, X_2, \ldots, X_S)$ is denoted by $x_1^S = (x_1, x_2, \ldots, x_S)$, and a specific realization of $(Z_1, Z_2, \ldots, Z_n)$ is denoted by $z_1^n = (z_1, z_2, \ldots, z_n)$.

In order to upper bound the right hand side of (225), we first condition on $\{X_1, X_2, \ldots, X_S\}$. In other words, we use

$$\mathbb{E}\left(\hat{H}(Z_1, Z_2, \ldots, Z_n) - \hat{H}(Z_1', Z_2, \ldots, Z_n)\right)^2 \quad (227)$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(\hat{H}(Z_1, Z_2, \ldots, Z_n) - \hat{H}(Z_1', Z_2, \ldots, Z_n)\right)^2 \Big| X_1^S\right]\right]. \quad (228)$$

The following lemma calculates the conditional distribution of Z_1 conditioned on $(X_1, X_2, \ldots, X_S)$.

Lemma 29. *The conditional distribution of Z_1 conditioned on $(X_1, X_2, \ldots, X_S)$ is given by the following discrete distribution:*

$$(X_1/n, X_2/n, \ldots, X_S/n). \quad (229)$$

Proof. By definition of conditional distribution, for any $k, 1 \leq k \leq S$, we have

$$\mathbb{P}(Z_1 = k | X_1^S = x_1^S)$$

$$= \frac{\mathbb{P}(Z_1 = k, X_1^S = x_1^S)}{\mathbb{P}(X_1^S = x_1^S)} \quad (230)$$

$$= \frac{\mathbb{P}(Z_1 = k)\mathbb{P}(X_1^S = x_1^S | Z_1 = k)}{\mathbb{P}(X_1^S = x_1^S)} \quad (231)$$

$$= \frac{p_k \binom{n-1}{x_1, x_2, \ldots, x_k-1, \ldots, x_S} p_k^{x_k-1} \prod_{i \neq k} p_i^{x_i}}{\binom{n}{x_1, x_2, \ldots, x_S} \prod_{1 \leq i \leq S} p_i^{x_i}} \quad (232)$$

$$= \frac{\binom{n-1}{x_1, x_2, \ldots, x_k-1, \ldots, x_S}}{\binom{n}{x_1, x_2, \ldots, x_S}} \quad (233)$$

$$= \frac{x_k}{n}, \quad (234)$$

where the multinomial coefficient $\binom{n}{x_1, x_2, \ldots, x_S}$ is defined as

$$\binom{n}{x_1, x_2, \ldots, x_S} = \frac{n!}{\prod_{i=1}^{S} x_i!}. \quad (235)$$

$\square$

Denoting $r(p) = -p \ln p$, we have $r(j/n) \triangleq \frac{-j}{n} \ln \frac{j}{n}$. We rewrite

$$\hat{H}(Z_1', Z_2, \ldots, Z_n) - \hat{H}(Z_1, Z_2, \ldots, Z_n) = D_- + D_+, \quad (236)$$

where

$$D_- = r\left(\frac{X_{Z_1} - 1}{n}\right) - r\left(\frac{X_{Z_1}}{n}\right) \quad (237)$$

$$D_+ = \begin{cases} r\left(\frac{X_{Z_1'} + 1}{n}\right) - r\left(\frac{X_{Z_1'}}{n}\right) & Z_1 \neq Z_1' \\ r\left(\frac{X_{Z_1'}}{n}\right) - r\left(\frac{X_{Z_1'} - 1}{n}\right) & Z_1 = Z_1' \end{cases} \quad (238)$$

Here, D_- is the change in $\hat{H}$ that occurs when Z_1 is removed according to the distribution specified in Lemma 29, and D_+ is the change in $\hat{H}$ that occurs when Z_1' is added back according to the true distribution P.

Now we compute $\mathbb{E}[D_-^2|X_1^S]$ and $\mathbb{E}[D_+^2|X_1^S]$. We have

$$\mathbb{E}[D_-^2|X_1^S] = \sum_{1 \leq i \leq S} \frac{X_i}{n}\left(r\left(\frac{X_i - 1}{n}\right) - r\left(\frac{X_i}{n}\right)\right)^2, \quad (239)$$

and

$$\mathbb{E}[D_+^2|X_1^S]$$

$$= \sum_{1 \leq i \leq S} p_i \frac{X_i}{n}\left(r\left(\frac{X_i}{n}\right) - r\left(\frac{X_i - 1}{n}\right)\right)^2$$

$$+ \sum_{1 \leq i \leq S} p_i\left(1 - \frac{X_i}{n}\right)\left(r\left(\frac{X_i + 1}{n}\right) - r\left(\frac{X_i}{n}\right)\right)^2. \quad (240)$$

Note that we interpret $\frac{X_i}{n}\left(r\left(\frac{X_i}{n}\right) - r\left(\frac{X_i-1}{n}\right)\right)^2$ as 0 when $X_i = 0$. Taking expectations of $\mathbb{E}[D_-^2|X_1^S]$ and $\mathbb{E}[D_+^2|X_1^S]$ with respect to X_1^S, we have

$$\mathbb{E}[D_-^2] = \sum_{1 \leq i \leq S} \sum_{1 \leq j \leq n} \frac{j}{n}\left(r\left(\frac{j-1}{n}\right) - r\left(\frac{j}{n}\right)\right)^2$$

$$\times \mathbb{P}(\mathsf{B}(n, p_i) = j) \quad (241)$$

and

$$\mathbb{E}[D_+^2]$$

$$= \sum_{1 \leq i \leq S} \sum_{0 \leq j \leq n} \left(\frac{j}{n}\left(r\left(\frac{j-1}{n}\right) - r\left(\frac{j}{n}\right)\right)^2\right.$$

$$\left. + \left(1 - \frac{j}{n}\right)\left(r\left(\frac{j}{n}\right) - r\left(\frac{j+1}{n}\right)\right)^2\right)$$

$$\times p_i \mathbb{P}(\mathsf{B}(n, p_i) = j). \quad (242)$$

After some algebra, one can show that $\mathbb{E}[D_+^2] = \mathbb{E}[D_-^2]$. It then follows from (225) that

$$\mathsf{Var}(H(P_n))$$

$$\leq \frac{n}{2} \cdot \mathbb{E}\left(\hat{H}(Z_1, Z_2, \ldots, Z_n) - \hat{H}(Z_1', Z_2, \ldots, Z_n)\right)^2 \quad (243)$$

$$= \frac{n}{2} \cdot \mathbb{E}(D_- + D_+)^2 \quad (244)$$

$$\leq n \cdot \mathbb{E}(D_-^2 + D_+^2) \quad (245)$$

$$\leq 2n\mathbb{E}D_-^2 \quad (246)$$

$$= 2n \cdot \sum_{1 \leq i \leq S} \mathbb{E}P_n(i)\left(r(P_n(i)) - r(P_n(i) - \frac{1}{n})\right)^2 \quad (247)$$

The proof above is an elaborate version of that in [14, App. B.3]. Now we proceed to obtain non-asymptotic upper bounds of (247). For $x \geq 1/n$, it follows from Taylor expansion with integral form residue that

$$(x - \frac{1}{n}) \ln(x - \frac{1}{n}) = x \ln x + (\ln x + 1)(-\frac{1}{n})$$

$$+ \int_x^{x - \frac{1}{n}} (x - \frac{1}{n} - u)\frac{1}{u} du. \quad (248)$$

Then, we have

$$
\left| r(P_n(i)) - r\left(P_n(i) - \frac{1}{n} \right) \right|
$$

$$
\leq \frac{|\ln P_n(i) + 1|}{n} + \left| \int_{P_n(i)}^{P_n(i) - \frac{1}{n}} \frac{P_n(i) - \frac{1}{n}}{u} du \right| + \frac{1}{n}
\tag{249}
$$

$$
\leq \frac{|\ln P_n(i) + 1| + 2}{n}.
\tag{250}
$$

Hence, we have

$$
\mathsf{Var}(H(P_n)) \leq 2n \cdot \sum_{1 \leq i \leq S} \mathbb{E} P_n(i) \left(\frac{|\ln P_n(i) + 1| + 2}{n} \right)^2.
\tag{251}
$$

Noting that $\ln P_n(i) \leq 0$, hence $0 \leq |\ln P_n(i) + 1| \leq 1 - \ln P_n(i)$. We have

$$
\mathsf{Var}(H(P_n)) \leq \frac{2}{n} \cdot \sum_{1 \leq i \leq S} \mathbb{E} P_n(i) \left(\ln P_n(i) - 3 \right)^2
\tag{252}
$$

$$
\leq \frac{2}{n} \sum_{1 \leq i \leq S} p_i (\ln p_i - 3)^2
\tag{253}
$$

$$
\leq \frac{2}{n} S \cdot \frac{1}{S} (-\ln S - 3)^2
\tag{254}
$$

$$
= \frac{2(\ln S + 3)^2}{n},
\tag{255}
$$

where we have used the fact that $x(\ln x - 3)^2$ is a concave function on $[0, 1]$.

I. Proof of Lemma 16

We apply the bounded differences inequality (Lemma 2). In our case, $F(\hat{P}_B)$ is a function of n independent random variables $\{Z_i\}_{1 \leq i \leq n}$ taking values in $\mathcal{Z} = \{1, 2, \cdots, S\}$. Changing one location of the sample would make some symbol with count j to have count $j + 1$, and another symbol with count i to have count $i - 1$. Then the absolute value of the total change in the functional estimator is

$$
\left| f\left(\frac{j + 1 + a}{n + Sa} \right) - f\left(\frac{j + a}{n + Sa} \right) \right.
$$

$$
\left. - f\left(\frac{i + a}{n + Sa} \right) + f\left(\frac{i - 1 + a}{n + Sa} \right) \right|
\tag{256}
$$

$$
\leq 2 \max_{1 \leq k \leq n} \left| f\left(\frac{k + a}{n + Sa} \right) - f\left(\frac{k - 1 + a}{n + Sa} \right) \right|.
\tag{257}
$$

In light of the Taylor expansion with integral form residue, we have that for $1 \geq x \geq t > 0$,

$$
(x - t) \ln(x - t) = x \ln x - t(\ln x + 1) + \int_x^{x-t} \frac{x - t - u}{u} du
\tag{258}
$$

so

$$
|(x - t) \ln(x - t) - x \ln x| \leq t|\ln x + 1|
$$

$$
+ \left| \int_x^{x-t} \frac{x - t}{u} du \right| + t
\tag{259}
$$

$$
\leq t|\ln x + 1| + 2t
\tag{260}
$$

$$
\leq t(3 - \ln x).
\tag{261}
$$

As a result,

$$
\max_{1 \leq k \leq n} \left| f\left(\frac{k + a}{n + Sa} \right) - f\left(\frac{k - 1 + a}{n + Sa} \right) \right|
$$

$$
\leq \max_{1 \leq k \leq n} \frac{1}{n + Sa} \left(3 - \ln\left(\frac{k + a}{n + Sa} \right) \right)
\tag{262}
$$

$$
\leq \frac{1}{n + Sa} \left(3 + \ln\left(\frac{n + Sa}{a + 1} \right) \right).
\tag{263}
$$

Hence, the bounded differences inequality shows that

$$
\mathsf{Var}\left(H(\hat{P}_B) \right)
$$

$$
\leq n \max_{2 \leq k \leq n} \left(f\left(\frac{k + a}{n + Sa} \right) - f\left(\frac{k - 1 + a}{n + Sa} \right) \right)^2
\tag{264}
$$

$$
\leq \frac{n}{(n + Sa)^2} \left(3 + \ln\left(\frac{n + Sa}{a + 1} \right) \right)^2,
\tag{265}
$$

which completes the proof of the first part.

To prove the second part, we use the Efron–Stein inequality (Lemma 1). Since $H(\hat{P}_B) = \hat{H}_B(Z_1, \cdots, Z_n)$ is invariant to any permutation of $(Z_1, Z_2, \cdots, Z_n)$, we know that the Efron–Stein inequality implies

$$
\mathsf{Var}\left(H(\hat{P}_B) \right)
$$

$$
\leq \frac{n}{2} \mathbb{E} \left(\hat{H}_B(Z_1', Z_2, \cdots, Z_n) - \hat{H}_B(Z_1, Z_2, \cdots, Z_n) \right)^2,
\tag{266}
$$

where Z_1' is an i.i.d. copy of Z_1.

Recall that

$$
X_i = \sum_{j=1}^{n} \mathbb{1}(Z_j = i), \qquad 1 \leq i \leq S.
\tag{267}
$$

For brevity, we denote the S-tuple $(X_1, \cdots, X_S)$ as X_1^S, and the n-tuple $(Z_1, \cdots, Z_n)$ as Z_1^n. A specific realization of $(X_1, \cdots, X_S)$ is denoted by $x_1^S = (x_1, \cdots, x_S)$, and a specific realization of $(Z_1, \cdots, Z_n)$ is denoted by $z_1^n = (z_1, \cdots, z_n)$. Then we have

$$
\mathbb{E} \left(\hat{H}_B(Z_1', Z_2, \cdots, Z_n) - \hat{H}_B(Z_1, Z_2, \cdots, Z_n) \right)^2
\tag{268}
$$

$$
= \sum_{x_1^S} \mathbb{P}(X_1^S = x_1^S)
$$

$$
\times \mathbb{E} \left[\left(\hat{H}_B(Z_1', Z_2, \cdots, Z_n) - \hat{H}_B(Z_1, Z_2, \cdots, Z_n) \right)^2 \Big| X_1^S = x_1^S \right].
\tag{269}
$$

In light of Lemma 29, we know that the conditional distribution of Z_1 conditioned on $(X_1, \cdots, X_S)$ is the discrete distribution $(X_1/n, X_2/n, \cdots, X_S/n)$. Denoting $r(p) = f(\frac{np+a}{n+Sa})$,

we can rewrite

$$\hat{H}_B(Z_1', Z_2, \cdots, Z_n) - \hat{H}_B(Z_1, Z_2, \cdots, Z_n) = D_- + D_+ \tag{270}$$

where

$$D_- = r\left(\frac{X_{Z_1} - 1}{n}\right) - r\left(\frac{X_{Z_1}}{n}\right) \tag{271}$$

$$D_+ = \begin{cases} r\left(\frac{X_{Z_1'}+1}{n}\right) - r\left(\frac{X_{Z_1'}}{n}\right) & Z_1 \neq Z_1' \\ r\left(\frac{X_{Z_1'}}{n}\right) - r\left(\frac{X_{Z_1'}-1}{n}\right) & Z_1 = Z_1' \end{cases}. \tag{272}$$

Here, D_- is the change in $\hat{H}_B$ that occurs when Z_1 is removed according to the distribution $(X_1/n, X_2/n, \cdots, X_S/n)$, and D_+ is the change in $\hat{H}$ that occurs when Z_1' is added back according to the true distribution P. Now we have

$$\mathbb{E}[D_-^2|X_1^S] = \sum_{i=1}^S \frac{X_i}{n}\left(r\left(\frac{X_i-1}{n}\right) - r\left(\frac{X_i}{n}\right)\right)^2 \tag{273}$$

$$\mathbb{E}[D_+^2|X_1^S] = \sum_{i=1}^S p_i \frac{X_i}{n}\left(r\left(\frac{X_i}{n}\right) - r\left(\frac{X_i-1}{n}\right)\right)^2$$
$$+ \sum_{i=1}^S p_i \left(1 - \frac{X_i}{n}\right)\left(r\left(\frac{X_i+1}{n}\right) - r\left(\frac{X_i}{n}\right)\right)^2 \tag{274}$$

where we define $r(x) = 0$ when $x \notin [0, 1]$. Then, by the law of iterated expectation, we know that

$$\mathbb{E}[D_-^2] = \sum_{i=1}^S \sum_{j=0}^n \frac{j}{n}\left(r\left(\frac{j-1}{n}\right) - r\left(\frac{j}{n}\right)\right)^2$$
$$\times \mathbb{P}(\mathsf{B}(n, p_i) = j) \tag{275}$$

$$\mathbb{E}[D_+^2] = \sum_{i=1}^S \sum_{j=0}^n \left(\frac{j}{n}\left(r\left(\frac{j-1}{n}\right) - r\left(\frac{j}{n}\right)\right)^2\right.$$
$$\left. + \left(1 - \frac{j}{n}\right)\left(r\left(\frac{j+1}{n}\right) - r\left(\frac{j}{n}\right)\right)^2\right)$$
$$\times p_i \mathbb{P}(\mathsf{B}(n, p_i) = j). \tag{276}$$

After some algebra we can show that $\mathbb{E}[D_-^2] = \mathbb{E}[D_+^2]$.

Hence, we have

$$\mathsf{Var}\left(H(\hat{P}_B)\right)$$
$$\leq \frac{n}{2}\mathbb{E}\left(D_- + D_+\right)^2 \leq n\mathbb{E}\left(D_-^2 + D_+^2\right) = 2n\mathbb{E}D_-^2 \tag{277}$$

$$= 2n\sum_{i=1}^S \mathbb{E}P_n(i)\left(r(P_n(i) - \frac{1}{n}) - r(P_n(i))\right)^2 \tag{278}$$

$$\leq 2n\sum_{i=1}^S \mathbb{E}P_n(i)\left(\frac{1}{n+Sa}\left[3 - \ln\left(\frac{nP_n(i)+a}{n+Sa}\right)\right]\right)^2 \tag{279}$$

$$= \frac{2n}{(n+Sa)^2}\sum_{i=1}^S \mathbb{E}P_n(i)\left(3 - \ln\left(\frac{nP_n(i)+a}{n+Sa}\right)\right)^2 \tag{280}$$

$$\leq \frac{2n}{(n+Sa)^2}\sum_{i=1}^S p_i\left(3 - \ln\left(\frac{np_i+a}{n+Sa}\right)\right)^2 \tag{281}$$

$$\leq \frac{2n}{(n+Sa)^2}S \cdot \frac{1}{S}\left(3 - \ln\left(\frac{n/S+a}{n+Sa}\right)\right)^2 \tag{282}$$

$$= \frac{2n(3 + \ln S)^2}{(n+Sa)^2} \tag{283}$$

where we have used the inequality (261) and Jensen's inequality due to

$$\frac{d^2}{dx^2}\left[x\left(\ln\left(\frac{nx+a}{n+Sa}\right) - 3\right)^2\right]$$
$$= \frac{n}{nx+a}\left[3\left(\ln\left(\frac{nx+a}{n+Sa}\right) - 3\right)\right.$$
$$\left. + \frac{nx}{nx+a}\left(4 - \ln\left(\frac{nx+a}{n+Sa}\right)\right)\right] \tag{284}$$

$$< 0. \tag{285}$$

J. Proof of Lemma 18

It is well known (see, e.g. [75, Cor. 10.4.2]) that if f is concave in $(0, 1)$, then

$$f(x) - B_n[f](x) \geq 0, \quad 0 \leq x \leq 1. \tag{286}$$

Hence we focus on deriving the other bound. For concave function $f_\alpha(x) = -x^\alpha, \alpha \in (1, 2)$, Taylor's polynomial of degree 5 at $x = x_0$ takes the form

$$Q_5(x) = -\frac{\alpha(\alpha-1)}{2}x_0^{\alpha-2}(x-x_0)^2$$
$$- \frac{\alpha(\alpha-1)(\alpha-2)}{6}x_0^{\alpha-3}(x-x_0)^3$$
$$- \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)}{24}x_0^{\alpha-4}(x-x_0)^4$$
$$- \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)(\alpha-4)}{120}x_0^{\alpha-5}(x-x_0)^5$$
$$+ \text{ affine terms of } x$$

We know that the Bernstein polynomial of any affine function on $[0, 1]$ is the affine function itself, hence it suffices to consider the non-affine part of $Q_5(x)$. [69, Prop. 4] showed the following results for Bernstein polynomials:

Lemma 30. *Let $0 \leq x_0 \leq 1$. Then we have*

$$B_n[(x - x_0)^2](x_0) = \frac{x_0(1 - x_0)}{n} \tag{287}$$

$$B_n[(x - x_0)^3](x_0) = \frac{x_0(1 - x_0)}{n^2}(1 - 2x_0) \tag{288}$$

$$B_n[(x - x_0)^4](x_0) = 3\frac{x_0^2(1 - x_0)^2}{n^2}$$
$$+ \frac{x_0(1 - x_0)}{n^3}[1 - 6x_0(1 - x_0)] \tag{289}$$

$$B_n[(x - x_0)^5](x_0) = \left(10\frac{x_0^2(1 - x_0)^2}{n^3}\right.$$
$$+ \frac{x_0(1 - x_0)}{n^4}[1 - 12x_0(1 - x_0)]\bigg)$$
$$\times (1 - 2x_0) \tag{290}$$

Applying Lemma 17 and Lemma 30, taking $x_0 = x$, we have the desired bound.

K. Proof of Lemma 19

It is well known (see, e.g. [75, Cor. 10.4.2]) that if f is concave in $(0, 1)$, then

$$f(x) - B_n[f](x) \geq 0, \quad 0 \leq x \leq 1. \tag{291}$$

Hence we focus on deriving the other bound. For function $f_\alpha(x) = x^\alpha$, Taylor's polynomial of degree 3 at $x = x_0$ takes the form

$$Q_3(x) = \frac{\alpha(\alpha - 1)}{2}x_0^{\alpha-2}(x - x_0)^2$$
$$+ \frac{\alpha(\alpha - 1)(\alpha - 2)}{6}x_0^{\alpha-3}(x - x_0)^3$$
$$+ \text{ affine terms of } x. \tag{292}$$

Applying Lemma 17 and Lemma 30, taking $x_0 = x$, we have

$$Q_3(x) - B_n[Q_3](x)$$
$$= -\frac{\alpha(\alpha - 1)}{2}x^{\alpha-2}\frac{x(1 - x)}{n}$$
$$- \frac{\alpha(\alpha - 1)(\alpha - 2)}{6}x^{\alpha-3}\frac{x(1 - x)}{n^2}(1 - 2x) \tag{293}$$

$$= \frac{\alpha(1 - \alpha)}{2n}x^{\alpha-2}(1 - x)\left(x - \frac{2 - \alpha}{3n}(1 - 2x)\right) \tag{294}$$

$$\geq \frac{\alpha(1 - \alpha)}{2n}x^{\alpha-2}(1 - x)\left(x - \frac{2 - \alpha}{3n}\right). \tag{295}$$

Hence, we have

$$f_\alpha(x) - B_n[f_\alpha](x) \geq Q_3(x) - B_n[Q_3](x) \tag{296}$$
$$\geq \frac{\alpha(1 - \alpha)}{2n}x^{\alpha-2}(1 - x)\left(x - \frac{2 - \alpha}{3n}\right). \tag{297}$$

L. Proof of Lemma 23

By setting $P = (1, 0, 0, \cdots, 0)$, we have $H(P) = 0$ and

$$H(\hat{P}_B) = -\frac{(S - 1)a}{n + Sa}\ln\left(\frac{a}{n + Sa}\right) - \frac{n + a}{n + Sa}\ln\left(\frac{n + a}{n + Sa}\right) \tag{298}$$

$$\geq \frac{(S - 1)a}{n + Sa}\ln\left(\frac{n + Sa}{a}\right), \tag{299}$$

hence we have obtained the first lower bound

$$\sup_{P \in \mathcal{M}_S}\left|\mathbb{E}_P H(\hat{P}_B) - H(P)\right| \geq \frac{(S - 1)a}{n + Sa}\ln\left(\frac{n + Sa}{a}\right). \tag{300}$$

If $n < Sa$, then

$$\sup_{P \in \mathcal{M}_S}\left|\mathbb{E}_P H(\hat{P}_B) - H(P)\right| \geq \frac{(S - 1)a}{2Sa}\ln S \tag{301}$$

$$\geq \frac{S - 1}{2S}\ln S. \tag{302}$$

If $n > 2ea$, then

$$\sup_{P \in \mathcal{M}_S}\left|\mathbb{E}_P H(\hat{P}_B) - H(P)\right| \geq \frac{(S - 1)a}{Sa + 2ea}\ln S \tag{303}$$

$$= \frac{S - 1}{S + 2e}\ln S. \tag{304}$$

From now on we assume $n \geq Sa, n \geq 2ea$. For $n \geq 15S$, it follows from applying Lemma 21 that

$$\sup_{P \in \mathcal{M}_S}\left|\mathbb{E}_P H(\hat{P}) - H(P)\right| \geq \frac{S - 1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2}. \tag{305}$$

If $n < 15S$, then it follows from applying Lemma 21 that one can essentially take $S = \lfloor n/15 \rfloor$ in (305), and obtain

$$\sup_{P \in \mathcal{M}_S}\left|\mathbb{E}_P H(\hat{P}) - H(P)\right| \geq \frac{\lfloor n/15 \rfloor}{2n} - \frac{1}{4n}. \tag{306}$$

It follows from a refinement result of Cover and Thomas [82, Thm. 17.3.3] that when $|\hat{p}_{B,i} - \hat{p}_i| \leq 1/2$ for all i (which is ensured by condition $n \geq Sa$), we have

$$|H(\hat{P}_B) - H(\hat{P})| \leq Sf\left(\frac{\|\hat{P}_B - \hat{P}\|_1}{S}\right), \tag{307}$$

where $f(x) = -x \ln x, x \in [0, 1]$. We have

$$\frac{1}{S}\|\hat{P}_B - \hat{P}\|_1 = \frac{1}{S}\sum_{i=1}^{S}\frac{|S\hat{p}_i - 1|a}{n + Sa} \tag{308}$$

$$\leq \frac{2(S - 1)a}{S(n + Sa)}, \tag{309}$$

where the last step follows from (192). Since we have assumed $n \geq 2ea$, we have $\frac{2a}{n + Sa} \leq \frac{2a}{n} \leq e^{-1}$. Since the function $f(x) = -x \ln x, x \in [0, 1]$ is monotonically increasing when $x \in [0, e^{-1}]$, we have

$$|H(\hat{P}_B) - H(\hat{P})| \leq \frac{2(S - 1)a}{n + Sa}\ln\left(\frac{n + Sa}{a}\right). \tag{310}$$

A combination of these two inequalities yield the second

lower bound

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right|$$

$$\geq \frac{S-1}{2n} + \frac{S^2}{20n^2} - \frac{1}{12n^2} - \frac{2(S-1)a}{n+Sa} \ln\left(\frac{n+Sa}{a}\right) \tag{311}$$

when $n \geq 15S$, and the second lower bound

$$\sup_{P \in \mathcal{M}_S} \left| \mathbb{E}_P H(\hat{P}_B) - H(P) \right|$$

$$\geq \frac{\lfloor n/15 \rfloor}{2n} - \frac{1}{4n} - \frac{2(S-1)a}{n+Sa} \ln\left(\frac{n+Sa}{a}\right) \tag{312}$$

when $n < 15S$.

Hence we are done by using these two lower bounds and the inequality $\max\{a, b\} \geq \frac{3a+b}{4}$.

APPENDIX C
PROOFS OF AUXILIARY LEMMAS

A. Proof of Lemma 27

Since nX is an integer, we have $(nX)^2 \geq (nX)^{2\alpha}, 0 < \alpha < 1$. Hence, for $p \leq 1/n$,

$$\text{Var}(X^\alpha) \leq \mathbb{E}X^{2\alpha} \tag{313}$$

$$= \frac{\mathbb{E}(nX)^{2\alpha}}{n^{2\alpha}} \tag{314}$$

$$\leq \frac{\mathbb{E}(nX)^2}{n^{2\alpha}} \tag{315}$$

$$= \frac{(np)^2 + np(1-p)}{n^{2\alpha}} \tag{316}$$

$$\leq \frac{2}{n^{2\alpha}} \wedge \frac{2p}{n^{2\alpha-1}}, \tag{317}$$

which completes the proof.

B. Proof of Lemma 28

Denoting $f(p) = p^\alpha, 0 < \alpha < 1$, we have

$\text{Var}(f(X))$

$$= \mathbb{E}f^2(X) - (\mathbb{E}f(X))^2 \tag{318}$$

$$= \mathbb{E}f^2(X) - f^2(p) + f^2(p) - (\mathbb{E}f(X))^2 \tag{319}$$

$$\leq |\mathbb{E}f^2(X) - f^2(p)| + |f^2(p) - (\mathbb{E}f(X) - f(p) + f(p))^2| \tag{320}$$

$$= |\mathbb{E}f^2(X) - f^2(p)| + |(\mathbb{E}f(X) - f(p))^2 + 2f(p)(\mathbb{E}f(X) - f(p))| \tag{321}$$

$$\leq |\mathbb{E}f^2(X) - f^2(p)| + |\mathbb{E}f(X) - f(p)|^2 + 2f(p)|\mathbb{E}f(X) - f(p)|. \tag{322}$$

Hence, it suffices to obtain bounds on $|\mathbb{E}f^2(X) - f^2(p)|$ and $|\mathbb{E}f(X) - f(p)|$. Denoting $r(x) = f^2(x)$, it follows from Taylor's formula and the integral representation of the remainder term that

$$r(X) = f^2(p) + r'(p)(X-p) + R_1(X;p) \tag{323}$$

$$R_1(X;p) = \int_p^X (X-u)r''(u)du = \frac{1}{2}r''(\eta_X)(X-p)^2 \tag{324}$$

where $\eta_X \in [\min\{X, p\}, \max\{X, p\}]$.

Similarly, we have

$$f(X) = f(p) + f'(p)(X-p) + R_2(X;p) \tag{325}$$

$$R_2(X;p) = \int_p^X (X-u)f''(u)du = \frac{1}{2}f''(\nu_X)(X-p)^2, \tag{326}$$

where $\nu_X \in [\min\{X, p\}, \max\{X, p\}]$.

Taking expectation on both sides with respect to X, where $nX \sim \mathsf{B}(n, p)$, we have

$$|\mathbb{E}f^2(X) - f^2(p)| = |\mathbb{E}R_1(X;p)|. \tag{327}$$

Similarly, we have

$$|\mathbb{E}f(X) - f(p)| = |\mathbb{E}R_2(X;p)|. \tag{328}$$

It is straightforward to show that

$$|r''(x)| = 2\alpha(2\alpha-1)x^{2\alpha-2} \leq 2x^{2\alpha-2}, \tag{329}$$

$$|f''(x)| = \alpha(1-\alpha)x^{\alpha-2} \leq \frac{1}{4}x^{\alpha-2}. \tag{330}$$

Now we are in the position to bound $|\mathbb{E}R_1(X;p)|$ and $|\mathbb{E}R_2(X;p)|$. For $|\mathbb{E}R_1(X;p)|$, we have

$|\mathbb{E}R_1(X;p)|$

$$\leq \mathbb{E}|R_1(X;p)| \tag{331}$$

$$= \mathbb{E}[|R_1(X;p)\mathbb{1}(X \geq p/2)|] + \mathbb{E}[R_1(X;p)\mathbb{1}(X < p/2)] \tag{332}$$

$$\leq \mathbb{E}\left[2(p/2)^{2\alpha-2}(X-p)^2\right] + \mathbb{E}[R_1(X;p)\mathbb{1}(X < p/2)] \tag{333}$$

$$\leq 8\frac{p^{2\alpha-1}}{n} + \sup_{x \leq p/2} |R_1(x;p)|\mathbb{P}(nX < np/2) \tag{334}$$

$$\leq 8\frac{p^{2\alpha-1}}{n} + \sup_{x \leq p/2} |R_1(x;p)|e^{-np/8}, \tag{335}$$

where in the last step we have used Lemma 26. Regarding $\sup_{x \leq p/2} |R_1(x;p)|$, for any $x \leq p/2$, we have

$$R_1(x;p) = \int_x^p (u-x)r''(u)du \leq \int_x^p (u-x)2u^{2\alpha-2}du \tag{336}$$

$$\leq 2\int_x^p u^{2\alpha-1}du \tag{337}$$

$$\leq 2\int_0^p u^{2\alpha-1}du \tag{338}$$

$$= \frac{p^{2\alpha}}{\alpha}. \tag{339}$$

Hence, we have

$$|\mathbb{E}R_1(X;p)| \leq \frac{8p^{2\alpha-1}}{n} + \frac{1}{\alpha}p^{2\alpha}e^{-np/8}. \tag{340}$$

Analogously, we obtain the following bound for $|\mathbb{E}R_2(X;p)|$:

$$|\mathbb{E}R_2(X;p)| \leq \frac{p^{\alpha-1}}{n} + \frac{1}{4\alpha}p^\alpha e^{-np/8}. \tag{341}$$

Plugging these estimates of $|\mathbb{E}R_1(X;p)|$ and $|\mathbb{E}R_2(X;p)|$

into (322), we have for $p \geq 1/n$,

$$\text{Var}(X^\alpha) \leq \frac{8p^{2\alpha-1}}{n} + \frac{1}{\alpha}p^{2\alpha}e^{-np/8}$$
$$+ \left(\frac{p^{\alpha-1}}{n} + \frac{1}{4\alpha}p^\alpha e^{-np/8} \right)^2$$
$$+ 2f(p)\left(\frac{p^{\alpha-1}}{n} + \frac{1}{4\alpha}p^\alpha e^{-np/8} \right) \quad (342)$$

$$\leq \frac{8p^{2\alpha-1}}{n} + \frac{1}{\alpha}p^{2\alpha}e^{-np/8} + \frac{2p^{2(\alpha-1)}}{n^2}$$
$$+ \frac{1}{8\alpha^2}p^{2\alpha}e^{-np/4} + 2p^\alpha \left(\frac{p^{\alpha-1}}{n} + \frac{1}{4\alpha}p^\alpha e^{-np/8} \right) \quad (343)$$

$$\leq \frac{10p^{2\alpha-1}}{n} + \frac{3}{2\alpha}p^{2\alpha}e^{-np/8} + \frac{2}{n^{2\alpha}} + \frac{1}{8\alpha^2}p^{2\alpha}e^{-np/4} \quad (344)$$

$$\leq \frac{10p^{2\alpha-1}}{n} + \frac{3}{2\alpha}\left(\frac{16\alpha}{en} \right)^{2\alpha} + \frac{2}{n^{2\alpha}} + \frac{1}{8\alpha^2}\left(\frac{8\alpha}{en} \right)^{2\alpha} \quad (345)$$

where we have used the following inequality in the last step: for $x \in (0,1)$ and any $c > 0$,

$$x^{2\alpha}e^{-cnx} \leq \left(\frac{2\alpha}{cen} \right)^{2\alpha}. \quad (346)$$

Note that if $0 < \alpha < 1/2$, we can upper bound $\frac{10p^{2\alpha-1}}{n}$ by $\frac{10}{n^{2\alpha}}$, since we have constrained $p \geq \frac{1}{n}$.

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

[8] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

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

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

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

[12] ——, "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.

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

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

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

[16] 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.

[17] P. Valiant and G. Valiant, "Estimating the unseen: improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, 2013, pp. 2157–2165.

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

[19] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.

[20] J. Jiao, Y. Han, and T. Weissman, "Beyond maximum likelihood: Boosting the Chow-Liu algorithm for large alphabets," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 321–325.

[21] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[22] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "Estimating Rényi entropy of discrete distributions," *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 38–56, 2017.

[23] Y. Wu and P. Yang, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *arXiv preprint arXiv:1504.01227*, 2015.

[24] Y. Han, J. Jiao, and T. Weissman, "Minimax rate-optimal estimation of divergences between discrete distributions," *arXiv preprint arXiv:1605.09124*, 2016.

[25] J. Jiao, Y. Han, and T. Weissman, "Minimax estimation of the L_1 distance," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 750–754.

[26] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence: Optimal minimax rate," *arXiv preprint arXiv:1607.02653*, 2016.

[27] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proceedings of the National Academy of Sciences*, p. 201607774, 2016.

[28] Y. Wu and P. Yang, "Sample complexity of the distinct elements problem," *arXiv preprint arXiv:1612.03375*, 2016.

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

[30] A. Carlton, "On the bias of information estimates." *Psychological Bulletin*, vol. 71, no. 2, p. 108, 1969.

[31] P. Grassberger, "Finite sample corrections to entropy and dimension estimates," *Physics Letters A*, vol. 128, no. 6, pp. 369–373, 1988.

[32] S. Zahl, "Jackknifing an index of diversity," *Ecology*, vol. 58, no. 4, pp. 907–913, 1977.

[33] J. Hausser and K. Strimmer, "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks," *The Journal of Machine Learning Research*, vol. 10, pp. 1469–1484, 2009.

[34] A. Chao and T.-J. Shen, "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample," *Environmental and ecological statistics*, vol. 10, no. 4, pp. 429–443, 2003.

[35] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, "Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data," *BMC bioinformatics*, vol. 5, no. 1, p. 118, 2004.

[36] P. Grassberger, "Entropy estimates from insufficient samplings," *arXiv preprint physics/0307138*, 2008.

[37] M. Vinck, F. P. Battaglia, V. B. Balakirsky, A. H. Vinck, and C. M. Pennartz, "Estimation of the entropy based on its polynomial representation," *Physical Review E*, vol. 85, no. 5, p. 051139, 2012.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," *Advances in neural information processing systems*, vol. 1, pp. 471–478, 2002.

[43] E. Archer, I. M. Park, and J. W. Pillow, "Bayesian estimation of discrete entropy with mixtures of stick-breaking priors," in *Advances in Neural Information Processing Systems*, 2012, pp. 2015–2023.

[44] I. Nemenman, W. Bialek, and R. d. R. van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Physical Review E*, vol. 69, no. 5, p. 056111, 2004.

[45] I. Nemenman, "Coincidences and estimation of entropies of random variables with large cardinalities," *Entropy*, vol. 13, no. 12, pp. 2013–2023, 2011.

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

[47] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6343–6354, 2015.

[48] I. A. Ibragimov and R. Z. Has' minskii, *Statistical estimation: asymptotic theory*. Springer-Verlag New York, 1981, vol. 2.

[49] 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.

[50] 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

[51] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.

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

[53] O. E. Barndorff-Nielsen and D. R. Cox, *Asymptotic techniques for use in statistics*. Chapman & Hall, 1989.

[54] C. G. Small, *Expansions and asymptotics for statistics*. CRC Press, 2010.

[55] B. Efron, "Bootstrap methods: another look at the Jackknife," *The Annals of Statistics*, pp. 1–26, 1979.

[56] P. Hall, *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 1992.

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

[58] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *Information Theory, IEEE Transactions on*, vol. 45, no. 4, pp. 1072–1081, 1999.

[59] P. Flajolet, "Singularity analysis and asymptotics of Bernoulli sums," *Theoretical Computer Science*, vol. 215, no. 1, pp. 371–381, 1999.

[60] J. Cichoń, Z. Golkebiewski, M. Kardas, and M. Klonowski, "On delta-method of moments and probabilistic sums," 2013.

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

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

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

[64] H. H. Gonska, *Quantitative Aussagen zur Approximation durch positive lineare Operatoren*. Gesamthochschule Duisburg, 1979.

[65] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 1372–1376.

[66] L. Strukov and A. Timan, "Mathematical expectation of continuous functions of random variables. smoothness and variance," *Siberian Mathematical Journal*, vol. 18, no. 3, pp. 469–474, 1977.

[67] H. Walk, "Probabilistic methods in the approximation by linear positive operators," in *Indagationes Mathematicae (Proceedings)*, vol. 83, no. 4. Elsevier, 1980, pp. 445–455.

[68] L. Hahn *et al.*, "A note on stochastic methods in connection with approximation theorems for positive linear operators," *Pacific J. Math*, vol. 101, pp. 307–319, 1981.

[69] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.

[70] P. Diaconis and S. Zabell, "Closed form summation for classical distributions: variations on a theme of de moivre," *Statistical Science*, pp. 284–302, 1991.

[71] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008, vol. 2.

[72] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

[73] R. D. Gill and B. Y. Levit, "Applications of the van Trees inequality: a Bayesian Cramér-Rao bound," *Bernoulli*, pp. 59–79, 1995.

[74] R. Paltanea, "On some constants in approximation by Bernstein operators," *General Mathematics*, vol. 16, no. 4, p. 137148, 2008.

[75] R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer, 1993, vol. 303.

[76] V. Totik, "Approximation by Bernstein polynomials," *American Journal of Mathematics*, pp. 995–1018, 1994.

[77] Z. Ditzian, "Polynomial approximation and $\omega_\varphi^r(f, t)$ twenty years later," *Surveys in Approximation Theory*, vol. 3, pp. 106–151, 2007.

[78] A. Wald, *Statistical decision functions*. Wiley, 1950.

[79] K. Joag-Dev and F. Proschan, "Negative association of random variables with applications," *The Annals of Statistics*, vol. 11, no. 1, pp. 286–295, March 1983.

[80] N. Batir, "Inequalities for the gamma function," *Archiv der Mathematik*, vol. 91, no. 6, pp. 554–563, 2008.

[81] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

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

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

Kartik Venkat (S'12) is currently a Research Associate in New York at PDT Partners, a quantitative investment manager. Kartik's research interests include statistical inference, information theory, machine learning and their inter-connections. Kartik received his Ph.D. in Electrical Engineering from Stanford University in 2015. Kartik also received a Masters Degree from the same university in 2012, and a Bachelors degree from the Indian Institute of Technology Kanpur in 2010, both in Electrical Engineering. Kartik was the recipient of the Thomas M. Cover Dissertation Award in 2016 awarded by the IEEE. Kartik was named the 2015 Marconi Society Paul Baran Young Scholar. His other honors include the Jack Keil Wolf Student Best Paper Award at the 2012 International Symposium on Information Theory, a Stanford Graduate Fellowship for Engineering and Sciences, the Numerical Technologies Founders Graduate Prize, and the National Talent Scholarship awarded by the Government of India.

Yanjun Han (S'14) received his B.Eng. degree with the highest honor in Electronic Engineering from Tsinghua University, Beijing, China in 2015, and a Master's degree in Electrical Engineering from Stanford University in 2017. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at Stanford University. His research interests include information theory and statistics, with applications in communications, data compression, and learning.

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

Tsachy's research is focused on information theory, statistical signal processing, the interplay between them, and their applications.

He is recipient of several best paper awards, and prizes for excellence in research.

He served on the editorial board of the IEEE TRANSACTIONS ON INFORMATION THEORY from Sept. 2010 to Aug. 2013, and currently serves on the editorial board of Foundations and Trends in Communications and Information Theory.