Optimal entropy estimation on large alphabets via best polynomial approximation

Yihong Wu and Pengkun Yang

***Abstract*—Consider the problem of estimating the Shannon entropy of a distribution on k elements from n independent samples. We show that the minimax mean-square error is within universal multiplicative constant factors of $(\frac{k}{n\log k})^2+\frac{\log^2 k}{n}$. This implies the recent result of Valiant-Valiant [1] that the minimal sample size for consistent entropy estimation scales according to $\Theta(\frac{k}{\log k})$. The apparatus of best polynomial approximation plays a key role in both the minimax lower bound and the construction of optimal estimators.**

***Index Terms*—large alphabet, high-dimensional statistics, entropy estimation, best polynomial approximation, functional estimation**

I. Introduction

Let P be a distribution over an alphabet of cardinality k. Let $X_1,\ldots,X_n$ be i.i.d. samples drawn from P. Without loss of generality, we shall assume that the alphabet is $[k]\triangleq\{1,\ldots,k\}$. To perform statistical inference on the unknown distribution P or any functional thereof, a sufficient statistic is the histogram $N\triangleq(N_1,\ldots,N_k)$, where

$$N_j=\sum_{i=1}^n \mathbf{1}_{\{X_i=j\}}$$

records the number of occurrences of $j\in[k]$ in the sample. Then $N\sim\text{Multinomial}(n,P)$.

The problem of interest is to estimate the Shannon entropy of distribution P:

$$H(P)=\sum_{i=1}^k p_i\log\frac{1}{p_i}.$$

Entropy estimation has many applications in various fields, such as neuroscience [2] and biomedical research [3], etc. To investigate the decision-theoretic fundamental limit, we consider the minimax quadratic risk of entropy estimation:

$$R^*(k,n)\triangleq\inf_{\hat H}\sup_{P\in\mathcal{M}_k}\mathbb{E}[(\hat H(N)-H(P))^2] \tag{1}$$

where $\mathcal{M}_k$ denotes the set of probability distributions on $[k]$. The goal of the paper is to provide non-asymptotic characterization of the minimax risk $R^*(k,n)$ within constant factors.

From a statistical standpoint, the problem of entropy estimation falls under the category of *functional estimation*, where we are not interested in directly estimating the high-dimensional parameter (the distribution P) per se, but rather a function thereof (the entropy $H(P)$). To estimate a function, perhaps the most natural idea is the "plug-in" approach, namely, first estimate the parameter and then substitute into the function. This leads to the commonly used plug-in estimator, i.e., the empirical entropy,

$$\hat H_{\text{plug-in}}=H(\hat P), \tag{2}$$

where $\hat P=(\hat p_1,\ldots,\hat p_k)$ denotes the empirical distribution with $\hat p_i=\frac{N_i}{n}$. As frequently observed in functional estimation problems, the plug-in estimator suffers from severe bias. Indeed, although $\hat H_{\text{plug-in}}$ is asymptotically efficient in the "fixed-P-large-n" regime, it can be highly suboptimal in high dimensions.

Our main result is the characterization of the minimax risk within universal constant factors:

Theorem 1. *If $n\gtrsim\frac{k}{\log k}$,*[1]

$$R^*(k,n)\asymp\left(\frac{k}{n\log k}\right)^2+\frac{\log^2 k}{n}. \tag{3}$$

If $n\lesssim\frac{k}{\log k}$, there exists no consistent estimators, i.e., $R^*(k,n)\gtrsim 1$.

To interpret the minimax rate (3), we note that the second term corresponds to the classical "parametric" term inversely proportional to $\frac{1}{n}$, which is governed by the variance and the central limit theorem (CLT). The first term corresponds to the squared bias, which is the main culprit in the regime of insufficient samples. Note that $R^*(k,n)\asymp(\frac{k}{n\log k})^2$ if and only if $n\lesssim\frac{k^2}{\log^4 k}$, where the bias dominates. As a consequence, the minimax rate in Theorem 1 implies that to estimate the entropy within ϵ bits with probability, say 0.9, the minimal sample size is given by

$$n\asymp\frac{\log^2 k}{\epsilon^2}\vee\frac{k}{\epsilon\log k}. \tag{4}$$

Next we evaluate the performance of plug-in estimator in terms of its worst-case mean-square error

$$R_{\text{plug-in}}(k,n)\triangleq\sup_{P\in\mathcal{M}_k}\mathbb{E}[(\hat H_{\text{plug-in}}(N)-H(P))^2]. \tag{5}$$

This research was supported by the National Science Foundation under Grant IIS-1447879, and CCF-1423088, and University of Illinois College of Engineering Strategic Research Initiatives. The authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL, {yihongwu,pyang14}@illinois.edu.

[1] For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n\gtrsim b_n$ or $b_n\lesssim a_n$ when $a_n\geq cb_n$ for some absolute constant c. Finally, we write $a_n\asymp b_n$ when both $a_n\gtrsim b_n$ and $a_n\lesssim b_n$ hold.

Analogous to Theorem 1 which applies to the optimal estimator, the risk of the plug-in estimator admits a similar characterization:

Proposition 1. *If $n \gtrsim k$, then*

$$R_{\textit{plug-in}}(k,n) \asymp \left(\frac{k}{n}\right)^2 + \frac{\log^2 k}{n}. \tag{6}$$

If $n \lesssim k$, then $\hat{H}_{\textit{plug-in}}$ is inconsistent, i.e., $R_{\textit{plug-in}}(k,n) \gtrsim 1$.

Note that the first and second term in the risk (6) again corresponds to the squared bias and variance, respectively. Comparing (3) and (6), we reach the following verdict on the plug-in estimator: Empirical entropy is rate-optimal, i.e., achieving a constant factor of the minimax risk, if and only if we are in the "data-rich" regime $n = \Omega(\frac{k^2}{\log^2 k})$. In the "data-starved" regime of $n = o(\frac{k^2}{\log^2 k})$, empirical entropy is strictly rate-suboptimal.

A. Previous results

Below we give a concise overview of the previous results on entropy estimation.

a) Fixed alphabet: For fixed distribution P and $n \to \infty$, Antos and Kontoyiannis showed that [4] the plug-in estimator is always consistent and the asymptotic variance of the plug-in estimator is obtained in [5]. However, the convergence rate of the bias can be arbitrarily slow on a possibly infinite alphabet. The asymptotic expansion of the bias is obtained in, e.g., [6], [7]:

$$\mathbb{E}[\hat{H}_{\text{plug-in}}(N)] = H(P) - \frac{k-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{i=1}^{k}\frac{1}{p_i}\right) + O(n^{-3}) \tag{7}$$

which inspired various types of bias correction to the plug-in estimator.

b) Large alphabet: It is well-known that to estimate the distribution P itself, say, under the total variation loss, we need at least $\Theta(k)$ samples. However, to estimate the entropy $H(P)$ which is a scalar function, it is unclear from first principles whether $n = \Theta(k)$ is necessary. Using non-constructive arguments, Paninski first proved that it is possible to consistently estimate the entropy using *sublinear* sample size, i.e., there exists $n_k = o(k)$, such that $R^*(k, n_k) \to 0$ as $k \to \infty$ [8]. Valiant proved that no consistent estimator exists, i.e., $R^*(k, n_k) \gtrsim 1$ if $n \lesssim \frac{k}{\exp(\sqrt{\log k})}$ [9]. The sharp scaling of the minimal sample size of consistent estimation is shown to be $\frac{k}{\log k}$ in the breakthrough results of Valiant and Valiant [1], [10]. However, the optimal sample size as a function of alphabet size k and estimation error ϵ has not been completely resolved. Indeed, an estimator based on linear programming is shown to achieve an additive error of ϵ using $\frac{k}{\epsilon^2 \log k}$ samples [11, Theorem 1], while $\frac{k}{\epsilon \log k}$ samples are shown to be necessary [10, Corollary 10]. This gap is partially amended in [12] by proposing a different estimator, which requires $\frac{k}{\epsilon \log k}$ samples but only valid when $\epsilon > k^{-0.03}$. Theorem 1 generalizes their result by characterizing the full minimax rate and the sharp sample complexity is given by (4).

We briefly discuss the difference between the lower bound strategy of [10] and ours. Since the entropy is a permutation-invariant functional of the distribution, a sufficient statistic for entropy estimation is the histogram of the histogram N:

$$h_i = \sum_{j=1}^{k} \mathbf{1}_{\{N_j = i\}}, \quad i \in [n], \tag{8}$$

also known as *fingerprint* [10], which is the number of symbols that appear exactly i times in the sample. A canonical approach to obtain minimax lower bounds for functional estimation is Le Cam's two-point argument [13, Chapter 2], i.e., finding two distributions which have very different entropy but induce almost the same distribution for the sufficient statistics, in this case, the histogram N_1^k or the fingerprints h_1^n, both of which have non-product distributions. A frequently used technique to reduce dependence is *Poisson sampling* (see Section II), where we relax the fixed sample size to a Poisson random variable with mean n. This does not change the statistical nature of the problem due to the exponential concentration of the Poisson distribution near its mean. Under the Poisson sampling model, the sufficient statistics $N_1, \ldots, N_k$ are independent Poissons with mean np_i; however, the entries of the fingerprint remain highly dependent. To contend with the difficulty of computing statistical distance between high-dimensional distributions with dependent entries, the major tool in [10] is a new CLT for approximating the fingerprint distribution by quantized Gaussian distribution, which are parameterized by the mean and covariance matrices and hence more tractable. This turns out to improve the lower bound in [9] obtained using Poisson approximation.

In contrast, in this paper we shall not deal with the fingerprint directly, but rather use the original sufficient statistics N_1^k due to their independence endowed by the Poissonized sampling. Our lower bound relies on choosing two random distributions (priors) with almost iid entries which effectively reduces the problem to one dimension, thus circumventing the hurdle of dealing with high-dimensional non-product distributions. The main intuition is that a random vector with iid entries drawn from a positive unit-mean distribution is not exactly but *sufficiently close* to a probability vector due to the law of large numbers, so that effectively it can be used as a prior in the minimax lower bound.

B. Best polynomial approximation

The proof of both the upper and the lower bound in Theorem 1 relies on the apparatus of *best polynomial approximation*. Our inspiration comes from previous work on functional estimation in Gaussian mean models [14], [15]. Nemirovski (credited in [16]) pioneered the use of polynomial approximation in functional estimation and showed that unbiased estimators for the truncated Taylor series of the smooth functionals is asymptotically efficient. This strategy is generalized to non-smooth functionals in [14] using best polynomial approximation and in [15] for estimating the ℓ_1-norm in Gaussian mean model.

On the constructive side, the main idea is to trade bias with variance. Under the iid sampling model, it is easy to show (see, e.g., [17, Proposition 8]) that to estimate a functional $f(P)$ using n samples, an unbiased estimator exists if and only if $f(P)$ is a polynomial in P of degree at most n. Similarly, under Poisson sample model, $f(P)$ admits an unbiased estimator if and only if f is real analytic. Consequently, there exists no unbiased entropy estimator with or without Poissonized sampling. Therefore, a natural idea is to approximate the entropy functional by polynomials which enjoy unbiased estimation, and reduce the bias to at most the uniform approximation error. The choice of the degree aims to strike a good bias-variance balance. Shortly before we posted this paper to arxiv, we learned that Jiao et al. [18] independently used the idea of best polynomial approximation in constructing rate-optimal estimators for Shannon entropy and power sums with a slightly different procedure.

While the use of best polynomial approximation on the constructive side is admittedly natural, the fact that it also arises in the optimal lower bound is perhaps surprising. As carried out in [14], [15], the strategy is to choose two priors with matching moments up to a certain degree, which ensures the impossibility to test. The minimax lower bound is then given by the maximal separation in the expected functional values subject to the moment matching condition. This problem is the *dual* of best polynomial approximation in the optimization sense. For entropy estimation, this approach yields the optimal minimax lower bound, although the argument is considerably more involved due to the constraint on the mean of the prior.

C. Notations

Throughout the paper all logarithms are with respect to the natural base and the entropy is measured in nats. Let $\mathrm{Poi}(\lambda)$ denote the Poisson distribution with mean λ whose probability mass function is $\mathrm{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}, j \in \mathbb{Z}_+$. Given a distribution P, its n-fold product is denoted by $P^{\otimes n}$. For a parametrized family of distributions $\{P_\theta\}$ and a prior π, the mixture is denoted by $\mathbb{E}_\pi[P_\theta] = \int P_\theta \pi(\mathrm{d}\theta)$. In particular, $\mathbb{E}[\mathrm{Poi}(U)]$ denotes the Poisson mixture with respect to the distribution of a positive random variable U. The total variation and Kullback-Leibler divergence between probability measures P and Q are respectively given by $\mathsf{TV}(P,Q) = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$ and $D(P\|Q) = \int \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}$.

All proofs are omitted due to space limitations and referred to [19].

II. Poisson Sampling

The multinomial distribution of the sufficient statistic $N = (N_1, \ldots, N_k)$ is difficult to analyze because of the dependency. A commonly used technique is the so-called *Poisson sampling*, where we relax the sample size n from being deterministic to a Poisson random variable n' with mean n. Under this model, we first draw the sample size $n' \sim \mathrm{Poi}(n)$, then draw n' i.i.d. samples from the distribution P. The main benefit is that now the sufficient statistics $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$ are independent, which significantly simplifies the analysis.

Analogous to the minimax risk (1), we define its counterpart under the Poisson sampling model:

$$\tilde{R}^*(k,n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}(\hat{H}(N) - H(P))^2, \tag{9}$$

where $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$ for $i = 1, \ldots, k$. In view of the exponential tail of Poisson distributions, the Poissonized sample size is concentrated near its mean n with high probability, which guarantees that the statistical performance as well as the minimax risk under Poisson sampling are provably close to that with fixed sample size. Indeed, the inequality

$$\tilde{R}^*(k,2n) - \exp(-n/4)\log^2 k \le R^*(k,n) \le 2\tilde{R}^*(k,n/2) \tag{10}$$

allows us to focus on the risk of the Poisson model.

III. Minimax Lower Bound

In this section we give converse results for entropy estimation and prove the lower bound part of Theorem 1. It suffices to show that the minimax risk is lower bounded by the two terms in (3) separately. This follows from combining Propositions 2 and 3 below.

Proposition 2. *For all $k, n \in \mathbb{N}$,*

$$R^*(k,n) \gtrsim \frac{\log^2 k}{n}. \tag{11}$$

Proposition 3. *If $n \ge \frac{ck}{\log k}$ for some $c > 0$, then*

$$R^*(k,n) \ge c' \left(\frac{k}{n \log k}\right)^2 \tag{12}$$

where c' only depends on c.

Proposition 2 follows from a simple application of Le Cam's *two-point method*: If two input distributions P and Q are sufficiently close such that it is impossible to reliably distinguish between them using n samples with error probability less than, say, $\frac{1}{2}$, then any estimator suffers a quadratic risk proportional to the separation of the functional values $|H(P) - H(Q)|^2$.

The remainder of this section is devoted to illustrating the broad strokes for proving Proposition 3. Since it can be shown that the best lower bound provided by the two-point method is $\frac{\log^2 k}{n}$, proving (12) requires more powerful techniques. To this end, we use a generalized version of Le Cam's method involving two *composite* hypotheses (also known as fuzzy hypothesis testing in [20]):

$$H_0 : H(P) \le t \quad \text{versus} \quad H_1 : H(P) \ge t + d, \tag{13}$$

which is more general than the two-point argument using only simple hypothesis testing. Similarly, if we can establish that no test can distinguish (13) reliably, then we obtain a lower bound for the quadratic risk on the order of d^2. By the minimax theorem, the optimal probability of error for the composite hypotheses test is given by the Bayesian version with respect to the least favorable prior. For (13) we need to choose a pair of priors, which, in this case, are distributions on the probability simplex $\mathcal{M}_k$, is to ensure the entropy values are separated.

A. *Construction of the priors*

The main idea for constructing the priors is as follows: First of all, the symmetry of the entropy functional implies that the least favorable prior must be permutation-invariant. This inspires us to use the following *iid construction*. For conciseness, we focus on the case of $n \asymp \frac{k}{\log k}$ for now and our goal is an $\Omega(1)$ lower bound. Let U be a $\mathbb{R}_+$-valued random variable with unit mean. Denote the random vector $\mathsf{P} = \frac{1}{k}(U_1, \ldots, U_k)$, consisting of iid copies of U. Note that P itself is *not* a probability distribution; however, the key observation is that, since $\mathbb{E}[U] = 1$, the law of large numbers implies P is *approximately* a probability distribution. Use some soft arguments we can show that the distribution of P can effectively serve as a prior.

Next we outline the main ingredients in Le Cam's method:

1) *Functional value separation*: Define $\phi(x) \triangleq x \log \frac{1}{x}$. Note that
$$H(\mathsf{P}) = \sum_{i=1}^{k} \phi\left(\frac{U_i}{k}\right) = \frac{1}{k}\sum_{i=1}^{k} \phi(U_i) + \frac{\log k}{k}\sum_{i=1}^{k} U_i,$$
which also concentrates near its mean $\mathbb{E}[H(\mathsf{P})] = \mathbb{E}[\phi(U)] + \mathbb{E}[U]\log k$. Therefore, given another random variable U' with unit mean, we can obtain P' similarly using iid copies of U'. Then with high probability, $H(\mathsf{P})$ and $H(\mathsf{P}')$ are separated by the difference in the respective means
$$\mathbb{E}[H(\mathsf{P})] - \mathbb{E}[H(\mathsf{P}')] = \mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')],$$
which we want to maximize.
2) *Indistinguishably*: Note that given P, the sufficient statistics satisfy $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Therefore, if P is drawn from the distribution of P, then $N = (N_1, \ldots, N_k)$ are iid distributed according the *Poisson mixture* $\mathbb{E}[\text{Poi}(\frac{n}{k}U)]$. Similarly, if P is drawn from the prior of P', then N is distributed according to $(\mathbb{E}[\text{Poi}(\frac{n}{k}U')])^{\otimes k}$. To establish the impossibility of testing, we need the total variation distance between the two k-product distributions to strictly bounded away from one, for which a sufficient condition is
$$\mathsf{TV}(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) \leq c/k \tag{14}$$
for some small c.

To conclude, we see that the iid construction fully exploits the independence blessed by the Poisson sampling, thereby reducing the problem to *one dimension*. This allows us to sidestep the difficulty encountered in [10] when dealing with fingerprints which are high-dimensional random vectors with dependent entries.

What remains is the following scalar problem: choose U, U' to maximize $|\mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')]|$ subject to the constraint (14). A commonly used proxy for bounding the total variation distance is *moment matching*, i.e., $\mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right]$ for all $j = 1, \ldots, L$. Together with some L_∞-norm constraints, a sufficient large L ensures the total variation bound (14). Combining the above steps, our lower bound is proportional to the value of the following convex optimization problem (in fact, infinite-dimensional linear programming):
$$\begin{aligned} \mathcal{F}_L(\lambda) \triangleq \sup\ & \mathbb{E}[\phi(U)] - \mathbb{E}[\phi(U')] \\ \text{s.t. } & \mathbb{E}[U] = \mathbb{E}[U'] = 1 \\ & \mathbb{E}\left[U^j\right] = \mathbb{E}\left[U'^j\right], \quad j = 1, \ldots, L, \\ & U, U' \in [0, \lambda] \end{aligned} \tag{15}$$
for some appropriately chosen $L \in \mathbb{N}$ and $\lambda > 1$ depending on n and k.

Finally, we connect the optimization problem (15) to the machinery of *best polynomial approximation*: Denote by $\mathcal{P}_L$ the set of polynomials of degree L and
$$E_L(f, I) \triangleq \inf_{p \in \mathcal{P}_L} \sup_{x \in I} |f(x) - p(x)|, \tag{16}$$
which is the best uniform approximation error of a function f over a finite interval I by polynomials of degree L. We prove that
$$\mathcal{F}_L(\lambda) \geq 2E_L(\log, [1/\lambda, 1]). \tag{17}$$
Due to the singularity of the logarithm at zero, the approximation error can be made bounded away from zero if λ grows *quadratically* with the degree L. Choosing $L \asymp \log k$ and $\lambda \asymp \log^2 k$ leads to the lower bound of $n \gtrsim \frac{k}{\log k}$ for consistent estimation. For $n \gg \frac{k}{\log k}$, the lower bound for the quadratic risk follows from relaxing the unit-mean constraint in (15) to $\mathbb{E}[U] = \mathbb{E}[U'] \leq 1$ and a simple scaling argument. We remark that analogous construction of priors and proof techniques have subsequently been used in [18] to obtain sharp minimax lower bound for estimating the power sum in which case the $\log p$ function is replaced by p^α.

IV. Optimal estimator via best polynomial approximation

As observed in various previous results as well as suggested by the minimax lower bound in Section III, the major difficulty of entropy estimation lies in the bias due to insufficient samples. Inspired by the observation in Section I-B that polynomials admit unbiased estimators, our main idea is to approximate ϕ by a polynomial of degree L, say g_L, for which we pay a price in bias at most the uniform approximation error. While the approximation error clearly decreases with the degree L, it is not unexpected that the variance of the unbiased estimator for $g_L(p_i)$ is increasing in both L and p_i. Therefore we only apply the polynomial approximation scheme to small p_i and directly use the plug-in estimator for large p_i, since the signal-to-noise ratio is sufficiently large.

Next we describe the estimator in detail. In view of the relationship (10) between the risks with fixed and Poisson sample size, we shall assume the Poisson sampling model to simplify the analysis. We split the samples equally and use the first half for selecting to use either the polynomial estimator or the plug-in estimator and the second half for estimation. Specifically, we draw two sets of samples independently, of which each has $\text{Poi}(n)$ samples. Count the samples in each set separately to obtain corresponding N, N'. Then N and N' are independent, where $N_i, N_i' \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}\,(np_i)$.

Let $c_0, c_1, c_2 > 0$ be constants to be specified. Let $L = \lfloor c_0 \log k \rfloor$. Denote the best polynomial of degree L to uniformly approximate ϕ on $[0,1]$ is $p_L(x) = \sum_{m=0}^{L} a_m x^m$. Through a change of variables, we see that the best polynomial of degree L to approximate ϕ on $[0, \frac{c_1 \log k}{n}]$ is

$$P_L(x) \triangleq \sum_{m=0}^{L} \frac{a_m n^{m-1}}{(c_1 \log k)^{m-1}} x^m + \left(\log \frac{n}{c_1 \log k} \right) x.$$

Define the factorial moment by $(x)_m \triangleq \frac{x!}{(x-m)!}$, which gives an unbiased estimator for the monomials of the Poisson mean: $\mathbb{E}[(X)_m] = \lambda^m$ where $X \sim \mathrm{Poi}(\lambda)$. Consequently, the following polynomial of degree L

$$g_L(N_i) \triangleq \frac{1}{n} \sum_{m=0}^{L} \frac{a_m}{(c_1 \log k)^{m-1}} (N_i)_m + \left(\log \frac{n}{c_1 \log k} \right) N_i$$

is an unbiased estimator for $P_L(p_i)$.

Define a bias-corrected plug-in estimator for ϕ by

$$\phi_0(N_i) = \phi\left(\frac{N_i}{n}\right) + \frac{1}{2n}. \tag{18}$$

Define a preliminary estimator of entropy by

$$\tilde{H} \triangleq \sum_{i=1}^{k} \left(g_L(N_i) \mathbf{1}_{\left\{N_i' \le c_2 \log k\right\}} + \phi_0(N_i) \mathbf{1}_{\left\{N_i' > c_2 \log k\right\}} \right), \tag{19}$$

where we apply the estimator from polynomial approximation if $N_i' \le c_2 \log k$ or the bias-corrected plug-in estimator otherwise (c.f. the asymptotic expansion (7) of the bias under the original sampling model). In view of the fact that $0 \le H(P) \le \log k$ for any distribution P with alphabet size k, we define our final estimator by:

$$\hat{H} = (\tilde{H} \vee 0) \wedge \log k,$$

Since (19) can be expressed in terms of a linear combination of the fingerprints (8) of the second sample and the coefficients can be pre-computed using fast best polynomial approximation algorithms (e.g., the Remez algorithm), it is clear that the estimator $\hat{H}$ can be computed in linear time in n.

The next result gives an upper bound on the above estimator under the Poisson sampling model, which, in view of the right inequality in (10) and Proposition 1, implies the upper bound on the minimax risk $R^*(n,k)$ in Theorem 1.

Proposition 4. *Assume that $\log n \le C \log k$ for some constant $C > 0$. Then there exists c_0, c_1, c_2 depending on C only, such that*

$$\sup_{P \in \mathcal{M}_k} \mathbb{E}[(H(P) - \hat{H}(N))^2] \lesssim \left(\frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n},$$

where $N = (N_1, \ldots, N_k) \overset{ind}{\sim} \mathrm{Poi}(np_i)$.

Remark 1. The estimator (19) uses the polynomial approximation of $x \log \frac{1}{x}$ for those masses below $\frac{\log k}{n}$ and the bias-corrected plug-in (18) otherwise. In view of the fact that the lower bound in Proposition 3 is based on a pair of randomized distributions whose masses are below $\frac{\log k}{n}$ (except for possibly a fixed large mass at the last element), this suggests that the main difficulty of the estimation tasks lies in those p_i's in the interval $[0, \frac{\log k}{n}]$, which are individually small but collectively contribute significantly to the entropy.

Remark 2. The estimator in (19) depends on the alphabet size k. In order to obtain an optimal adaptive estiamtor, simply replace all $\log k$ by $\log n$ and the optimal rate in Proposition 4 continues to hold in the regime of $\frac{k}{\log k} \lesssim n \lesssim \frac{k^2}{\log^2 k}$ where the plug-in estimator fails to attain the minimax rate.

References

[1] 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*, 2011, pp. 685–694.

[2] F. Rieke, W. Bialek, D. Warland, and R. d. R. van Steveninck, "Spikes: Exploring the neural code," 1999.

[3] A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti, "Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1282–1291, 2001.

[4] 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.

[5] G. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory of Probability & Its Applications*, vol. 4, no. 3, pp. 333–336, 1959.

[6] G. A. Miller, "Note on the bias of information estimates," *Information theory in psychology: Problems and methods*, vol. 2, pp. 95–100, 1955.

[7] B. Harris, "The statistical estimation of entropy in the non-parametric case," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Springer Netherlands, 1975, vol. 16, pp. 323–355.

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

[9] P. Valiant, "Testing symmetric properties of distributions," in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, ser. STOC '08, 2008, pp. 383–392.

[10] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy," in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, 2010, p. 179.

[11] 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.

[12] 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.

[13] L. Le Cam, *Asymptotic methods in statistical decision theory*. New York, NY: Springer-Verlag, 1986.

[14] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probability theory and related fields*, vol. 113, no. 2, pp. 221–253, 1999.

[15] T. Cai and M. G. Low, "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional," *The Annals of Statistics*, vol. 39, no. 2, pp. 1012–1041, 2011.

[16] I. Ibragimov, A. Nemirovskii, and R. Khas'minskii, "Some problems on nonparametric estimation in gaussian white noise," *Theory of Probability & Its Applications*, vol. 31, no. 3, pp. 391–406, 1987.

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

[18] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *arXiv:1406.6956v5*, 2014.

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

[20] A. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY: Springer Verlag, 2009.