Estimating Renyi Entropy of Discrete Distributions

Jayadev Acharya, *Member, IEEE*, Alon Orlitsky, *Fellow, IEEE*, Ananda Theertha Suresh, Himanshu Tyagi, *Member, IEEE*

***Abstract*—It was shown recently that estimating the Shannon entropy $H(\mathrm{p})$ of a discrete k-symbol distribution p requires $\Theta(k/\log k)$ samples, a number that grows near-linearly in the support size. In many applications $H(\mathrm{p})$ can be replaced by the more general Rényi entropy of order α, $H_\alpha(\mathrm{p})$. We determine the number of samples needed to estimate $H_\alpha(\mathrm{p})$ for all α, showing that $\alpha < 1$ requires a super-linear, roughly $k^{1/\alpha}$ samples, noninteger $\alpha > 1$ requires a near-linear k samples, but, perhaps surprisingly, integer $\alpha > 1$ requires only $\Theta(k^{1-1/\alpha})$ samples. Furthermore, developing on a recently established connection between polynomial approximation and estimation of additive functions of the form $\sum_x f(\mathrm{p}_x)$, we reduce the sample complexity for non-integer values of α by a factor of $\log k$ compared to the empirical estimator. The estimators achieving these bounds are simple and run in time linear in the number of samples. Our lower bounds provide explicit constructions of distributions with different Rényi entropies that are hard to distinguish.**

I. Introduction

A. Shannon and Rényi entropies

One of the most commonly used measure of randomness of a distribution p over a discrete set $\mathcal{X}$ is its *Shannon entropy*

$$H(\mathrm{p}) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} \mathrm{p}_x \log \frac{1}{\mathrm{p}_x}.$$

The estimation of Shannon entropy has several applications, including measuring genetic diversity [37], quantifying neural activity [32], [29], network anomaly detection [20], and others. It was recently shown that estimating the Shannon entropy of a discrete distribution p over k elements to a given additive accuracy requires $\Theta(k/\log k)$ independent samples from p [33], [41]; see [16], [43] for subsequent extensions. This number of samples grows near-linearly with the alphabet size and is only a logarithmic factor smaller than the $\Theta(k)$ samples needed to learn p itself to within a small total variation distance.

A popular generalization of Shannon entropy is the *Rényi entropy* of order $\alpha \geq 0$, defined for $\alpha \neq 1$ by

$$H_\alpha(\mathrm{p}) \stackrel{\text{def}}{=} \frac{1}{1-\alpha} \log \sum_{x\in\mathcal{X}} \mathrm{p}_x^\alpha$$

and for $\alpha = 1$ by

$$H_1(\mathrm{p}) \stackrel{\text{def}}{=} \lim_{\alpha\to 1} H_\alpha(\mathrm{p}).$$

It was shown in the seminal paper [36] that Rényi entropy of order 1 is Shannon entropy, namely $H_1(\mathrm{p}) = H(\mathrm{p})$, and for all other orders it is the unique extension of Shannon entropy when of the four requirements in Shannon entropy's axiomatic definition, continuity, symmetry, and normalization are kept but grouping is restricted to only additivity over independent random variables (*cf*. [13]).

Rényi entropy too has many applications. It is often used as a bound on Shannon entropy [26], [29], [12], and in many applications it replaces Shannon entropy as a measure of randomness [7], [24], [3]. It is also of interest in its own right, with diverse applications to unsupervised learning [44], [15], source adaptation [22], image registration [21], [28], and password guess-ability [3], [35], [10] among others. In particular, the Rényi entropy of order 2, $H_2(\mathrm{p})$, measures the quality of random number generators [19], [30], determines the number of unbiased bits that can be extracted from a physical source of randomness [14], [6], helps test graph expansion [8] and closeness of distributions [5], [34], and characterizes the number of reads needed to reconstruct a DNA sequence [27].

Motivated by these and other applications, unbiased and heuristic estimators of Rényi entropy have been studied in the physics literature following [9], and asymptotically consistent and normal estimates were proposed in [45], [18]. However, no systematic study of the complexity of estimating Rényi entropy is available. For example, it was hitherto unknown if the number of samples needed to estimate the Rényi entropy of a given order α differs from that required for Shannon entropy, or whether it varies with the order α, or how it depends on the alphabet size k.

An initial version of this paper [1] was presented at the ACM Symposium on Discrete Algorithms (SODA), 2015.

J. Acharya is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (email: acharya@cornell.edu).

A. Orlitsky is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (email: alon@ucsd.edu).

A. T. Suresh is with Google Research, New York (email: theertha@google.com).

H. Tyagi is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (email: htyagi@ece.iisc.ernet.in).

Part of this research was done when J. Acharya was at the EECS department at the Massachusetts Institute of Technology, supported by an MITEI-Shell grant, and A. T. Suresh was a student at the University of California, San Diego.

B. Definitions and results

We answer these questions by showing that the number of samples needed to estimate $H_\alpha(\mathrm{p})$ falls into three different ranges. For $\alpha < 1$ it grows super-linearly with k, for $1 < \alpha \notin \mathbb{N}$ it grows almost linearly with k, and most interestingly, for the popular orders $1 < \alpha \in \mathbb{N}$ it grows as $\Theta(k^{1-1/\alpha})$, which is much less than the sample complexity of estimating Shannon entropy.

To state the results more precisely we need a few definitions. A Rényi-entropy *estimator* for distributions over support set $\mathcal{X}$ is a function $f : \mathcal{X}^* \to \mathbb{R}$ mapping a sequence of samples drawn from a distribution to an estimate of its entropy. The sample complexity of an estimator f for distributions over k elements is defined as

$$S_\alpha^f(k,\delta,\epsilon) \stackrel{\text{def}}{=} \min_n \{n : \mathrm{p}\{|H_\alpha(\mathrm{p}) - f(X^n)| > \delta\} < \epsilon, \quad \forall \mathrm{p} \text{ with } \|\mathrm{p}\|_0 \le k\},$$

i.e., the minimum number of samples required by f to estimate $H_\alpha(\mathrm{p})$ of any k-symbol distribution p to a given additive accuracy δ with probability greater than $1-\epsilon$. The *sample complexity* of estimating $H_\alpha(\mathrm{p})$ is then

$$S_\alpha(k,\delta,\epsilon) \stackrel{\text{def}}{=} \min_f S_\alpha^f(k,\delta,\epsilon),$$

the least number of samples any estimator needs to estimate $H_\alpha(\mathrm{p})$ for all k-symbol distributions p, to an additive accuracy δ and with probability greater than $1-\epsilon$. This is a min-max definition where the goal is to obtain the *best* estimator for the *worst* distribution.

The desired accuracy δ and confidence $1-\epsilon$ are typically fixed. We are therefore most interested[1] in the dependence of $S_\alpha(k,\delta,\epsilon)$ on the alphabet size k and omit the dependence of $S_\alpha(k,\delta,\epsilon)$ on δ and ϵ to write $S_\alpha(k)$. In particular, we are interested in the *large alphabet* regime and focus on the essential growth rate of $S_\alpha(k)$ as a function of k for large k with fixed δ and ϵ. Using the standard asymptotic notations, let $S_\alpha(k) = O(k^\beta)$ indicate that for some constant c which may depend on α, δ, and ϵ, for all sufficiently large k, $S_\alpha(k,\delta,\epsilon) \le c \cdot k^\beta$. Similarly, $S_\alpha(k) = \Theta(k^\beta)$ adds the corresponding $\Omega(k^\beta)$ lower bound for $S_\alpha(k,\delta,\epsilon)$, for all sufficiently small δ and ϵ. Finally, extending the $\tilde{\Omega}$ notation[2], we let $S_\alpha(k) = \tilde{\tilde{\Omega}}(k^\beta)$ indicate that for every sufficiently small ϵ and arbitrary $\eta > 0$, there exist c and δ depending on η such that for all k sufficiently large $S_\alpha(k,\delta,\epsilon) > ck^{\beta-\eta}$, namely $S_\alpha(k)$ grows polynomially in k with exponent not less than $\beta - \eta$ for $\delta \le \delta_\eta$.

We show that $S_\alpha(k)$ behaves differently in three ranges of α. For $0 \le \alpha < 1$,

$$\tilde{\tilde{\Omega}}\left(k^{1/\alpha}\right) \le S_\alpha(k) \le O\left(\frac{k^{1/\alpha}}{\log k}\right),$$

namely the sample complexity grows super-linearly in k and estimating the Rényi entropy of these orders is even more difficult than estimating the Shannon entropy. In fact, the upper bound follows from a corresponding result on estimation of power sums considered in [16], [43] which uses the best polynomial approximation based estimator (see Section III-C for further discussion) and is proved in Theorem 13; the lower bound is proved in Theorem 22.

For $1 < \alpha \notin \mathbb{N}$,

$$\tilde{\tilde{\Omega}}(k) \le S_\alpha(k) \le O\left(\frac{k}{\log k}\right),$$

namely as with Shannon entropy, the sample complexity grows roughly linearly in the alphabet size. Once again, the upper bound uses the aforementioned polynomial approximation estimator of [16], [43] and is proved in Theorem 12; the lower bound is proved in Theorem 21.

For $1 < \alpha \in \mathbb{N}$,

$$S_\alpha(k,\delta,\epsilon) = \Theta\left(k^{1-1/\alpha}\right),$$

and in particular, the sample complexity is *strictly sub-linear* in the alphabet size. The upper and lower bounds are shown in Theorems 11 and 15, respectively. Unlike the previous two cases, the upper bound for integer $\alpha > 1$ is attained by a simple bias-corrected version of the empirical estimator. Figure 1 illustrates our results for different ranges of α.

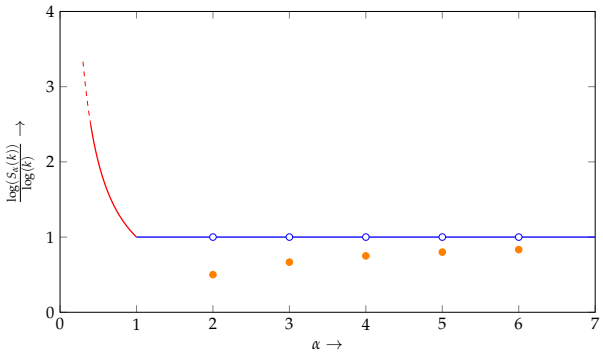

Fig. 1: Exponent of k in $S_\alpha(k)$ as a function of α.

Of the three ranges, the most frequently used, and coincidentally the one for which the results are most surprising, is the last with $\alpha = 2, 3, \ldots$. Some elaboration is in order.

First, for all integral $\alpha > 1$, $H_\alpha(\mathrm{p})$ can be estimated with a sublinear number of samples. The most commonly used Rényi entropy, $H_2(\mathrm{p})$, can be estimated within δ using just $\Theta\left(\sqrt{k}\right)$ samples, and hence Rényi entropy can be estimated much more efficiently than Shannon Entropy, a useful property for large-alphabet applications such as language processing genetic analysis.

Also, note that Rényi entropy is continuous in the order α. Yet the sample complexity is discontinuous

[1] Whenever a more refined result indicating the dependence of sample complexity on both k and δ is available, we shall use the more elaborate $S_\alpha(k,\delta,\epsilon)$ notation.

[2] The notations $\tilde{O}$, $\tilde{\Omega}$, and $\tilde{\Theta}$ hide poly-logarithmic factors.

at integer orders. While this makes the estimation of the popular integer-order entropies easier, it may seem contradictory. For instance, to approximate $H_{2.001}(\mathrm{p})$ one could approximate $H_2(\mathrm{p})$ using significantly fewer samples. The reason for this is that the Rényi entropy, while continuous in α, is not uniformly continuous. In fact, as shown in Example 2, the difference between say $H_2(\mathrm{p})$ and $H_{2.001}(\mathrm{p})$ may increase to infinity when the alphabet-size increases.

While the bounds for sample complexity described above capture the essence of our results, our complete results are more elaborate. For the case of integer $\alpha > 1$, we provide a complete characterization of $S_\alpha(k, \delta, \epsilon)$ including the precise dependence on k as well as δ for every k greater than a constant and δ smaller than a constant. For noninteger α, our upper bound reflects the dependence of $S_\alpha(k, \delta, \epsilon)$ on both k and δ, but holds only in the large alphabet regime when k is sufficiently large for a fixed δ and ϵ. The exact characterization of sample complexity for every k and δ for noninteger α remains open.

In the conference version [1] of this paper, weaker upper bounds for the case of noninteger α were obtained using the empirical estimator which simply *plugs-in* the normalized empirical-frequency in the formula for entropy. In this version, we provide a complete characterization of the sample complexity of the empirical estimator for every k greater than a constant and δ smaller than a constant. In particular, we show that the empirical estimator requires strictly more samples than the polynomial approximation estimator, in general.

It should also be noted that the estimators achieving the upper bounds in this paper are simple and run in time linear in the number of samples. Furthermore, the estimators are *universal* in that they do not require the knowledge of k. On the other hand, the lower bounds on $S_\alpha(k)$ hold even if the estimator knows k.

C. The estimators

The *power sum* of order α of a distribution p over $\mathcal{X}$ is

$$P_\alpha(\mathrm{p}) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} \mathrm{p}_x^\alpha,$$

and is related to the Rényi entropy for $\alpha \neq 1$ via

$$H_\alpha(\mathrm{p}) = \frac{1}{1-\alpha} \log P_\alpha(\mathrm{p}).$$

Hence estimating $H_\alpha(\mathrm{p})$ to an additive accuracy of $\pm\delta$ is equivalent to estimating $P_\alpha(\mathrm{p})$ to a multiplicative accuracy of $2^{\pm\delta\cdot(1-\alpha)}$. Furthermore, if $\delta(\alpha-1) \leq 1/2$ then estimating $P_\alpha(\mathrm{p})$ to multiplicative accuracy of $1 \pm \delta(1-\alpha)/2$ ensures a $\pm\delta$ additive accurate estimate of $H_\alpha(\mathrm{p})$.

We construct estimators for the power-sums of distributions with a multiplicative-accuracy of $(1 \pm \delta)$ and hence obtain an additive-accuracy of $\Theta(\delta)$ for Rényi entropy estimation. We consider the following three different estimators for different ranges of α and with different performance guarantees.

a) Empirical estimator: The *empirical*, or *plug-in*, estimator of $P_\alpha(\mathrm{p})$ is given by

$$\widehat{P}_\alpha^{\mathrm{e}} \stackrel{\text{def}}{=} \sum_x \left(\frac{N_x}{n}\right)^\alpha. \tag{1}$$

For $\alpha \neq 1$, $\widehat{P}_\alpha^{\mathrm{e}}$ is a not an unbiased estimator of $P_\alpha(\mathrm{p})$. We show in Corollary 27 that for $\alpha < 1$ the sample complexity of the empirical estimator is $\Theta\left(\max\left\{\left(\frac{k}{\delta}\right)^{\frac{1}{\alpha}}, \frac{k^{\frac{1-\alpha}{\alpha}}}{\delta^2}\right\}\right)$ and in Corollary 25 that for $\alpha > 1$ it is $\Theta\left(\max\left\{\frac{k}{\delta}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right\}\right)$.

b) Bias-corrected estimator: For integral $\alpha > 1$, the *bias-corrected* estimator for $P_\alpha(\mathrm{p})$ is

$$\widehat{P}_\alpha^{\mathrm{u}} \stackrel{\text{def}}{=} \sum_x \frac{N_x^{\underline{\alpha}}}{n^\alpha}, \tag{2}$$

where for integers N and $r > 0$, $N^{\underline{r}} \stackrel{\text{def}}{=} N(N-1)\ldots(N-r+1)$. A variation of this estimator was proposed first in [4] for estimating moments of frequencies in a sequence using random samples drawn from it. Corollary 16 shows that for $1 < \alpha \in \mathbb{N}$, $\widehat{P}_\alpha^{\mathrm{u}}$ estimates $P_\alpha(\mathrm{p})$ within a factor of $1 \pm \delta$ using $\Theta\left(\frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right)$ samples.

c) Polynomial approximation estimator: To obtain a logarithmic improvement in $S_\alpha(k)$, we consider the polynomial approximation estimator proposed in [43], [16] for different problems, concurrently to a conference version [1] of this paper. The polynomial approximation estimator first considers the *best polynomial approximation* of degree d to y^α for the interval $y \in [0, 1]$ [39]. Suppose this polynomial is given by $a_0 + a_1 y + a_2 y^2 + \ldots + a_d y^d$. We roughly divide the samples into two parts. Let N'_x and N_x be the multiplicities of x in the first and second parts respectively. The polynomial approximation estimator uses the empirical estimate of p_x^α for large N'_x, but estimates a polynomial approximation of p_x^α for a small N'_x; the integer powers of p_x in the latter in turn is estimated using the bias-corrected estimator.

The estimator is roughly of the form

$$\widehat{P}_\alpha^{d,\tau} \stackrel{\text{def}}{=} \sum_{x: N'_x \leq \tau} \left(\sum_{m=0}^{d} \frac{a_m (2\tau)^{\alpha-m} N_x^{\underline{m}}}{n^\alpha}\right) + \sum_{x: N'_x > \tau} \frac{N_x^{\underline{\alpha}}}{n^\alpha}, \tag{3}$$

where d and τ are both $O(\log n)$ and chosen appropriately. Theorem 12 and Theorem 13 show that for $\alpha > 1$ and $\alpha < 1$, respectively, the sample complexity of $\widehat{P}_\alpha^{d,\tau}$ is $O(k/\delta^{1/\alpha} \log k)$ and $O(k^{\frac{1}{\alpha}}/\delta^{1/\alpha} \log k)$, when k is sufficiently large (depending on δ), resulting in a reduction in sample complexity of $O(\log k)$ over the empirical estimator.

Note that while the results described above characterize $S_\alpha(k, \delta, \epsilon)$ for every k and δ only for the case of integer $\alpha > 1$, and the general problem of characterizing $S_\alpha(k, \delta, \epsilon)$ remains open, we identify the exponent of k

in $S_\alpha(k)$ for every α, *i.e.*, we can characterize the limit

$$E_\alpha = \lim_{k\to\infty} \frac{\log S_\alpha(k,\delta,\epsilon)}{\log k}$$

for every fixed δ and ϵ and show that

$$E_\alpha = \begin{cases} \frac{1}{\alpha}, & 0 < \alpha < 1, \\ 1, & 1 < \alpha \notin \mathbb{N}, \\ 1 - \frac{1}{\alpha}, & 1 < \alpha \in \mathbb{N}. \end{cases}$$

Furthermore, the empirical estimator attains the optimal exponent for $\alpha \notin \mathbb{N}$, but has a suboptimal exponent for $1 < \alpha \in \mathbb{N}$. In this latter regime, the bias-corrected estimator attains the optimal exponent. While the exponent captures a very coarse-level behavior of the sample complexity $S_\alpha(k,\delta,\epsilon)$, it is an important indicator of the behavior in the large alphabet regime. In fact, we provide a complete characterization of the dependence of sample complexity of the empirical estimator on k and δ for every $\alpha \neq 1$ and that of the bias-corrected estimator for integer $\alpha > 1$. results. Our results are summarized in Table I. Note that the $\overset{\approx}{\Omega}$ form of our general lower bounds is interesting for the large alphabet case, when k is sufficiently large for a fixed δ. In Theorem 15, a simple alternative lower bound is established for $\alpha > 1$ showing that $S_\alpha(k,\delta,\epsilon) \geq \Omega\left(\frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right)$. A similar lower bound for $\alpha < 1$ showing $S_\alpha(k,\delta,\epsilon) \geq \Omega\left(\frac{k^{\frac{1-\alpha}{\alpha}}}{\delta^2}\right)$ is established in Theorem 26. When $\alpha > 1$ and $k \leq \delta^{-\alpha}$ or when $1/2 \leq \alpha < 1$ and $k \leq \delta^{\frac{1-2\alpha}{\alpha}}$ these bounds are tight, and show that the empirical estimator attains the optimal sample complexity up to constant factors. However, the polynomial approximation estimator strictly outperforms the empirical estimator in the large alphabet regime; the latter does not even attain the optimal exponent in the dependence of sample complexity on k for integer $\alpha > 1$. We have only obtained the results for the polynomial approximation in this large alphabet regime, and the problem of characterizing the exact sample complexity of polynomial approximation estimator remains open, along with that of characterizing the exact sample complexity of estimating $H_\alpha(\mathrm{p})$ for noninteger $\alpha > 0$.

D. Organization

The rest of the paper is organized as follows. Section II presents basic properties of power sums of distributions and moments of Poisson random variables, which may be of independent interest. Upper bounds for sample complexity of our proposed estimators are given in Section III, and examples and simulation of the proposed estimators are given in Section IV. Section V contains general lower bounds for the sample complexity of estimating Rényi entropy and lower bounds for the sample complexity of the empirical estimator. Furthermore, in the Appendix we analyze the performance of the empirical estimator for power-sum estimation with an additive-accuracy.

II. Technical preliminaries

A. Bounds on power sums

Consider a distribution p over $[k] = \{1,\dots,k\}$. Since Rényi entropy is a measure of randomness (see [36] for a detailed discussion), it is maximized by the uniform distribution and the following inequalities hold:

$$0 \leq H_\alpha(\mathrm{p}) \leq \log k, \quad \alpha \neq 1,$$

or equivalently

$$1 \leq P_\alpha(\mathrm{p}) \leq k^{1-\alpha}, \quad \alpha < 1 \quad \text{, and} \tag{4}$$
$$k^{1-\alpha} \leq P_\alpha(\mathrm{p}) \leq 1, \quad \alpha > 1. \tag{5}$$

Furthermore, for $\alpha > 1$, $P_{\alpha+\beta}(\mathrm{p})$ and $P_{\alpha-\beta}(\mathrm{p})$ can be bounded in terms of $P_\alpha(\mathrm{p})$, using the monotonicity of norms and of Hölder means (see, for instance, [11]).

Lemma 1. (i) *For every $\alpha \geq 0$ and $\beta \geq 0$,*

$$P_{\alpha+\beta}(\mathrm{p}) \leq P_\alpha(\mathrm{p})^{\frac{\alpha+\beta}{\alpha}}.$$

(ii) *For every $\alpha \geq 0$,*

$$P_{2\alpha}(\mathrm{p}) \leq P_\alpha(\mathrm{p})^2.$$

(iii) *For $\alpha > 0$ and $0 \leq \beta \leq \alpha$,*

$$P_{\alpha-\beta}(\mathrm{p}) \leq k^{\frac{\beta}{\alpha}} P_\alpha(\mathrm{p})^{\frac{\alpha-\beta}{\alpha}}.$$

(iv) *For $\alpha \geq 1$ and $0 \leq \beta \leq \alpha$,*

$$P_{\alpha+\beta}(\mathrm{p}) \leq k^{(\alpha-1)(\alpha-\beta)/\alpha}\, P_\alpha(\mathrm{p})^2,$$

and

$$P_{\alpha-\beta}(\mathrm{p}) \leq k^\beta\, P_\alpha(\mathrm{p}).$$

Proof. (i) holds by the monotonicity of norms; (ii) follows upon choosing $\alpha = \beta$. For (iii) note that by the monotonicity of Hölder means

$$\left(\frac{1}{k}\sum_x \mathrm{p}_x^{\alpha-\beta}\right)^{\frac{1}{\alpha-\beta}} \leq \left(\frac{1}{k}\sum_x \mathrm{p}_x^{\alpha}\right)^{\frac{1}{\alpha}},$$

which yields (iii) by rearranging the terms. Property (iv) is obtained by (i) and (iii) together with (5). ∎

B. Bounds on moments of a Poisson random variable

Let $\mathrm{Poi}(\lambda)$ be the Poisson distribution with parameter λ. We consider Poisson sampling where $N \sim \mathrm{Poi}(n)$ samples are drawn from the distribution p and the multiplicities used in the estimation are based on the sequence $X^N = X_1, ..., X_N$ instead of X^n. Under Poisson sampling, the multiplicities N_x are distributed as $\mathrm{Poi}(n\mathrm{p}_x)$ and are all independent, leading to simpler analysis. To facilitate our analysis under Poisson sampling, we note a few properties of the moments of a Poisson random variable.

We start with the expected value and the variance of falling powers of a Poisson random variable.

Lemma 2. *Let $X \sim \mathrm{Poi}(\lambda)$. Then, for all $r \in \mathbb{N}$*

$$\mathbb{E}[X^{\underline{r}}] = \lambda^r$$

Range of α	Empirical	Bias-corrected	Polynomial	Lower bounds
$0 < \alpha < 1$	$\Theta\left(\max\left\{\left(\frac{k}{\delta}\right)^{\frac{1}{\alpha}}, \frac{k^{\frac{1-\alpha}{\alpha}}}{\delta^2}\right\}\right)$		$O\left(\frac{k^{1/\alpha}}{\delta^{1/\alpha}\log k}\right)$	$\widetilde{\widetilde{\Omega}}\left(k^{\frac{1}{\alpha}}\right)$
$\alpha > 1$, $\alpha \notin \mathbb{N}$	$\Theta\left(\max\left\{\frac{k}{\delta}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right\}\right)$		$O\left(\frac{k}{\delta^{1/\alpha}\log k}\right)$	$\widetilde{\widetilde{\Omega}}\left(k\right)$
$\alpha > 1$, $\alpha \in \mathbb{N}$	$\Theta\left(\max\left\{\frac{k}{\delta}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right\}\right)$	$O\left(\frac{k^{1-1/\alpha}}{\delta^2}\right)$		$\Omega\left(\frac{k^{1-1/\alpha}}{\delta^2}\right)$

TABLE I: Performance of estimators and general lower bounds for estimating Rényi entropy .

and

$$\mathbb{V}\mathrm{ar}[X^{\underline{r}}] \le \lambda^r\left((\lambda+r)^r - \lambda^r\right).$$

Proof. The expectation is

$$\begin{aligned}\mathbb{E}[X^{\underline{r}}] &= \sum_{i=0}^{\infty}\mathrm{Poi}(\lambda,i)\cdot i^{\underline{r}}\\ &= \sum_{i=r}^{\infty} e^{-\lambda}\cdot\frac{\lambda^i}{i!}\cdot\frac{i!}{(i-r)!}\\ &= \lambda^r\sum_{i=0}^{\infty} e^{-\lambda}\cdot\frac{\lambda^i}{i!}\\ &= \lambda^r.\end{aligned}$$

The variance satisfies

$$\begin{aligned}\mathbb{E}\left[(X^{\underline{r}})^2\right] &= \sum_{i=0}^{\infty}\mathrm{Poi}(\lambda,i)\cdot(i^{\underline{r}})^2\\ &= \sum_{i=r}^{\infty} e^{-\lambda}\cdot\frac{\lambda^i}{i!}\frac{i!^2}{(i-r)!^2}\\ &= \lambda^r\sum_{i=0}^{\infty} e^{-\lambda}\cdot\frac{\lambda^i}{i!}\cdot(i+r)^{\underline{r}}\\ &= \lambda^r\cdot\mathbb{E}[(X+r)^{\underline{r}}]\\ &\le \lambda^r\cdot\mathbb{E}\left[\sum_{j=0}^{r}\binom{r}{j}X^{\underline{j}}\cdot r^{r-j}\right]\\ &= \lambda^r\cdot\sum_{j=0}^{r}\binom{r}{j}\cdot\lambda^j\cdot r^{r-j}\\ &= \lambda^r(\lambda+r)^r,\end{aligned}$$

where the inequality follows from

$$(X+r)^{\underline{r}} = \prod_{j=1}^{r}[(X+1-j)+r] \le \sum_{j=0}^{r}\binom{r}{j}\cdot X^{\underline{j}}\cdot r^{r-j}.$$

Therefore,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X^{\underline{r}}] &= \mathbb{E}\left[(X^{\underline{r}})^2\right] - [\,\mathbb{E}\,X^{\underline{r}}]^2\\ &\le \lambda^r\cdot\left((\lambda+r)^r-\lambda^r\right).\end{aligned}$$

■

The next result establishes a bound on the moments of a Poisson random variable.

Lemma 3. *For $\beta > 0$ and $X \sim \mathrm{Poi}(\lambda)$, there exists a constant C_β depending only on β such that*

$$\mathbb{E}\left[X^\beta\right] \le C_\beta\max\{\lambda,\lambda^\beta\}.$$

Proof. For $\lambda \le 1$,

$$\begin{aligned}\mathbb{E}\left[X^\beta\right] = \sum_{i=1}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\cdot i^\beta &= \lambda\cdot\sum_{i=1}^{\infty} e^{-\lambda}\frac{\lambda^{i-1}}{i!}i^\beta\\ &\le \lambda\sum_{i=1}^{\infty}\frac{i^\beta}{i!},\end{aligned}$$

which proves the claim since the summation on the right-side is bounded.

For $\lambda > 1$, let $Z = \max\{\lambda^{1/\beta},\lambda\}$. Then,

$$\begin{aligned}\mathbb{E}\left[\frac{X^\beta}{Z^\beta}\right] &\le \mathbb{E}\left[\left(\frac{X}{Z}\right)^{\lceil\beta\rceil} + \left(\frac{X}{Z}\right)^{\lfloor\beta\rfloor}\right]\\ &= \frac{1}{Z^{\lceil\beta\rceil}}\sum_{i=1}^{\lceil\beta\rceil}\lambda^i\left\{{\lceil\beta\rceil \atop i}\right\} + \frac{1}{Z^{\lfloor\beta\rfloor}}\sum_{i=1}^{\lfloor\beta\rfloor}\lambda^i\left\{{\lfloor\beta\rfloor \atop i}\right\}\\ &\le \sum_{i=0}^{\lceil\beta\rceil}\left\{{\lceil\beta\rceil \atop i}\right\} + \sum_{i=0}^{\lfloor\beta\rfloor}\left\{{\lfloor\beta\rfloor \atop i}\right\},\end{aligned}$$

where $\left\{{m \atop i}\right\}$ denotes the Stirling number of the second kind. The first inequality follows upon considering the two cases $X \le Z$ and $X > Z$, the equality uses a well-known formula for integer moments of a Poisson random variable, and the second inequality holds since $\lambda > 1$ and $\lambda/Z \le 1$. Multiplying both sides by Z^β yields the bound[3]. ■

We close this section with a bound for $|\mathbb{E}[X^\alpha] - \lambda^\alpha|$, which will be used in the next section and is also of independent interest.

Lemma 4. *For $X \sim \mathrm{Poi}(\lambda)$, there exists a constant C_α depending only on α such that*

$$|\mathbb{E}[X^\alpha] - \lambda^\alpha| \le \begin{cases} C_\alpha\max\{\lambda,\lambda^\alpha\}, & \lambda < 1,\\ C_\alpha\lambda^{\alpha-1}, & \lambda \ge 1.\end{cases}$$

In particular,

$$|\mathbb{E}[X^\alpha] - \lambda^\alpha| \le \begin{cases} C_\alpha & \alpha < 1,\\ C_\alpha\left(1+\lambda^{\alpha-1}\right), & \alpha > 1.\end{cases}$$

Proof. For $\lambda < 1$, the claimed bound simply follows by Lemma 3 upon noting that

$$|\mathbb{E}[X^\alpha] - \lambda^\alpha| \le \mathbb{E}[X^\alpha] + \lambda^\alpha.$$

[3]All the constants in this proof can be shown to be less than $e + O(\beta^\beta)$.

Also, for $\alpha \le 1$, $(1+y)^\alpha \ge 1+\alpha y - y^2$ for all $y \in [-1,\infty]$. Hence,

$$\begin{aligned} X^\alpha &= \lambda^\alpha \left(1 + \left(\frac{X}{\lambda} - 1\right)\right)^\alpha \\ &\ge \lambda^\alpha \left(1 + \alpha\left(\frac{X}{\lambda} - 1\right) - \left(\frac{X}{\lambda} - 1\right)^2\right). \end{aligned}$$

Taking expectations on both sides,

$$\begin{aligned} \mathbb{E}[X^\alpha] &\ge \lambda^\alpha \left(1 + \alpha \mathbb{E}\left[\frac{X}{\lambda} - 1\right] - \mathbb{E}\left[\left(\frac{X}{\lambda} - 1\right)^2\right]\right) \\ &= \lambda^\alpha \left(1 - \frac{1}{\lambda}\right). \end{aligned}$$

Since x^α is a concave function and X is nonnegative, the previous bound yields

$$|\mathbb{E}[X^\alpha] - \lambda^\alpha| = \lambda^\alpha - \mathbb{E}[X^\alpha] \le \min\left\{\lambda^\alpha, \lambda^{\alpha-1}\right\},$$

which implies our claimed bound for $\alpha \le 1$.

It remains to establish the bound for the case $\alpha > 1$ and $\lambda \ge 1$. For this case, observe that

$$\left(\frac{X}{\lambda}\right)^\alpha = \left(1 + \frac{X-\lambda}{\lambda}\right)^\alpha \le e^{\frac{\alpha(X-\lambda)}{\lambda}}.$$

Taking expectation on both sides,

$$\mathbb{E}[X^\alpha] \le \lambda^\alpha e^{-\alpha} \mathbb{E}[e^{\alpha X/\lambda}] = \lambda^\alpha e^{-\alpha} e^{\lambda(e^{\alpha/\lambda}-1)}.$$

Furthermore by convexity, $\mathbb{E}[X^\alpha] \ge \lambda^\alpha$. Hence,

$$\begin{aligned} |\mathbb{E}[X^\alpha] - \lambda^\alpha| &= \mathbb{E}[X^\alpha] - \lambda^\alpha \\ &\le \lambda^\alpha \left[e^{-\alpha} e^{\lambda(e^{\alpha/\lambda}-1)} - 1\right] \\ &\le e^{e^\alpha} \lambda^{\alpha-1}, \end{aligned}$$

where the last inequality holds since for $\lambda \ge 1$

$$\lambda(e^{\alpha/\lambda} - 1) \le \alpha + (e^\alpha/\lambda),$$

and hence

$$\begin{aligned} \lambda\left(e^{-\alpha} e^{\lambda(e^{\alpha/\lambda}-1)} - 1\right) &\le \lambda\left(e^{\frac{e^\alpha}{\lambda}} - 1\right) \\ &\le e^\alpha + \frac{1}{\lambda}\left(e^{e^\alpha} - 1 - e^\alpha\right) \\ &\le e^{e^\alpha}. \end{aligned}$$

The weaker alternative form follows since for $\alpha < 1$ and $\lambda > 1$, $\lambda^{\alpha-1} \le 1$. ■

C. Polynomial approximation of x^α

In this section, we review a bound on the error in approximating x^α by a d-degree polynomial over a bounded interval. Let $\mathcal{P}_d$ denote the set of all polynomials of degree less than or equal to d over $\mathbb{R}$. For a continuous function $f(x)$ and $\lambda > 0$, let

$$E_d(f, [0,\lambda]) \stackrel{\text{def}}{=} \inf_{q \in \mathcal{P}_d} \max_{x \in [0,\lambda]} |q(x) - f(x)|.$$

Lemma 5 ([39]). *There is a constant c'_α such that for any $d > 0$,*

$$E_d(x^\alpha, [0,1]) \le \frac{c'_\alpha}{d^{2\alpha}}.$$

To obtain an estimator which does not require a knowledge of the support size k, we seek a polynomial approximation $q_\alpha(x)$ of x^α with $q_\alpha(0) = 0$. Such a polynomial $q_\alpha(x)$ can be obtained by a minor modification of the polynomial $q'_\alpha(x) = \sum_{j=0}^d q_j x^j$ satisfying the error bound in Lemma 5. Specifically, we use the polynomial $q_\alpha(x) = q'_\alpha(x) - q_0$ for which the approximation error is bounded as

$$\begin{aligned} \max_{x\in[0,1]} |q_\alpha(x) - x^\alpha| &\le |q_0| + \max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &= |q'_\alpha(0) - 0^\alpha| + \max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &\le 2 \max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &= \frac{2c'_\alpha}{d^{2\alpha}} \\ &\stackrel{\text{def}}{=} \frac{c_\alpha}{d^{2\alpha}}. \end{aligned} \tag{6}$$

To bound the variance of the proposed polynomial approximation estimator, we require a bound on the absolute values of the coefficients of $q_\alpha(x)$. The following inequality due to Markov serves this purpose.

Lemma 6 ([23]). *Let $p(x) = \sum_{j=0}^d c_j x^j$ be a degree-d polynomial so that $|p(x)| \le 1$ for all $x \in [-1,1]$. Then for all $j = 0, \ldots, m$*

$$\max_j |c_j| \le (\sqrt{2}+1)^d.$$

Since $|x^\alpha| \le 1$ for $x \in [0,1]$, the approximation bound (6) implies $|q_\alpha(x)| < 1 + \frac{c_\alpha}{d^{2\alpha}}$ for all $x \in [0,1]$. It follows from Lemma 6 that

$$\max_m |a_m| < \left(1 + \frac{c_\alpha}{d^{2\alpha}}\right)(\sqrt{2}+1)^d. \tag{7}$$

III. Upper bounds on sample complexity

In this section, we analyze the performances of the estimators we proposed in Section I-C. Our proofs are based on bounding the bias and the variance of the estimators under Poisson sampling. We first describe our general recipe and then analyze the performance of each estimator separately.

Let $X_1, \ldots, X_n$ be n independent samples drawn from a distribution p over k symbols. Consider an estimate $f_\alpha(X^n) = \frac{1}{1-\alpha} \log \widehat{\mathrm{P}}_\alpha(n, X^n)$ of $H_\alpha(\mathrm{p})$ which depends on X^n only through the multiplicities and the sample size. Here $\widehat{\mathrm{P}}_\alpha(n, X^n)$ is the corresponding estimate of $P_\alpha(\mathrm{p})$ – as discussed in Section I, small additive error in the estimate $f_\alpha(X^n)$ of $H_\alpha(\mathrm{p})$ is equivalent to small multiplicative error in the estimate $\widehat{\mathrm{P}}_\alpha(n, X^n)$ of $P_\alpha(\mathrm{p})$. For simplicity, we analyze a randomized estimator $\tilde{f}_\alpha$

described as follows: For $N \sim \mathrm{Poi}(n/2)$, let

$$\tilde{f}_\alpha(X^n) = \begin{cases} \text{constant}, & N > n, \\ \frac{1}{1-\alpha}\log \widehat{\mathrm{P}}_\alpha(n/2, X^N), & N \le n. \end{cases}$$

The following reduction to Poisson sampling is easy to show.

Lemma 7. (Poisson approximation 1) *For $n \ge 6\log(2/\epsilon)$ and $N \sim \mathrm{Poi}(n/2)$,*

$$\begin{aligned} &\mathbb{P}\left(|H_\alpha(\mathrm{p}) - \tilde{f}_\alpha(X^n)| > \delta\right) \\ &\le \mathbb{P}\left(|H_\alpha(\mathrm{p}) - \frac{1}{1-\alpha}\log \widehat{\mathrm{P}}_\alpha(\frac{n}{2}, X^N)| > \delta\right) + \frac{\epsilon}{2}. \end{aligned}$$

It remains to bound the probability on the right-side above, which can be done provided the bias and the variance of the estimator are bounded.

Lemma 8. *For $N \sim \mathrm{Poi}(n)$, let the power sum estimator $\widehat{\mathrm{P}}_\alpha = \widehat{\mathrm{P}}_\alpha(n, X^N)$ have bias and variance satisfying*

$$\begin{aligned} \left|\mathbb{E}\left[\widehat{\mathrm{P}}_\alpha\right] - P_\alpha(\mathrm{p})\right| &\le \frac{\delta}{2} P_\alpha(\mathrm{p}), \\ \mathbb{V}\mathrm{ar}\left[\widehat{\mathrm{P}}_\alpha\right] &\le \frac{\delta^2}{12} P_\alpha(\mathrm{p})^2. \end{aligned}$$

Then, there exists an estimator $\widehat{\mathrm{P}}'_\alpha$ that uses $36\mathrm{Poi}(n\log(2/\epsilon))$ samples and ensures

$$\mathbb{P}\left(\left|\widehat{\mathrm{P}}'_\alpha - P_\alpha(\mathrm{p})\right| > \delta\, P_\alpha(\mathrm{p})\right) \le \epsilon.$$

Proof. By Chebyshev's Inequality

$$\begin{aligned} &\mathbb{P}\left(\left|\widehat{\mathrm{P}}_\alpha - P_\alpha(\mathrm{p})\right| > \delta\, P_\alpha(\mathrm{p})\right) \\ &\le \mathbb{P}\left(\left|\widehat{\mathrm{P}}_\alpha - \mathbb{E}\left[\widehat{\mathrm{P}}_\alpha\right]\right| > \frac{\delta}{2} P_\alpha(\mathrm{p})\right) \le \frac{1}{3}. \end{aligned}$$

To reduce the probability of error to ϵ, we use the estimate $\widehat{\mathrm{P}}_\alpha$ repeatedly for $O(\log(1/\epsilon))$ independent samples X^N and take the estimate $\widehat{\mathrm{P}}'_\alpha$ to be the *sample median* of the resulting estimates[4]. Specifically, let $\widehat{\mathrm{P}}_1, ..., \widehat{\mathrm{P}}_t$ denote t-estimates of $P_\alpha(\mathrm{p})$ obtained by applying $\widehat{\mathrm{P}}_\alpha$ to independent sequences X^N, and let $\mathbb{1}_{\mathcal{E}_i}$ be the indicator function of the event $\mathcal{E}_i = \{|\widehat{\mathrm{P}}_i - P_\alpha(\mathrm{p})| > \delta\, P_\alpha(\mathrm{p})\}$. By the analysis above we have $\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_i}\right] \le 1/3$ and hence by Hoeffding's inequality

$$\mathbb{P}\left(\sum_{i=1}^{t} \mathbb{1}_{\mathcal{E}_i} > \frac{t}{2}\right) \le \exp(-t/18).$$

On choosing $t = 18\log(1/\epsilon)$ and noting that if more than half of $\widehat{\mathrm{P}}_1, ..., \widehat{\mathrm{P}}_t$ satisfy $|\widehat{\mathrm{P}}_i - P_\alpha(\mathrm{p})| \le \delta\, P_\alpha(\mathrm{p})$, then their median must also satisfy the same condition, we get that the median estimate satisfies the required error bound by using $\mathrm{Poi}(18n\log 1/\epsilon)$ samples. The claimed bound follows by Lemma 7. ∎

In the remainder of the section, we bound the bias and the variance for our estimators when the number of samples n are of the appropriate order. Denote by f^{e}_α, f^{u}_α, and $f^{d,\tau}_\alpha$, respectively, the empirical estimator $\frac{1}{1-\alpha}\log \widehat{P}^{\mathrm{e}}_\alpha$, the bias-corrected estimator $\frac{1}{1-\alpha}\log \widehat{P}^{\mathrm{u}}_\alpha$, and the polynomial approximation estimator $\frac{1}{1-\alpha}\log \widehat{P}^{d,\tau}_\alpha$. We begin by analyzing the performances of f^{e}_α and f^{u}_α and build-up on these steps to analyze $f^{d,\tau}_\alpha$.

[4]This technique is often referred to as the *median trick*.

A. Performance of empirical estimator

The empirical estimator was presented in (1). Using the Poisson sampling recipe given above, we derive upper bound for the sample complexity of the empirical estimator by bounding its bias and variance. The resulting bound for $\alpha > 1$ is given in Theorem 9 and for $\alpha < 1$ in Theorem 10.

Theorem 9. *For $\alpha > 1$, there exists a constant c_α depending only on α such that for every $0 < \delta < 1$, $k \in \mathbb{N}$, and $0 < \epsilon < 1$, the estimator f^e_α satisfies*

$$S^{f^e_\alpha}_\alpha(k, \delta, \epsilon) \le c_\alpha \max\left\{\frac{k}{\delta}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right\}.$$

Proof. Denote $\lambda_x \stackrel{\mathrm{def}}{=} n\mathrm{p}_x$. For $\alpha > 1$, using Lemma 4 we get

$$\begin{aligned} \left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right| &\le \frac{1}{n^\alpha}\sum_x |\mathbb{E}[N_x^\alpha] - \lambda_x^\alpha| \\ &\le \frac{C_\alpha}{n^\alpha}\sum_x \left(1 + \lambda_x^{\alpha-1}\right) \\ &= C_\alpha\left(\frac{k}{n^\alpha} + \frac{P_{\alpha-1}(\mathrm{p})}{n}\right) \\ &\le C_\alpha\left[\left(\frac{k}{n}\right)^\alpha + \frac{k}{n}\right] P_\alpha(\mathrm{p}), \end{aligned} \tag{8}$$

where the previous inequality noting that $k^{1-\alpha} \le P_\alpha(\mathrm{p})$ by (5) and $P_{\alpha-1}(\mathrm{p}) \le kP_\alpha(\mathrm{p})$ by Lemma 1(iv).

Similarly, using the independence of multiplicities under Poisson sampling, we have

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{V}\mathrm{ar}[N_x^\alpha] \\ &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - [\mathbb{E}N_x^\alpha]^2 \\ &\le \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha}, \end{aligned} \tag{9}$$

where the previous inequality is from Jensen's inequality since z^α is convex and $\mathbb{E}[N_x] = \lambda_x$. Therefore, by Lemma 4,

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] &\le \frac{C_\alpha}{n^{2\alpha}}\sum_x \left(1 + \lambda_x^{2\alpha-1}\right) \\ &= C_\alpha\left[\frac{k}{n^{2\alpha}} + \frac{P_{2\alpha-1}(\mathrm{p})}{n}\right] \\ &\le C_\alpha\left[\left(\frac{k}{n}\right)^{2\alpha} + \frac{k^{\frac{\alpha-1}{\alpha}}}{n}\right] P_\alpha(\mathrm{p})^2, \end{aligned} \tag{10}$$

where the last inequality follows upon noting that $k^{1-2\alpha} \leq P_{2\alpha}(\mathrm{p}) \leq P_\alpha(\mathrm{p})^2$ by (5) and $P_{2\alpha-1}(\mathrm{p}) \leq k^{\frac{\alpha-1}{\alpha}} P_\alpha(\mathrm{p})^2$ Lemma 1(iv). The claim follows from Lemma 8 upon choosing n to be sufficiently large. ■

Theorem 10. *For $0 < \alpha < 1$, there exists a constant c_α depending only on α such that for every $0 < \delta < 1$, $k \in \mathbb{N}$, and $0 < \epsilon < 1$, the estimator f^e_α satisfies*

$$S^{f^e_\alpha}_\alpha(k, \delta, \epsilon) \leq c_\alpha \max\left\{ \left(\frac{k}{\delta}\right)^{\frac{1}{\alpha}}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2} \right\}.$$

Proof. Proceeding as in the proof of Theorem 9, by Lemma 4 we have

$$\begin{aligned}\left| \mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p}) \right| &\leq \frac{1}{n^\alpha} \sum_x |\mathbb{E}[N_x^\alpha] - \lambda_x^\alpha| \\ &\leq \frac{2C_\alpha k}{n^\alpha} \\ &\leq 2C_\alpha P_\alpha(\mathrm{p}) \left(\frac{k^{1/\alpha}}{n}\right)^\alpha \qquad (11)\end{aligned}$$

where the previous inequality uses $1 \leq P_\alpha(\mathrm{p})$ from (4). For bounding the variance, note that

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] \\ =&\frac{1}{n^{2\alpha}} \sum_x \mathbb{V}\mathrm{ar}[N_x^\alpha] \\ =&\frac{1}{n^{2\alpha}} \sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - [\mathbb{E}N_x^\alpha]^2 \\ =&\frac{1}{n^{2\alpha}} \sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha} + \frac{1}{n^{2\alpha}} \sum_x \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2. \qquad (12)\end{aligned}$$

Consider the first term on the right-side. For $\alpha \leq 1/2$, it is bounded above by 0 since $z^{2\alpha}$ is concave in z, and for $\alpha > 1/2$ (10) yields

$$\begin{aligned}\frac{1}{n^{2\alpha}} \sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha} &\leq C_\alpha \left[\frac{k}{n^{2\alpha}} + \frac{P_{2\alpha-1}(\mathrm{p})}{n}\right] \\ &\leq C_\alpha \left[\frac{k}{n^{2\alpha}} + \frac{k^{\frac{1-\alpha}{\alpha}}}{n}\right] P_\alpha(\mathrm{p})^2,\end{aligned}$$

where we have used $P_{2\alpha-1}(\mathrm{p}) \leq k^{\frac{1-\alpha}{\alpha}} P_\alpha(\mathrm{p})^2$, which holds by Lemma 1(iii) and (5).

For the second term, we have

$$\begin{aligned}&\sum_x \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2 \\ =& \sum_x (\lambda_x^\alpha - \mathbb{E}[N_x^\alpha])(\lambda_x^\alpha + \mathbb{E}[N_x^\alpha]) \\ \leq& 2C_\alpha n^\alpha P_\alpha(\mathrm{p}) \left(\frac{k^{1/\alpha}}{n}\right)^\alpha \sum_x (\lambda_x^\alpha + \mathbb{E}[N_x^\alpha]) \\ \leq& 4C_\alpha n^{2\alpha} P_\alpha(\mathrm{p})^2 \left(\frac{k^{1/\alpha}}{n}\right)^\alpha,\end{aligned}$$

where the first inequality is by (11) and the final inequality holds since $\mathbb{E}[N_x^\alpha] \leq \lambda_x^\alpha$ by the concavity of z^α in z. The claim follows from Lemma 8. ■

B. Performance of bias-corrected estimator for integral α

To reduce the sample complexity for integer orders $\alpha > 1$ to below k, we follow the development of Shannon entropy estimators. Shannon entropy was first estimated via an empirical estimator, analyzed in, for instance, [2]. However, with $o(k)$ samples, the bias of the empirical estimator remains high [33]. This bias is reduced by the Miller-Madow correction [25], [33], but even then, $O(k)$ samples are needed for a reliable Shannon-entropy estimation [33].

Similarly, we reduce the bias for Rényi entropy estimators using *unbiased estimators* for p_x^α for integral α. We first describe our estimator, and in Theorem 11 we show that for $1 < \alpha \in \mathbb{N}$, $\widehat{P}^{\mathrm{u}}_\alpha$ estimates $P_\alpha(\mathrm{p})$ using $O(k^{1-1/\alpha}/\delta^2)$ samples. Theorem 15 in Section V shows that this number is optimal up to constant factors.

Consider the estimator for $P_\alpha(\mathrm{p})$ given by

$$\widehat{P}^{\mathrm{u}}_\alpha \stackrel{\mathrm{def}}{=} \sum_x \frac{N_x^{\underline{\alpha}}}{n^\alpha},$$

which is unbiased since by Lemma 2,

$$\mathbb{E}\left[\widehat{P}^{\mathrm{u}}_\alpha\right] = \sum_x \mathbb{E}\left[\frac{N_x^{\underline{\alpha}}}{n^\alpha}\right] = \sum_x p_x^\alpha = P_\alpha(\mathrm{p}).$$

Our *bias-corrected* estimator for $H_\alpha(\mathrm{p})$ is

$$\hat{H}_\alpha = \frac{1}{1-\alpha} \log \widehat{P}^{\mathrm{u}}_\alpha.$$

The next result provides a bound for the number of samples needed for the bias-corrected estimator.

Theorem 11. *For an integer $\alpha > 1$, there exists a constant c_α depending only on α such that for every $0 < \delta < 1$, $k \in \mathbb{N}$, and $0 < \epsilon < 1$, the estimator f^u_α satisfies*

$$S^{f^u_\alpha}_\alpha(k, \delta, \epsilon) \leq c_\alpha \left(\frac{k^{(\alpha-1)/\alpha}}{\delta^2} \log \frac{1}{\epsilon}\right).$$

Proof. Since the bias is 0, we only need to bound the variance to use Lemma 8. To that end, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\frac{\sum_x N_x^{\underline{\alpha}}}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}} \sum_x \mathbb{V}\mathrm{ar}\left[N_x^{\underline{\alpha}}\right] \\ &\leq \frac{1}{n^{2\alpha}} \sum_x \left(\lambda_x^\alpha (\lambda_x + \alpha)^\alpha - \lambda_x^{2\alpha}\right) \\ &= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} \sum_x \binom{\alpha}{r} \alpha^{\alpha-r} {\lambda_x}^{\alpha+r} \\ &= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} n^{\alpha+r} \binom{\alpha}{r} \alpha^{\alpha-r} P_{\alpha+r}(\mathrm{p}), \qquad (13)\end{aligned}$$

where the inequality uses Lemma 2. It follows from

Lemma 1(iv) that

$$\frac{1}{n^{2\alpha}}\frac{\mathbb{V}\mathrm{ar}\left[\sum_x N_x^{\underline{\alpha}}\right]}{P_\alpha(\mathrm{p})^2} \le \frac{1}{n^{2\alpha}}\sum_{r=0}^{\alpha-1} n^{\alpha+r}\binom{\alpha}{r}\alpha^{\alpha-r}\frac{P_{\alpha+r}(\mathrm{p})}{P_\alpha(\mathrm{p})^2}$$
$$\le \sum_{r=0}^{\alpha-1} n^{r-\alpha}\binom{\alpha}{r}\alpha^{\alpha-r}k^{(\alpha-1)(\alpha-r)/\alpha}$$
$$\le \sum_{r=0}^{\alpha-1}\left(\frac{\alpha^2 k^{(\alpha-1)/\alpha}}{n}\right)^{\alpha-r},$$

which is less than $\delta^2/12$ if $(\alpha^2 k^{1-1/\alpha}/n) \le 11\delta^2/144$. The claim follows by Lemma 8. ■

C. The polynomial approximation estimator

Concurrently with a conference version of this paper [1], a polynomial approximation based approach was proposed in [16] and [43] for estimating *additive functions* of the form $\sum_x f(\mathrm{p}_x)$. As seen in Theorem 11, polynomials of probabilities have succinct unbiased estimators. Motivated by this observation, instead of estimating f, these papers consider estimating a polynomial that is a *good approximation* to f. The underlying heuristic for this approach is that the difficulty in estimation arises from small probability symbols since empirical estimation is nearly optimal for symbols with large probabilities. On the other hand, there is no loss in estimating a polynomial approximation of the function of interest for symbols with small probabilities.

In particular, [16] considered the problem of estimating power sums $P_\alpha(\mathrm{p})$ up to additive accuracy and showed that $O\left(k^{1/\alpha}/\log k\right)$ samples suffice for $\alpha < 1$. Since $P_\alpha(\mathrm{p}) \ge 1$ for $\alpha < 1$, this in turn implies a similar sample complexity for estimating $H_\alpha(\mathrm{p})$ for $\alpha < 1$. On the other hand, $\alpha > 1$, the power sum $P_\alpha(\mathrm{p}) \le 1$ and can be small (*e.g.*, it is $k^{1-\alpha}$ for the uniform distribution). In fact, we show in the Appendix that additive-accuracy estimation of power sum is easy for $\alpha > 1$ and has a constant sample complexity. Therefore, additive guarantees for estimating the power sums are insufficient to estimate the Rényi entropy . Nevertheless, our analysis of the polynomial estimator below shows that it attains the $O(\log k)$ improvement in sample complexity over the empirical estimator even for the case $\alpha > 1$.

We first give a brief description of the polynomial estimator of [43] and then in Theorem 12 prove that for $\alpha > 1$ the sample complexity of $\widehat{P}_\alpha^{d,\tau}$ is $O(k/\log k)$. For completeness, we also include a proof for the case $\alpha < 1$, which is slightly different from the one in [16].

Let N_1, N_2 be independent $\mathrm{Poi}(n)$ random variables. We consider Poisson sampling with two set of samples drawn from p, first of size N_1 and the second N_2. Note that the total number of samples $N = N_1 + N_2 \sim \mathrm{Poi}(2n)$. The polynomial approximation estimator uses different estimators for different estimated values of symbol probability p_x. We use the first N_1 samples for comparing the symbol probabilities p_x with τ/n and the second is used for estimating p_x^α. Specifically, denote by N_x and N'_x the number of appearances of x in the N_1 and N_2 samples, respectively. Note that both N_x and N'_x have the same distribution $\mathrm{Poi}(n\mathrm{p}_x)$. Let τ be a threshold, and d be the degree chosen later. Given a threshold τ, the polynomial approximation estimator is defined as follows:

> $N'_x > \tau$: For all such symbols, estimate p_x^α using the empirical estimate $(N_x/n)^\alpha$.
> $N'_x \le \tau$: Suppose $q(x) = \sum_{m=0}^d a_m x^m$ is the polynomial satisfying Lemma 5. Since we expect p_x to be less than $2\tau/n$ in this case, we estimate p_x^α using an unbiased estimate of[5] $(2\tau/n)^\alpha q(n\mathrm{p}_x/2\tau)$, namely
>
> $$\left(\sum_{m=0}^d \frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right).$$

Therefore, for a given τ and d the combined estimator $\widehat{P}_\alpha^{d,\tau}$ is

$$\widehat{P}_\alpha^{d,\tau} \stackrel{\mathrm{def}}{=} \sum_{x:N'_x\le\tau}\left(\sum_{m=0}^d \frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right) + \sum_{x:N'_x>\tau}\left(\frac{N_x}{n}\right)^\alpha.$$

Denoting by $\hat{\mathrm{p}}_x$ the estimated probability of the symbol x, note that the polynomial approximation estimator relies on the empirical estimator when $\hat{\mathrm{p}}_x > \tau/n$ and uses the the bias-corrected estimator for estimating each term in the polynomial approximation of p_x^α when $\hat{\mathrm{p}}_x \le \tau/n$.

We derive upper bounds for the sample complexity of the polynomial approximation estimator. The bounds are valid in the large alphabet regime where k is sufficiently large for a fixed δ.

Theorem 12. *For $\alpha > 1$, $\delta > 0$, $0 < \epsilon < 1$, there exist constants c_1 and c_2 such that the estimator $\widehat{P}_\alpha^{d,\tau}$ with $\tau = c_1 \log n$ and $d = c_2 \log n$ satisfies*

$$S_\alpha^{\widehat{P}_\alpha^{d,\tau}}(k,\delta,\epsilon) \le O\left(\frac{k}{\log k}\frac{\log(1/\epsilon)}{\delta^{1/\alpha}}\right).$$

Proof. We follow the approach in [43] closely. Choose $\tau = c^* \log n$ such that with probability at least $1-\epsilon$ the events $N'_x > \tau$ and $N'_x \le \tau$ do not occur for all symbols x satisfying $\mathrm{p}_x \le \tau/(2n)$ and $\mathrm{p}_x > 2\tau/n$, respectively. Or equivalently, with probability at least $1-\epsilon$ all symbols x such that $N'_x > \tau$ satisfy $\mathrm{p}_x > \tau/(2n)$ and all symbols such that $N'_x \le \tau$ satisfy $\mathrm{p}_x \le 2\tau/n$. We condition on this event throughout the proof. For concreteness, we choose $c^* = 4$, which is a valid choice for $n > 20\log(1/\epsilon)$ by the Poisson tail bound and the union bound.

Let $q(x) = \sum_{m=0}^d a_m x^m$ satisfy the polynomial approximation error bound guaranteed by Lemma 5, *i.e.*,

$$\max_{x\in(0,1)} |q(x) - x^\alpha| < c_\alpha/d^{2\alpha} \tag{14}$$

To bound the bias of $\widehat{P}_\alpha^{d,\tau}$, note first that for $N'_x < \tau$ (assuming $\mathrm{p}_x \le 2\tau/nsmp$ and estimating

[5] Note that if $|q(x) - x^\alpha| < \epsilon$ for all $x \in [0,1]$, then $|\eta^\alpha q(x/\eta) - x^\alpha| < \eta^\alpha\epsilon$ for all $x \in [0,\eta]$.

$(2\tau/n)^\alpha q(np_x/2\tau))$

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{m=0}^{d}\frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right]-p_x^\alpha\right| \\
&= \left|\sum_{m=0}^{d}a_m\left(\frac{2\tau}{n}\right)^{\alpha-m}p_x^m-p_x^\alpha\right| \\
&= \frac{(2\tau)^\alpha}{n^\alpha}\left|\sum_{m=0}^{d}a_m\left(\frac{np_x}{2\tau}\right)^m-\left(\frac{np_x}{2\tau}\right)^\alpha\right| \\
&= \frac{(2\tau)^\alpha}{n^\alpha}\left|q\left(\frac{np_x}{2\tau}\right)-\left(\frac{np_x}{2\tau}\right)^\alpha\right| \\
&< \frac{(2\tau)^\alpha c_\alpha}{(nd^2)^\alpha}, \qquad (15)
\end{aligned}$$

where (15) uses (14) and $np_x/(2\tau)\le 1$.

For $N'_x>\tau$, the bias of empirical part of the power sum is bounded as Suppose $p_x>\tau/(2n)$, and $\tau>2$. Applying Lemma 4 for $\lambda=np_x$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\frac{N_x}{n}\right)^\alpha\right]-p_x^\alpha\right| \overset{(a)}{\le} \frac{1}{n^\alpha}C_\alpha\cdot(np_x)^{\alpha-1} \\
&= p_x^\alpha C_\alpha\frac{1}{np_x}\le p_x^\alpha\frac{2C_\alpha}{\tau},
\end{aligned}$$

where the last inequality uses $p_x>\tau/(2n)$, which holds for $N'_x>\tau$. Using the triangle inequality and applying the bounds above to each term, we obtain the following bound on the bias of $\widehat{P}_\alpha^{d,\tau}$:

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{P}_\alpha\right]-P_\alpha(p)\right| &\le \frac{k(2\tau)^\alpha c_\alpha}{(nd^2)^\alpha}+P_\alpha(p)\frac{2C_\alpha}{\tau} \\
&\le P_\alpha(p)\left[c_\alpha\left(\frac{k\cdot 2\tau}{nd^2}\right)^\alpha+\frac{2C_\alpha}{\tau}\right], \qquad (16)
\end{aligned}$$

where the last inequality uses $k<k^\alpha P_\alpha(p)$ from (5).

For variance, independence of multiplicities under Poisson sampling gives

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{P}_\alpha\right] &= \sum_{x:N'_x\le\tau}\mathbb{V}\mathrm{ar}\left(\sum_{m=0}^{d}\frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right) \qquad (17)\\
&\quad+\sum_{x:N_x>\tau}\mathbb{V}\mathrm{ar}\left(\frac{N_x}{n}\right)^\alpha.
\end{aligned}$$

Let $a=\max_m|a_m|$. By Lemma 2, for any x with $p_x\le 2\tau/n$,

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}\left(\sum_{m=0}^{d}\frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right) \\
&\le a^2d^2\max_{1\le m\le d}\left\{\frac{(2\tau)^{2\alpha-2m}}{n^{2\alpha}}\mathbb{V}\mathrm{ar}N_x^{\underline{m}}\right\} \\
&\overset{(a)}{\le} a^2d^2\max_{1\le m\le d}\left\{\frac{(2\tau)^{2\alpha-2m}}{n^{2\alpha}}(np_x)^m((np_x+m)^m-np_x^m)\right\} \\
&\overset{(b)}{\le}\frac{a^2d^2(2\tau+d)^{2\alpha}}{n^{2\alpha}}, \qquad (18)
\end{aligned}$$

where (a) is from Lemma 2, and (b) from plugging $np_x\le 2\tau$. Furthermore, using similar steps as (9) together with Lemma 4, for x with $p_x>\tau/(2n)$ we get

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\left(\frac{N_x}{n}\right)^\alpha\right] &\le \frac{1}{n^{2\alpha}}\left(\mathbb{E}\left[N_x^{2\alpha}\right]-\lambda_x^{2\alpha}\right), \\
&\le \frac{1}{n^{2\alpha}}\cdot C_{2\alpha}(np_x)^{2\alpha-1} \qquad (19)\\
&= C_{2\alpha}\frac{p_x^{2\alpha}}{np_x} \\
&\le 2C_{2\alpha}\frac{p_x^{2\alpha}}{\tau}
\end{aligned}$$

The two bounds above along with Lemma 1 and (5) yield

$$\mathbb{V}\mathrm{ar}\left[\widehat{P}_\alpha\right]\le P_\alpha(p)^2\left[\frac{a^2d^2(2\tau+d)^{2\alpha}}{n}\left(\frac{k}{n}\right)^{2\alpha-1}+\frac{2C_{2\alpha}}{\tau}\right]. \qquad (20)$$

For $d=\tau/8=\frac{1}{2}\log n$, the second terms in (16) are $o(1)$ in n which gives[6]

$$\left|\mathbb{E}\left[\widehat{P}_\alpha\right]-P_\alpha(p)\right| = P_\alpha(p)\left(c_\alpha\left(\frac{32k}{(n\log n)}\right)^\alpha+o(1)\right).$$

Recall from (7) that $a<(1+c_\alpha/d^{2\alpha})(\sqrt{2}+1)^d$, and therefore, $a^2=O((\sqrt{2}+1)^{\log n})=n^{c_0}$ for some $c_0<1$. Using (20) we get

$$\mathbb{V}\mathrm{ar}\left[\widehat{P}_\alpha\right]=O\left(P_\alpha(p)^2\frac{n^{c_0}\log^{2\alpha+2}n}{n}\left(\frac{k}{n}\right)^{2\alpha-1}\right).$$

Therefore, the result follows from Lemma 8 for k sufficiently large. ■

We now prove an analogous result for $\alpha<1$.

Theorem 13. *For $\alpha<1$, $\delta>0$, $0<\epsilon<1$, there exist constants c_1 and c_2 such that the estimator $\widehat{P}_\alpha^{d,\tau}$ with $\tau=c_1\log n$ and $d=c_2\log n$ satisfies*

$$S_\alpha^{\widehat{P}_\alpha^{d,\tau}}(k,\delta,\epsilon)\le O\left(\frac{k^{1/\alpha}}{\log k}\frac{\log(1/\epsilon)}{\alpha^2\delta^{1/\alpha}}\right).$$

Proof. We proceed as in the previous proof and set τ to be $4\log n$. The contribution to the bias of the estimator for a symbol x with $N'_x<\tau$ remains bounded as in (15). For a symbol x with $N'_x>\tau$, the bias contribution of the empirical estimator is bounded as

$$\left|\mathbb{E}\left[\left(\frac{N_x}{n}\right)^\alpha\right]-p_x^\alpha\right|\le\frac{C_\alpha p_x^{\alpha-1}}{n}\le\frac{2C_\alpha p_x^\alpha}{\tau},$$

where the first inequality is by Lemma 4 and the second uses $p_x>\tau/(2n)$, which holds if $N'_x>\tau$. Thus, we obtain the following bound on the bias of $\widehat{P}_\alpha^{d,\tau}$:

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{P}_\alpha\right]-P_\alpha(p)\right| &\le \frac{k(2\tau)^\alpha c_\alpha}{(nd^2)^\alpha}+\frac{2}{\tau}P_\alpha(p) \\
&\le P_\alpha(p)\left[c_\alpha\left(\frac{k^{1/\alpha}\cdot 2\tau}{nd^2}\right)^\alpha+\frac{2}{\tau}\right],
\end{aligned}$$

[6]This approximation is valid in the large alphabet regime, where k is sufficiently large for a fixed δ.

where the last inequality is by (4).

To bound the variance, first note that bound (18) still holds of $\mathrm{p}_x \le 2\tau/n$. To bound the contribution to the variance from the terms with $n\mathrm{p}_x > \tau/2$, we follow the steps in the proof of Theorem 10. In particular, (12) gives

$$\mathbb{V}\mathrm{ar}\left[\sum_{x:N'_x>\tau}\frac{N_x^\alpha}{n^\alpha}\right] \le \frac{1}{n^{2\alpha}}\left(\sum_{x:N'_x>\tau}\mathbb{E}\left[N_x^{2\alpha}\right]-\lambda_x^{2\alpha}\right) + \frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\left(\lambda_x^{2\alpha}-[\mathbb{E}N_x^\alpha]^2\right). \quad (21)$$

We consider each term separately. The first term is at most zero for $\alpha \le 1/2$. For $\alpha > 1/2$, using Lemma 4,

$$\begin{aligned}\frac{1}{n^{2\alpha}}\left(\sum_{x:N'_x>\tau}\mathbb{E}\left[N_x^{2\alpha}\right]-\lambda_x^{2\alpha}\right) &\le \sum_{x:N'_x>\tau} C_{2\alpha}\frac{(\mathrm{p}_x)^{2\alpha}}{n\mathrm{p}_x}\\ &\le 2C_{2\alpha}\frac{P_{2\alpha}(\mathrm{p})}{\tau}\\ &\le 2C_{2\alpha}\frac{P_\alpha(\mathrm{p})^2}{\tau}.\end{aligned}$$

For the second term, we have

$$\begin{aligned}&\frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\lambda_x^{2\alpha}-[\mathbb{E}N_x^\alpha]^2 \quad (22)\\ &= \frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\left(\lambda_x^\alpha-\mathbb{E}[N_x^\alpha]\right)\left(\lambda_x^\alpha+\mathbb{E}[N_x^\alpha]\right)\\ &\le \frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\left(\lambda_x^{\alpha-1}\right)\left(2\lambda_x^\alpha\right)\\ &= 2\sum_{x:N'_x>\tau}\frac{\mathrm{p}_x^{2\alpha}}{n\mathrm{p}_x}\\ &\le \frac{4}{\tau}P_\alpha(\mathrm{p})^2,\end{aligned}$$

where the first inequality follows from Lemma 4 and concavity of z^α in z and the second from $n\mathrm{p}_x > \tau/2$ and Lemma 1.

Thus, the contribution of the terms corresponding to $N'_x > \tau$ in the bias and the variance are $P_\alpha(\mathrm{p})\cdot o(1)$ and $P_\alpha(\mathrm{p})^2\cdot o(1)$, respectively, and can be ignored. Choosing $d = \frac{\alpha}{2}\log n$ and combining the observations above, we get the following bound for the bias:

$$\left|\mathbb{E}\left[\widehat{\mathrm{P}}_\alpha\right]-P_\alpha(\mathrm{p})\right| = P_\alpha(\mathrm{p})\left(c_\alpha\left(\frac{32k^{1/\alpha}}{n\log n\alpha^2}\right)^\alpha+o(1)\right),$$

and, using (18), the following bound for the variance:

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\widehat{\mathrm{P}}_\alpha\right] \quad (23)\\ &\le k\frac{a^2d^2(2\tau+d)^{2\alpha}}{n^{2\alpha}}+P_\alpha(\mathrm{p})^2\cdot o(1)\\ &\le P_\alpha(\mathrm{p})^2\left[\left(\frac{a^2}{n^\alpha}\right)(9\log n)^{2\alpha+2}\left(\frac{k^{\frac{1}{\alpha}}}{n}\right)^\alpha+o(1)\right].\end{aligned}$$

Here a^2 is the largest squared coefficient of the approximating polynomial and, by (7), is $O(2^{2c_0d}) = O(n^{c_0\alpha})$ for some $c_0 < 1$. Thus, $a^2 = o(n^\alpha)$ and the proof follows by Lemma 8. ■

IV. Examples and experiments

We begin by computing Rényi entropy for uniform and Zipf distributions; the latter example illustrates the lack of uniform continuity of $H_\alpha(\mathrm{p})$ in α.

Example 1. The *uniform distribution* U_k over $[k] = \{1,\dots,k\}$ is given by

$$p_i = \frac{1}{k} \quad \text{for } i \in [k].$$

Its Rényi entropy for every order $1 \ne \alpha \ge 0$, and hence for all $\alpha \ge 0$, is

$$H_\alpha(U_k) = \frac{1}{1-\alpha}\log\sum_{i=1}^k\frac{1}{k^\alpha} = \frac{1}{1-\alpha}\log k^{1-\alpha} = \log k.$$

Example 2. The *Zipf distribution* $Z_{\beta,k}$ for $\beta > 0$ and $k \in [k]$ is given by

$$\mathrm{p}_i = \frac{i^{-\beta}}{\sum_{j=1}^k j^{-\beta}} \quad \text{for } i \in [k].$$

Its Rényi entropy of order $\alpha \ne 1$ is

$$H_\alpha(Z_{\beta,k}) = \frac{1}{1-\alpha}\log\sum_{i=1}^k i^{-\alpha\beta} - \frac{\alpha}{1-\alpha}\log\sum_{i=1}^k i^{-\beta}.$$

Table II summarizes the leading term $g(k)$ in the approximation[7] $H_\alpha(Z_{\beta,k}) \sim g(k)$.

	$\beta<1$	$\beta=1$	$\beta>1$
$\alpha\beta<1$	$\log k$	$\frac{1-\alpha\beta}{1-\alpha}\log k$	$\frac{1-\alpha\beta}{1-\alpha}\log k$
$\alpha\beta=1$	$\frac{\alpha-\alpha\beta}{\alpha-1}\log k$	$\frac{1}{2}\log k$	$\frac{1}{1-\alpha}\log\log k$
$\alpha\beta>1$	$\frac{\alpha-\alpha\beta}{\alpha-1}\log k$	$\frac{\alpha}{\alpha-1}\log\log k$	constant

TABLE II: The leading terms $g(k)$ in the approximations $H_\alpha(Z_{\beta,k}) \sim g(k)$ for different values of $\alpha\beta$ and β. The case $\alpha\beta = 1$ and $\beta = 1$ corresponds to the Shannon entropy of $Z_{1,k}$.

In particular, for $\alpha > 1$

$$H_\alpha(Z_{1,k}) = \frac{\alpha}{1-\alpha}\log\log k + \Theta\left(\frac{1}{k^{\alpha-1}}\right) + c(\alpha),$$

and the difference $|H_2(\mathrm{p}) - H_{2+\epsilon}(\mathrm{p})|$ is $O(\epsilon\log\log k)$. Therefore, even for very small ϵ this difference is unbounded and approaches infinity in the limit as k goes to infinity.

We now illustrate the performance of the proposed estimators for various distributions for $\alpha = 2$ in Figures 2 and $\alpha = 1.5$ in Figures 3. For $\alpha = 2$, we compare the performance of bias-corrected and empirical estimators. For $\alpha = 1.5$, we compare the performance of the polynomial-approximation and the empirical estimator.

[7] We say $f(n) \sim g(n)$ to denote $\lim_{n\to\infty} f(n)/g(n) = 1$.

For the polynomial-approximation estimator, the threshold τ is chosen as $\tau = \ln(n)$ and the approximating polynomial degree is chosen as $d = \lceil 1.5\tau \rceil$.

We test the performance of these estimators over six different distributions: the uniform distribution, a step distribution with half of the symbols having probability $1/(2k)$ and the other half have probability $3/(2k)$, Zipf distribution with parameter $3/4$ ($p_i \propto i^{-3/4}$), Zipf distribution with parameter $1/2$ ($p_i \propto i^{-1/2}$), a randomly generated distribution using the uniform prior on the probability simplex, and another one generated using the Dirichlet-$1/2$ prior.

In both the figures the true value is shown in black and the estimated values are color-coded, with the solid line representing their mean estimate and the shaded area corresponding to one standard deviation. As expected, bias-corrected estimators outperform empirical estimators for $\alpha = 2$ and polynomial-approximation estimators perform better than empirical estimators for $\alpha = 1.5$.

V. Lower bounds on sample complexity

We now establish lower bounds on $S_\alpha(k, \delta, \epsilon)$. The proof is based on exhibiting two distributions p and q with $H_\alpha(\mathrm{p}) \neq H_\alpha(\mathrm{q})$ such that the set of N_x's have very similar distribution from p and q, if fewer samples than the claimed lower bound are available. This method is often referred to as *Le Cam's two-point method* (see, for instance, [46]). The key idea is summarized in the following result which is easy to derive.

Lemma 14. *If for two distributions* p *and* q *on* $\mathcal{X}$ *and* $n \in \mathbb{N}$ *the total variation distance* $\|\mathrm{p}^n - \mathrm{q}^n\| < \epsilon$, *then for every function* $\hat{f}$, *either*

$$\mathrm{p}\left(|H_\alpha(\mathrm{p}) - \hat{f}(X^n)| \geq \frac{|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})|}{2}\right) \geq \frac{1-\epsilon}{2},$$

or

$$\mathrm{q}\left(|H_\alpha(\mathrm{q}) - \hat{f}(X^n)| \geq \frac{|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})|}{2}\right) \geq \frac{1-\epsilon}{2}.$$

A. Lower bound for integer α

We first prove a lower bound for integers $\alpha > 1$ which matches the upper bound in Theorem 11 up to a constant factor. In fact, the bound is valid for any $\alpha > 1$.

Theorem 15. *Given an* $1 < \alpha$ *and* $0 < \epsilon < 1$, *there exists a constant* c *depending on* α *and* ϵ *such that for every* $\delta > 0$ *sufficiently small (depending only on* α*) and every* k *sufficiently large (depending only on* α*)*

$$S_\alpha(k, \delta, \epsilon) \geq c\left(\frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right).$$

Proof. We rely on Lemma 14 and exhibit two distributions p and q with appropriate properties. Specifically, consider the following distributions p and q over $[k]$: $\mathrm{p}_1 = 1/k^{1-1/\alpha}$, and for $x = 2, \ldots, k$, $\mathrm{p}_x = (1 - \mathrm{p}_1)/(k-1)$; $\mathrm{q}_1 = (1+\delta)/k^{1-1/\alpha}$, and for $x = 2, \ldots, k$, $\mathrm{q}_x = (1 - \mathrm{q}_1)/(k-1)$. Then, we have

$$P_\alpha(\mathrm{p}) = \mathrm{p}_1^\alpha + \frac{(1-\mathrm{p}_1)^\alpha}{(k-1)^{\alpha-1}},$$

and

$$P_\alpha(\mathrm{q}) = \mathrm{q}_1^\alpha + \frac{(1-\mathrm{q}_1)^\alpha}{(k-1)^{\alpha-1}}.$$

By noting that $(1+\delta)^\alpha \geq (1+\alpha\delta)$ for $\alpha > 1$ and using Taylor's approximation

$$\begin{aligned}
&P_\alpha(\mathrm{q}) - P_\alpha(\mathrm{p}) \\
&= \mathrm{q}_1^\alpha - \mathrm{p}_1^\alpha + \frac{[(1-\mathrm{q}_1)^\alpha - (1-\mathrm{p}_1)^\alpha]}{(k-1)^{\alpha-1}} \\
&\geq \frac{\delta}{k^{\alpha-1}} - \frac{1}{(k-1)^{\alpha-1}} \cdot \frac{\alpha\delta}{k^{1-1/\alpha}} \cdot \left(1 - \frac{1}{k^{1-1/\alpha}}\right)^{\alpha-1} \\
&\geq \frac{\delta}{2k^{\alpha-1}},
\end{aligned}$$

where the last inequality holds if k is larger than a constant depending on α. Therefore, for k sufficiently large

$$\frac{|P_\alpha(\mathrm{q}) - P_\alpha(\mathrm{p})|}{P_\alpha(\mathrm{p})} \geq \frac{\delta}{4},$$

and so, $|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| \geq \delta/(1-\alpha)8$ for δ sufficiently small. To complete the proof, we show that there exists a constant C_ϵ such that $\|\mathrm{p}^n - \mathrm{q}^n\| \leq \epsilon$ if $n \leq C_\epsilon k^{1-1/\alpha}/\delta^2$. To that end, we bound the squared Hellinger distance between p^n and q^n given by

$$h^2(\mathrm{p}, \mathrm{q}) = 2 - 2\sum_x \sqrt{\mathrm{p}_x \mathrm{q}_x} = \sum_x (\sqrt{\mathrm{p}_x} - \sqrt{\mathrm{q}_x})^2.$$

Therefore, for $\delta < 1$ and k sufficiently large so that $p_1, q_1 \leq 1/2$,

$$\begin{aligned}
h^2(\mathrm{p}, \mathrm{q}) &= (\sqrt{p_1} - \sqrt{q_1})^2 + \left(\sqrt{1-p_1} - \sqrt{1-q_1}\right)^2 \\
&= \frac{(p_1 - q_1)^2}{\left(\sqrt{p_1} + \sqrt{q_1}\right)^2} + \frac{(p_1 - q_1)^2}{\left(\sqrt{1-p_1} + \sqrt{1-q_1}\right)^2} \\
&\leq 2\frac{(p_1 - q_1)^2}{\left(\sqrt{p_1} + \sqrt{q_1}\right)^2} \\
&\leq 2\frac{(p_1 - q_1)^2}{p_1} \\
&= \frac{2\delta^2}{k^{1-1/\alpha}}.
\end{aligned}$$

The required bound for $\|\mathrm{p}^n - \mathrm{q}^n\|$ follows using the standard steps (*cf.* [46]) below:

$$\begin{aligned}
\|\mathrm{p}^n - \mathrm{q}^n\| &\leq \sqrt{h^2(\mathrm{p}, \mathrm{q})} \\
&= \sqrt{1 - \left(1 - \frac{1}{2}h^2(\mathrm{p}, \mathrm{q})\right)^n} \\
&\leq \sqrt{\frac{n}{2}h^2(\mathrm{p}, \mathrm{q})}.
\end{aligned}$$

■

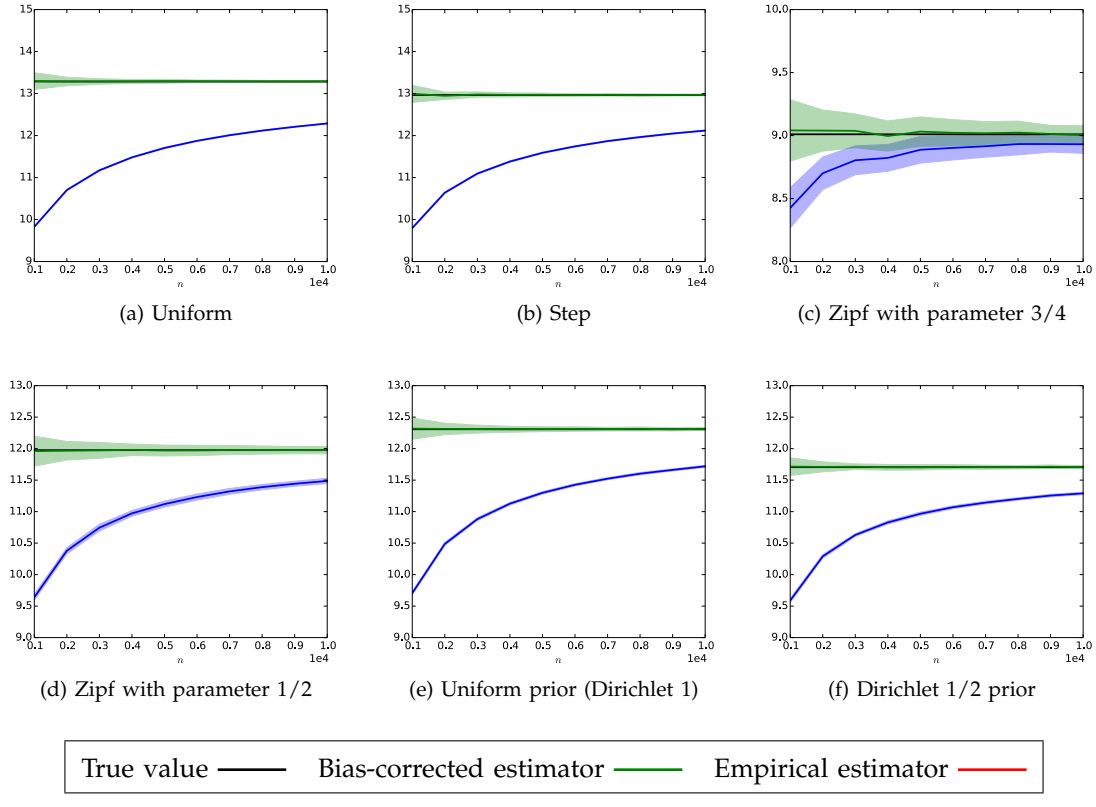

Fig. 2: Rényi entropy estimates for order 2 for support 10000, number of samples ranging from 1000 to 10000, averaged over 100 trials.

As a corollary, the result above and the upper bound of Theorem 11 yields the following characterization of $S_\alpha(k,\delta,\epsilon)$.

Corollary 16. *Given an $1 < \alpha \in \mathbb{N}$ and $0 < \epsilon < 1$, for every $\delta > 0$ sufficiently small (depending only on α) and every k sufficiently large (depending only on α)*

$$S_\alpha(k,\delta,\epsilon) = \Theta\left(\frac{k^{(\alpha-1)/\alpha}}{\delta^2}\right),$$

where constants implied by Θ depend only on ϵ and α.

B. Lower bound for noninteger α

Next, we lower bound $S_\alpha(k)$ for noninteger $\alpha > 1$ and show that it must be almost linear in k. While we still rely on Lemma 14 for our lower bound, we take recourse to Poisson sampling to simplify our calculations.

Lemma 17. (Poisson approximation 2) *Suppose there exist $\delta, \epsilon > 0$ such that, with $N \sim \mathrm{Poi}(2n)$, for all estimators $\hat{f}$ we have*

$$\max_{\mathrm{p}\in\mathcal{P}} \mathbb{P}\left(|H_\alpha(\mathrm{p}) - \hat{f}_\alpha(X^N)| > \delta\right) > \epsilon,$$

where $\mathcal{P}$ is a fixed family of distributions. Then, for all fixed length estimators $\tilde{f}$

$$\max_{\mathrm{p}\in\mathcal{P}} \mathbb{P}\left(|H_\alpha(\mathrm{p}) - \tilde{f}_\alpha(X^n)| > \delta\right) > \frac{\epsilon}{2},$$

when $n > 4\log(2/\epsilon)$.

Also, it will be convenient to replace the observations X^N with its *profile* $\Phi = \Phi(X^N)$ [31], i.e., $\Phi = (\Phi_1, \Phi_2, \ldots)$ where Φ_l is the number of elements x that appear l times in the sequence X^N. The following well-known result says that for estimating $H_\alpha(\mathrm{p})$, it suffices to consider only the functions of the profile.

Lemma 18. (Sufficiency of profiles). *Consider an estimator $\hat{f}$ such that*

$$\mathbb{P}\left(|H_\alpha(\mathrm{p}) - \hat{f}(X^N)| > \delta\right) \le \epsilon, \quad \textit{for all } \mathrm{p}.$$

Then, there exists an estimator $\tilde{f}(X^N) = \tilde{f}(\Phi)$ such that

$$\mathbb{P}\left(|H_\alpha(\mathrm{p}) - \tilde{f}(\Phi)| > \delta\right) \le \epsilon, \quad \textit{for all } \mathrm{p}.$$

Thus, lower bounds on the sample complexity will follow upon showing a contradiction for the second inequality above when the number of samples n is sufficiently small. We obtain the required contradiction by using Lemma 14 upon showing there are distributions p and q of support-size k such that the following hold:

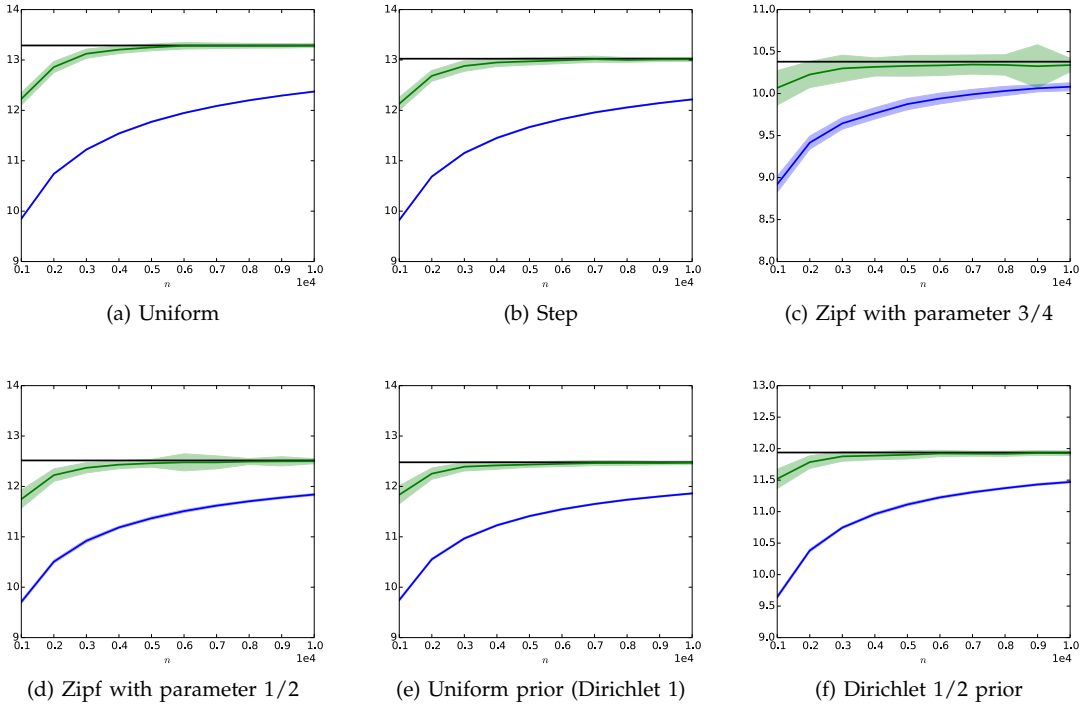

Fig. 3: Rényi entropy estimates for order 1.5 for support 10000, number of samples ranging from 1000 to 10000, averaged over 100 trials.

(i) There exists $\delta > 0$ such that

$$|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| > \delta; \tag{24}$$

(ii) denoting by p_Φ and q_Φ, respectively, the distributions on the profiles under Poisson sampling corresponding to underlying distributions p and q, there exist $\epsilon > 0$ such that

$$\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\| < \epsilon, \tag{25}$$

if $n < k^{c(\alpha)}$.

Therefore, it suffices to find two distributions p and q with different Rényi entropies and with small total variation distance between the distributions of their profiles, when n is sufficiently small. For the latter requirement, we recall a result of [42] that allows us to bound the total variation distance in (25) in terms of the differences of power sums $|P_a(\mathrm{p}) - P_a(\mathrm{q})|$.

Theorem 19. *[42] Given distributions* p *and* q *such that*

$$\max_x \max\{\mathrm{p}_x; \mathrm{q}_x\} \leq \frac{\epsilon}{40n},$$

for Poisson sampling with $N \sim \mathrm{Poi}(n)$*, it holds that*

$$\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\| \leq \frac{\epsilon}{2} + 5\sum_a n^a |P_a(\mathrm{p}) - P_a(\mathrm{q})|.$$

It remains to construct the required distributions p and q, satisfying (24) and (25) above. By Theorem 19, the total variation distance $\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\|$ can be made small by ensuring that the power sums of distributions p and q are matched, that is, we need distributions p and q with different Rényi entropies and identical power sums for as large an order as possible. To that end, for every positive integer d and every vector $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$, associate with $\mathbf{x}$ a distribution $\mathrm{p}^{\mathbf{x}}$ of support-size dk such that

$$\mathrm{p}^{\mathbf{x}}_{ij} = \frac{|x_i|}{k\|\mathbf{x}\|_1}, \quad 1 \leq i \leq d,\ 1 \leq j \leq k.$$

Note that

$$H_\alpha(\mathrm{p}^{\mathbf{x}}) = \log k + \frac{\alpha}{\alpha - 1} \log \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_\alpha},$$

and for all a

$$P_a\left(\mathrm{p}^{\mathbf{x}}\right) = \frac{1}{k^{a-1}} \left(\frac{\|\mathbf{x}\|_a}{\|\mathbf{x}\|_1}\right)^a.$$

We choose the required distributions p and q, respectively, as $\mathrm{p}^{\mathbf{x}}$ and $\mathrm{p}^{\mathbf{y}}$, where the vectors $\mathbf{x}$ and $\mathbf{y}$ are given by the next result.

Lemma 20. *For every* $d \in \mathbb{N}$ *and* α *not integer, there exist*

positive vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *such that*

$$\begin{aligned}&\|\mathbf{x}\|_r = \|\mathbf{y}\|_r, \quad 1 \le r \le d-1,\\ &\|\mathbf{x}\|_d \ne \|\mathbf{y}\|_d,\\ &\|\mathbf{x}\|_\alpha \ne \|\mathbf{y}\|_\alpha.\end{aligned}$$

Proof. Let $\mathbf{x} = (1, ..., d))$. Consider the polynomial

$$p(z) = (z - x_1)...(z - x_d),$$

and $q(z) = p(z) - \Delta$, where Δ is chosen small enough so that $q(z)$ has d positive roots. Let $y_1, ..., y_d$ be the roots of the polynomial $q(z)$. By Newton-Girard identities, while the sum of dth power of roots of a polynomial does depend on the constant term, the sum of first $d-1$ powers of roots of a polynomial do not depend on it. Since $p(z)$ and $q(z)$ differ only by a constant, it holds that

$$\sum_{i=1}^{d} x_i^r = \sum_{i=1}^{d} y_i^r, \quad 1 \le r \le d-1,$$

and that

$$\sum_{i=1}^{d} x_i^d \ne \sum_{i=1}^{d} y_i^d.$$

Furthermore, using a first order Taylor approximation, we have

$$y_i - x_i = \frac{\Delta}{p'(x_i)} + o(\Delta),$$

and for any differentiable function g,

$$g(y_i) - g(x_i) = g'(x_i)(y_i - x_i) + o(|y_i - x_i|).$$

It follows that

$$\sum_{i=1}^{d} g(y_i) - g(x_i) = \sum_{i=1}^{d} \frac{g'(x_i)}{p'(x_i)}\Delta + o(\Delta),$$

and so, the left side above is nonzero for all Δ sufficiently small provided

$$\sum_{i=1}^{d} \frac{g'(x_i)}{p'(x_i)} \ne 0.$$

Upon choosing $g(x) = x^\alpha$, we get

$$\sum_{i=1}^{d} \frac{g'(x_i)}{p'(x_i)} = \frac{\alpha}{d!} \sum_{i=1}^{d} \left(\binom{d}{i} \right) (-1)^{d-i}\, i^\alpha.$$

Denoting the right side above by $h(\alpha)$, note that $h(i) = 0$ for $i = 1, ..., d-1$. Since $h(\alpha)$ is a linear combination of d exponentials, it cannot have more than $d-1$ zeros (see, for instance, [40]). Therefore, $h(\alpha) \ne 0$ for all $\alpha \notin \{1, ..., d-1\}$; in particular, $\|\mathbf{x}\|_\alpha \ne \|\mathbf{y}\|_\alpha$ for all Δ sufficiently small. ■

We are now in a position to prove our converse results.

Theorem 21. *Given a noninteger* $\alpha > 1$*, for any fixed* $0 < \epsilon < 1/2$*, we have*

$$S_\alpha(k, \delta, \epsilon) = \tilde{\tilde{\Omega}}\,(k).$$

Proof. For a fixed d, let distributions p and q be as in the previous proof. Then, as in the proof of Theorem 21, inequality (24) holds by Lemma 20 and (25) holds by Theorem 19 if $n < C_2 k^{(d-1)/d}$. The theorem follows since d can be arbitrary large. ■

Finally, we show that $S_\alpha(k)$ must be super-linear in k for $\alpha < 1$.

Theorem 22. *Given* $0 < \alpha < 1$*, for every* $0 < \epsilon < 1/2$*, we have*

$$S_\alpha(k, \delta, \epsilon) = \tilde{\tilde{\Omega}}\left(k^{1/\alpha}\right).$$

Proof. Consider distributions p and q on an alphabet of size $kd + 1$, where

$$\mathrm{p}_{ij} = \frac{\mathrm{p}^{\mathbf{x}}_{ij}}{k^\beta} \text{ and } \mathrm{q}_{ij} = \frac{\mathrm{p}^{\mathbf{x}}_{ij}}{k^\beta}, \quad 1 \le i \le d,\, 1 \le j \le k,$$

where the vectors $\mathbf{x}$ and $\mathbf{y}$ are given by Lemma 20 and β satisfies $\alpha(1+\beta) < 1$, and

$$\mathrm{p}_0 = \mathrm{q}_0 = 1 - \frac{1}{k^\beta}.$$

For this choice of p and q, we have

$$\begin{aligned}P_a(\mathrm{p}) &= \left(1 - \frac{1}{k^\beta}\right)^a + \frac{1}{k^{a(1+\beta)-1}} \left(\frac{\|\mathbf{x}\|_a}{\|\mathbf{x}\|_1}\right)^a,\\ H_\alpha(\mathrm{p}) &= \frac{1-\alpha(1+\beta)}{1-\alpha} \log k + \frac{\alpha}{1-\alpha} \log \frac{\|\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_1}\\ &\quad + O(k^{a(1+\beta)-1}),\end{aligned}$$

and similarly for q, which further yields

$$|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| = \frac{\alpha}{1-\alpha} \left| \log \frac{\|\mathbf{x}\|_\alpha}{\|\mathbf{y}\|_\alpha} \right| + O(k^{a(1+\beta)-1}).$$

Therefore, for sufficiently large k, (24) holds by Lemma 20 since $\alpha(1+\beta) < 1$, and for $n < C_2 k^{(1+\beta-1/d)}$ we get (25) by Theorem 19 as

$$\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\| \le \frac{\epsilon}{2} + 5 \sum_{a \ge d} \left(\frac{n}{k^{1+\beta-1/a}} \right)^a \le \epsilon.$$

The theorem follows since d and $\beta < 1/\alpha - 1$ are arbitrary. ■

C. Sample complexity of empirical estimator

We now derive lower bounds for the sample complexity of the empirical estimator of $H_\alpha(\mathrm{p})$ and characterize it up to constant factors.

Theorem 23. *Given* $\alpha > 1$*, there exists a constant* c_α *such that for every* δ *sufficiently small (depending only on* α*) and every* k *sufficiently large*

$$S^{fe}_\alpha(k, \delta, 0.9) \ge c_\alpha \left(\frac{k}{\delta}\right).$$

Proof. We prove the lower bound for the uniform distribution over k symbols in two steps. We first show that for any constant $c_1 > 1$ if $n < k/c_1$ then the additive approximation error is at least δ with probability one,

for every $\delta < \log c_1$. Then, assuming that $n \geq k/c_1$, we show that the additive approximation error is at least δ with probability greater than 0.9 if $n < k/\delta$.

For the first claim, we assume without loss of generality that $n \leq k$, since otherwise the proof is complete. Note that for $\alpha > 1$ the function $(\mathrm{p}_i - y)^\alpha + (\mathrm{p}_j + y)^\alpha$ is decreasing in y for all y such that $(\mathrm{p}_i - y) > (\mathrm{p}_j + y)$. Thus, the minimum value of $\sum_x \left(\frac{N_x}{n}\right)^\alpha$ is attained when each N_x is either 0 or 1. It follows that

$$\widehat{P}_\alpha^{\mathrm{e}} = \sum_x \left(\frac{N_x}{n}\right)^\alpha \geq \frac{1}{n^{\alpha-1}},$$

which is the same as

$$H_\alpha(\mathrm{p}) - \frac{1}{\alpha-1}\log\frac{1}{\widehat{P}_\alpha^{\mathrm{e}}} \geq \log\frac{k}{n}.$$

Hence, for any $c_1 > 1$ and $n < k/c_1$ and any $0 \leq \delta \leq \log c_1$, the additive approximation error is more than δ with probability one.

Moving to the second claim, suppose now $n > k/c_1$. We first show that with high probability, the multiplicities of a linear fraction of k symbols should be at least a factor of standard deviation higher than the mean. Specifically, let

$$A = \sum_x \mathbb{1}\left(N_x \geq \frac{n}{k} + c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right).$$

Then,

$$\begin{aligned}\mathbb{E}[A] &= \sum_x \mathbb{E}\left[\mathbb{1}\left(N_x \geq \frac{n}{k} + c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right)\right] \\ &= k \cdot p\left(N_x \geq \frac{n}{k} + c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right) \\ &\geq k \cdot Q(c_2),\end{aligned}$$

where Q denotes the Q-function, *i.e.*, the tail of the standard normal random variable, and the final inequality uses Slud's inequality [38, Theorem 2.1].

Note that A is a function of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, and changing any one X_i changes A by at most 2. Hence, by McDiarmid's inequality,

$$\Pr(A \geq \mathbb{E}[A] - \sqrt{8n}) \geq 1 - e^{-4} \geq 0.9.$$

Therefore, for all k sufficiently large (depending on δ) and denoting $c = Q(c_2)/2$, at least ck symbols occur more than $\frac{n}{k} + c_2\sqrt{\frac{n}{k}}$ times with probability greater than 0.9. Using the fact that $(\mathrm{p}_i - y)^\alpha + (\mathrm{p}_j + y)^\alpha$ is decreasing if $(\mathrm{p}_i - y) > (\mathrm{p}_j + y)$ once more, we get

$$\begin{aligned}&\sum_{x\in\mathcal{X}} \frac{N_x^\alpha}{n^\alpha} \\ &= \sum_{x:N_x\geq t} \frac{N_x^\alpha}{n^\alpha} + \sum_{x:N_x<t} \frac{N_x^\alpha}{n^\alpha} \\ &\geq ck\left(\frac{1}{k} + c_2\sqrt{\frac{1}{nk}}\right)^\alpha + (1-c)k\left(\frac{1}{k} - \frac{cc_2}{1-c}\sqrt{\frac{1}{nk}}\right)^\alpha \\ &= \frac{1}{k^{\alpha-1}}\left[c\left(1 + c_2\sqrt{\frac{k}{n}}\right)^\alpha + (1-c)\left(1 - \frac{cc_2}{1-c}\sqrt{\frac{k}{n}}\right)^\alpha\right] \\ &\geq \frac{1}{k^{\alpha-1}}\left[c\left(1 + c_2\sqrt{\frac{k}{n}}\right)^\alpha + (1-c)\left(1 - \frac{\alpha cc_2}{1-c}\sqrt{\frac{k}{n}}\right)\right] \\ &\geq \frac{1}{k^{\alpha-1}}\left[c\left(1 + \alpha c_2\sqrt{\frac{k}{n}} + c_4\frac{k}{n}\right)\right. \\ &\quad \left. + (1-c)\left(1 - \frac{\alpha cc_2}{1-c}\sqrt{\frac{k}{n}}\right)\right] \\ &= \frac{1}{k^{\alpha-1}}\left(1 + cc_4\frac{k}{n}\right)\end{aligned}$$

where the second inequality is by Bernoulli's inequality and the third inequality holds for every $c_4 \leq \alpha(\alpha-1)(c_2\sqrt{c_1})^{\alpha-2}/2$. Therefore, with probability ≥ 0.9,

$$H_\alpha(\mathrm{p}) - \frac{1}{\alpha-1}\log\frac{1}{\widehat{P}_\alpha^{\mathrm{e}}} \geq \frac{1}{\alpha-1}\log\left(1 + cc_4\frac{k}{n}\right),$$

which yields the desired bound. ∎

Theorem 24. *Given $0 < \alpha < 1$, there exists a constant c_α such that for every δ sufficiently small (depending only on α) and every k*

$$S_\alpha^{f_\alpha^e}(k, \delta, 0.9) \geq c_\alpha\left(\frac{k^{1/\alpha}}{\delta^{1/\alpha}}\right).$$

Proof. We proceed as in the proof of the previous lemma. However, instead of using the uniform distribution, we use a distribution which has one "heavy element" and is uniform conditioned on the occurrence of the remainder. The key observation is that there will be roughly n^α occurrences of the "light elements". Thus, when we account for the error in the estimation of the contribution of light elements to the power sum, we can replace n with $n^{1/\alpha}$ in our analysis of the previous lemma, which yields the required bound for sample complexity.

Specifically, consider a distribution with one heavy element 0 such that

$$\mathrm{p}_0 = 1 - \frac{\delta}{n^{1-\alpha}}, \quad \text{and} \quad \mathrm{p}_i = \frac{\delta}{kn^{1-\alpha}}, \quad 1 \leq i \leq k.$$

Thus,

$$P_\alpha(\mathrm{p}) = \left(1 - \frac{\delta}{n^{1-\alpha}}\right)^\alpha + \delta^\alpha\left(\frac{k}{n^\alpha}\right)^{1-\alpha}. \tag{26}$$

We begin by analyzing the estimate of the second term

in power sum, namely

$$\sum_{i\in[k]} \left(\frac{N_i}{n}\right)^{\alpha}.$$

Let $R = \sum_{i\in[k]} N_i$ be the total number of occurrences of light elements. Since R is a binomial $(n, \delta n^{\alpha-1})$ random variable, for every constant $c > 0$

$$\mathbb{P}\left(1-c < \frac{R}{\delta n^{\alpha}} < 1+c\right) \geq 1 - \frac{1}{c^2 n}.$$

In the remainder of the proof, we shall assume that this large probability event holds.

As in the proof of the previous lemma, we first prove a δ independent lower bound for sample complexity. To that end, we fix $\delta = 1$ in the definition of p. Assuming $(1+c)n^{\alpha} \leq k$, which implies $R \leq k$, and using the fact that $(\mathrm{p}_i - y)^{\alpha} - (\mathrm{p}_j + y)^{\alpha}$ is increasing in y if $(\mathrm{p}_i - y) > (\mathrm{p}_j + y)$, we get

$$\begin{aligned}
\widehat{P}^{\mathrm{e}}_{\alpha} &\leq 1 + \left(\frac{R}{n}\right)^{\alpha} \sum_{i\in[k]} \left(\frac{N_i}{R}\right)^{\alpha} \\
&\leq 1 + \frac{(1+c)^{\alpha}}{n^{\alpha(1-\alpha)}} \sum_{i\in[k]} \left(\frac{N_i}{R}\right)^{\alpha} \\
&\leq 1 + \frac{(1+c)^{\alpha}}{n^{\alpha(1-\alpha)}} R^{1-\alpha} \\
&\leq 3,
\end{aligned}$$

where the last inequality uses $R \leq (1+c)n^{\alpha} \leq 2n^{\alpha}$. Thus, the empirical estimate is at most 3 with probability close to 1 when k (and therefore n) large. It follows from (26) that

$$H_{\alpha}(\mathrm{p}) - \frac{1}{1-\alpha} \log \widehat{P}^{\mathrm{e}}_{\alpha} \geq \log \frac{k}{3n^{\alpha}}.$$

Therefore, for all $c_1 > 1$, $\delta < \log 3c_1$ and k sufficiently large, at least $(k/c_1)^{1/\alpha}$ samples are needed to get a δ-additive approximation of $H_{\alpha}(\mathrm{p})$ with probability of error less than $1 - 1/(c^2 n)$. Note that we only needed to assume $R \leq (10/9)n^{\alpha}$, an event with probability greater than 0.9, to get the contradiction above. Thus, we may assume that $n \geq (k/c_1)^{1/\alpha}$. Under this assumption, for k sufficiently large, n is sufficiently large so that $(1-c)n^{\alpha} \leq R \leq (1+c)n^{\alpha}$ holds with probability arbitrarily close to 1.

Next, assuming that $n \geq (k/c_1)^{1/\alpha}$, we obtain a δ-dependent lower bound for sample complexity of the empirical estimator. We use the p mentioned above with a general δ and assume that the large probability event

$$(1-c) \leq \frac{R}{\delta n^{\alpha}} \leq (1+c) \tag{27}$$

holds. Note that conditioned on each value of R, the random variables $(N_i, i \in [k])$ have a multinomial distribution with uniform probabilities, *i.e.*, these random variables behave as if we drew R i.i.d. samples from a uniform distribution on $[k]$ elements. Thus, we can follow the proof of the previous lemma *mutatis mutandis*. We now define A as

$$A = \sum_{x} \mathbb{1}\left(N_x \leq \frac{n}{k} - c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right).$$

and satisfies Then,

$$\begin{aligned}
\mathbb{E}[A\,] &= \sum_{x} \mathbb{E}\left[\mathbb{1}\left(N_x \leq \frac{n}{k} - c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right)\right] \\
&= k \cdot p\left(N_x \leq \frac{n}{k} - c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right).
\end{aligned}$$

To lower bound $p\left(N_x \leq \frac{n}{k} - c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right)$, Slud's inequality is no longer available (since it may not hold for $\mathtt{Bin}(n,p)$ with $p > 1/2$ and that is the regime of interest for the lower tail probability bounds needed here). Instead we take recourse to a combination of Bohman's inequality and Anderson-Samuel inequality, as suggested in [38, Eqns. (i) and (ii)]. It can be verified that the condition for [38, Eqns. (ii)] holds, and therefore,

$$p\left(N_x \leq \frac{n}{k} - c_2\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}\right) \geq Q(c_2).$$

Continuing as in the proof of the previous lemma, we get that the following holds with conditional probability greater than 0.9 given each value of R satisfying (27):

$$\begin{aligned}
\sum_{i\in[k]} \left(\frac{N_x}{R}\right)^{\alpha} &\leq k^{1-\alpha}\left(1 - c_3\frac{k}{R}\right) \\
&\leq k^{1-\alpha}\left(1 - c_4\frac{k}{\delta n^{\alpha}}\right),
\end{aligned}$$

where c_3 is a sufficiently small constant such that $(1+x)^{\alpha} \leq 1 + \alpha x - c_3 x^2$ for all $x \geq 0$ and $c_4 = c_3/(1+c)$. Thus,

$$\begin{aligned}
\widehat{P}^{\mathrm{e}}_{\alpha} &\leq 1 + \left(\frac{R}{n}\right)^{\alpha} \sum_{i\in[k]} \left(\frac{N_i}{R}\right)^{\alpha} \\
&\leq 1 + \left(\frac{R}{n}\right)^{\alpha} k^{1-\alpha}\left(1 - c_4\frac{k}{\delta n^{\alpha}}\right) \\
&\leq 1 + (1+c)^{\alpha}\,\delta^{\alpha}\left(\frac{k}{n^{\alpha}}\right)^{1-\alpha}\left(1 - c_4\frac{k}{\delta n^{\alpha}}\right).
\end{aligned}$$

Denoting $y = (k/n^{\alpha})$ and choosing c_1 and c small enough such that $\widehat{P}^{\mathrm{e}}_{\alpha} \leq 2$, for all sufficiently large n we get from (26) that

$$\begin{aligned}
\frac{P_{\alpha}(\mathrm{p})}{\widehat{P}^{\mathrm{e}}_{\alpha}} &\geq \frac{1 - \delta + y^{1-\alpha}}{1 + \delta^{\alpha}(1+c)^{\alpha} y^{1-\alpha} - (1+c)^{\alpha} c_4 \delta^{\alpha-1} y^{2-\alpha}} \\
&\geq \frac{1 - \delta + y^{1-\alpha}}{1 + y^{1-\alpha} - \delta^{\alpha-1} y^{2-\alpha}} \\
&\geq 1 - \frac{\delta}{2} + \frac{\delta^{\alpha-1} y^{2-\alpha}}{2},
\end{aligned}$$

where the second inequality uses the fact that $\delta^{\alpha}(1+c)^{\alpha} y^{1-\alpha} - (1+c)^{\alpha} c_4 \delta^{\alpha-1} y^{2-\alpha}$ is negative, $c_4 > 1$ and

$\delta < 1$. Therefore, $\frac{P_\alpha(\mathrm{p})}{\widehat{P}^{\mathrm{e}}_\alpha} \geq 1+\delta$ if $y^{2-\alpha} \geq 3\delta^{2-\alpha}$, which completes the proof. ■

Note that for a fixed δ when k is sufficiently large depending on δ or equivalently δ is greater than a function of k, the lower bounds in Theorem 23 and 24 match (up to constants) the upper bounds of Theorem 9 and 10, respectively. In fact, for $\alpha > 1$ the gap between the bounds can be fixed by using the general lower bound of Theorem 15, which constitutes the dominant lower bound for small δ. The resulting characterization of the sample complexity of the sample complexity of the empirical estimator for $\alpha > 1$ is summarized in the next corollary.

Corollary 25. *Given $\alpha > 1$, for every $\delta > 0$ sufficiently small (depending only on α) and every k,*

$$S_\alpha^{f_\alpha^e}(k, \delta, 0.9) = \Theta\left(\max\left\{\frac{k}{\delta}, \frac{k^{\frac{\alpha-1}{\alpha}}}{\delta^2}\right\}\right).$$

In particular, the sample complexity of the empirical estimator is optimal up to constant factors when $k \leq \delta^{-\alpha}$.

To obtain a similar characterization for $\alpha < 1$, we provide a companion result for Theorem 15.

Theorem 26. *Given an $0 < \alpha < 1$ and $0 < \epsilon < 1$, there exists a constant c depending on α and ϵ such that for every $\delta > 0$ sufficiently small (depending only on α) and every k sufficiently large (depending only on α)*

$$S_\alpha(k, \delta, \epsilon) \geq c\left(\frac{k^{\frac{1-\alpha}{\alpha}}}{\delta^2}\right).$$

The proof is very similar to that of Theorem 15 and is based applying Le Cam's two point method to the following distributions p and q over $[k]$: $\mathrm{p}_1 = 1/k^{\frac{1-\alpha}{\alpha}}$, and for $x = 2, \ldots, k$, $\mathrm{p}_x = (1-\mathrm{p}_1)/(k-1)$ and $\mathrm{q}_1 = (1+\delta)\mathrm{p}_1$, and for $x = 2, \ldots, k$, $\mathrm{q}_x = (1-\mathrm{q}_1)/(k-1)$; we omit the details. The following corollary is immediate by combining the previous result with the lower bound of Theorem 24 and the upper bound of Theorem 10.

Corollary 27. *Given $0 < \alpha < 1$, for every $\delta > 0$ sufficiently small (depending only on α) and every k,*

$$S_\alpha^{f_\alpha^e}(k, \delta, 0.9) = \Theta\left(\max\left\{\left(\frac{k}{\delta}\right)^{\frac{1}{\alpha}}, \frac{k^{\frac{1-\alpha}{\alpha}}}{\delta^2}\right\}\right).$$

In particular, the sample complexity of the empirical estimator is optimal up to constant factors when $1/2 \leq \alpha < 1$ and $k \leq \delta^{\frac{1-2\alpha}{\alpha}}$.

Acknowledgements

The authors thank Chinmay Hegde and Piotr Indyk for helpful discussions and suggestions. The authors are also indebted to an anonymous reviewer for providing many useful suggestions to extend the scope of this paper. In an earlier submission we had focused on the large alphabet regime with a fixed δ and large k. The reviewer suggested to include results for arbitrary k and δ whenever possible and provided a proof sketch for characterizing the sample complexity of empirical estimator. In particular, the current form of Lemma 4 is a strengthening of our original result and was suggested by the anonymous reviewer. This stronger form was essential in obtaining a characterization of sample complexity of empirical estimator for arbitrary k and δ. We thank the anonymous reviewer for providing a very thorough review and suggesting many ideas for extending this work.

Appendix: Estimating power sums

The broader problem of estimating smooth functionals of distributions was considered in [41]. Independently and concurrently with this work, [16] considered estimating more general functionals and applied their technique to estimating the power sums of a distribution to a given additive accuracy. Letting $S_\alpha^{P+}(k)$ denote the number of samples needed to estimate $P_\alpha(\mathrm{p})$ to a given additive accuracy, [16] showed that for $\alpha < 1$,

$$S_\alpha^{P+}(k) = \Theta\left(\frac{k^{1/\alpha}}{\log k}\right), \tag{28}$$

and [17] showed that for $1 < \alpha < 2$,

$$S_\alpha^{P+}(k) \leq O\left(k^{2/\alpha - 1}\right).$$

In fact, using techniques similar to multiplicative guarantees on $P_\alpha(\mathrm{p})$ we show that for $S_\alpha^{P+}(k)$ is a constant independent of k for all $k > 1$.

Since $P_\alpha(\mathrm{p}) > 1$ for $\alpha < 1$, power sum estimation to a fixed additive accuracy implies also a fixed multiplicative accuracy, and therefore

$$S_\alpha(k) = \Theta(S_\alpha^{P\times}(k)) \leq O(S_\alpha^{P+}(k)),$$

namely for estimation to an additive accuracy, Rényi entropy requires fewer samples than power sums. Similarly, $P_\alpha(\mathrm{p}) < 1$ for $\alpha > 1$, and therefore

$$S_\alpha(k) = \Theta(S_\alpha^{P\times}(k)) \geq \Omega(S_\alpha^{P+}(k)),$$

namely for an additive accuracy in this range, Rényi entropy requires more samples than power sums.

It follows that the power sum estimation results in [16], [17] and the Rényi-entropy estimation results in this paper complement each other in several ways. For example, for $\alpha < 1$,

$$\begin{aligned}\widetilde{\widetilde{\Omega}}\left(k^{1/\alpha}\right) \leq S_\alpha(k) = \Theta(S_\alpha^{P\times}(k)) \leq & O(S_\alpha^{P+}(k)) \\ \leq & O\left(\frac{k^{1/\alpha}}{\log k}\right),\end{aligned}$$

where the first inequality follows from Theorem 22 and the last follows from the upper-bound (28) derived in [16] using a *polynomial approximation estimator*. Hence, for $\alpha < 1$, estimating power sums to additive and

multiplicative accuracy require a comparable number of samples.

On the other hand, for $\alpha > 1$, Theorems 9 and 21 imply that for non integer α, $\widetilde{\widetilde{\Omega}}(k) \leq S_\alpha^{P\times}(k) \leq O(k)$, while in the Appendix we show that for $1 < \alpha$, $S_\alpha^{P+}(k)$ is a constant. Hence in this range, power sum estimation to a multiplicative accuracy requires considerably more samples than estimation to an additive accuracy.

We now show that the empirical estimator requires a constant number of samples to estimate $P_\alpha(\mathrm{p})$ independent of k, *i.e.*, $S_\alpha^{P+}(k) = O(1)$. In view of Lemma 8, it suffices to bound the bias and variance of the empirical estimator. Concurrently with this work, similar results were obtained in an updated version of [16].

As before, we consider Poisson sampling with $N \sim \mathrm{Poi}(n)$ samples. The *empirical* or *plug-in* estimator of $P_\alpha(\mathrm{p})$ is

$$\widehat{P}_\alpha^{\mathrm{e}} \stackrel{\mathrm{def}}{=} \sum_x \left(\frac{N_x}{n}\right)^\alpha.$$

The next result shows that the bias and the variance of the empirical estimator are $o(1)$.

Lemma 28. *For an appropriately chosen constant $c > 0$, the bias and the variance of the empirical estimator are bounded above as*

$$\left|\widehat{P}_\alpha^e - P_\alpha(\mathrm{p})\right| \leq 2c\max\{n^{-(\alpha-1)}, n^{-1/2}\},$$
$$\mathbb{V}\mathrm{ar}[\widehat{\mathrm{P}}_\alpha] \leq 2c\max\{n^{-(2\alpha-1)}, n^{-1/2}\},$$

for all $n \geq 1$.

Proof. Denoting $\lambda_x = n\mathrm{p}_x$, we get the following bound on the bias for an appropriately chosen constant c:

$$\begin{aligned}&\left|\widehat{P}_\alpha^{\mathrm{e}} - P_\alpha(\mathrm{p})\right| \\ &\leq \frac{1}{n^\alpha}\sum_{\lambda_x\leq 1}|\mathbb{E}[N_x^\alpha] - \lambda_x| + \frac{1}{n^\alpha}\sum_{\lambda_x>1}|\mathbb{E}[N_x^\alpha] - \lambda_x| \\ &\leq \frac{c}{n^\alpha}\sum_{\lambda_x\leq 1}\lambda_x + \frac{c}{n^\alpha}\sum_{\lambda_x>1}\left(\lambda_x + \lambda_x^{\alpha-1/2}\right),\end{aligned}$$

where the last inequality holds by Lemma 4 and Lemma 2since x^α is convex in x. Noting $\sum_i \lambda_x = n$, we get

$$\left|\widehat{P}_\alpha^{\mathrm{e}} - P_\alpha(\mathrm{p})\right| \leq \frac{c}{n^{\alpha-1}} + \frac{c}{n^\alpha}\sum_{\lambda_x>1}\lambda_x^{\alpha-1/2}.$$

Similarly, proceeding as in the proof of Theorem 9, the variance of the empirical estimator is bounded as

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\widehat{\mathrm{P}}_\alpha] &= \frac{1}{n^{2\alpha}}\sum_{x\in\mathcal{X}}\mathbb{E}\left[N_x^{2\alpha}\right] - \mathbb{E}[N_x^\alpha]^2 \\ &\leq \frac{1}{n^{2\alpha}}\sum_{x\in\mathcal{X}}\left|\mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha}\right| \\ &\leq \frac{c}{n^{2\alpha-1}} + \frac{c}{n^{2\alpha}}\sum_{\lambda_x>1}\lambda_x^{2\alpha-1/2}.\end{aligned}$$

The proof is completed upon showing that

$$\sum_{\lambda_x>1}\lambda_x^{\alpha-1/2} \leq \max\{n, n^{\alpha-1/2}\}, \quad \alpha > 1.$$

To that end, note that for $\alpha < 3/2$

$$\sum_{\lambda_x>1}\lambda_x^{\alpha-1/2} \leq \sum_{\lambda_x>1}\lambda_x \leq n, \quad \alpha < 3/2.$$

Further, since $x^{\alpha-1/2}$ is convex for $\alpha \geq 3/2$, the summation above is maximized when one of the λ_x's is n and the remaining equal 0 which yields

$$\sum_{\lambda_x>1}\lambda_x^{\alpha-1/2} \leq n^{\alpha-1/2}, \quad \alpha \geq 3/2,$$

and completes the proof. ∎

References

[1] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating rényi entropy," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015, pp. 1855–1869.

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

[3] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 99–105, 1996.

[4] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, "Sampling algorithms: lower bounds and applications," in *Proceedings on 33rd Annual ACM Symposium on Theory of Computing, July 6-8, 2001, Heraklion, Crete, Greece*, 2001, pp. 266–275.

[5] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing closeness of discrete distributions," *J. ACM*, vol. 60, no. 1, p. 4, 2013.

[6] C. Bennett, G. Brassard, C. Crepeau, and U. Maurer, "Generalized privacy amplification," *IEEE Transactions on Information Theory*, vol. 41, no. 6, Nov 1995.

[7] I. Csiszár, "Generalized cutoff rates and Renyi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.

[8] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 7, no. 20, 2000.

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

[10] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 70–78, 2011.

[11] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities. 2nd edition*. Cambridge University Press, 1952.

[12] N. J. A. Harvey, J. Nelson, and K. Onak, "Sketching and streaming entropy via approximation theory," in *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, 2008, pp. 489–498.

[13] V. M. Ilic and M. S. Stankovic, "A unified characterization of generalized information and certainty measures," *CoRR*, vol. abs/1310.4896, 2013. [Online]. Available: http://arxiv.org/abs/1310.4896

[14] R. Impagliazzo and D. Zuckerman, "How to recycle random bits," in *FOCS*, 1989.

[15] R. Jenssen, K. Hild, D. Erdogmus, J. Principe, and T. Eltoft, "Clustering using Renyi's entropy," in *Proceedings of the International Joint Conference on Neural Networks*. IEEE, 2003.

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

[17] J. Jiao, K. Venkat, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *CoRR*, vol. abs/1406.6959, 2014.

[18] D. Källberg, N. Leonenko, and O. Seleznjev, "Statistical inference for rényi entropy functionals," *CoRR*, vol. abs/1103.4977, 2011.
[19] D. E. Knuth, *The Art of Computer Programming, Volume III: Sorting and Searching*. Addison-Wesley, 1973.
[20] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang, "Data streaming algorithms for estimating entropy of network traffic," *SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 1, pp. 145–156, Jun. 2006.
[21] B. Ma, A. O. H. III, J. D. Gorman, and O. J. J. Michel, "Image registration with minimum spanning tree algorithm," in *ICIP*, 2000, pp. 481–484.
[22] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Multiple source adaptation and the Renyi divergence," *CoRR*, vol. abs/1205.2628, 2012.
[23] V. Markov, "On functions deviating least from zero in a given interval," *Izdat. Imp. Akad. Nauk, St. Petersburg*, pp. 218–258, 1892.
[24] J. Massey, "Guessing and entropy," in *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, Jun 1994, pp. 204–.
[25] G. A. Miller, "Note on the bias of information estimates," *Information theory in psychology: Problems and methods*, vol. 2, pp. 95–100, 1955.
[26] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 193–196, 1989.
[27] A. Motahari, G. Bresler, and D. Tse, "Information theory of dna shotgun sequencing," *Information Theory, IEEE Transactions on*, vol. 59, no. 10, pp. 6273–6289, Oct 2013.
[28] H. Neemuchwala, A. O. Hero, S. Z., and P. L. Carson, "Image registration methods in high-dimensional space," *Int. J. Imaging Systems and Technology*, vol. 16, no. 5, pp. 130–145, 2006.
[29] I. Nemenman, W. Bialek, and R. R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Physical Review E*, vol. 69, pp. 056 111–056 111, 2004.
[30] P. C. V. Oorschot and M. J. Wiener, "Parallel collision search with cryptanalytic applications," *Journal of Cryptology*, vol. 12, pp. 1–28, 1999.
[31] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," 2004.
[32] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
[33] ——, "Estimating entropy on m bins given fewer than m samples," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2200–2203, 2004.
[34] ——, "A coincidence-based test for uniformity given very sparsely sampled discrete data," vol. 54, no. 10, pp. 4750–4755, 2008.
[35] C.-E. Pfister and W. Sullivan, "Renyi entropy, guesswork moments, and large deviations," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2794–2800, Nov 2004.
[36] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961, pp. 547–561.
[37] P. S. Shenkin, B. Erman, and L. D. Mastrandrea, "Information-theoretical entropy as a measure of sequence variability," *Proteins*, vol. 11, no. 4, pp. 297–313, 1991.
[38] E. V. Slud, "Distribution Inequalities for the Binomial Law," *The Annals of Probability*, vol. 5, no. 3, pp. 404–412, 1977.
[39] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, 1963.
[40] T. Tossavainen, "On the zeros of finite sums of exponential functions," *Australian Mathematical Society Gazette*, vol. 33, no. 1, pp. 47–50, 2006.
[41] G. Valiant and P. Valiant, "Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts," 2011.
[42] P. Valiant, "Testing symmetric properties of distributions," in *STOC*, 2008.
[43] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *CoRR*, vol. abs/1407.0381v1, 2014.
[44] D. Xu, "Energy, entropy and information potential for neural computation," Ph.D. dissertation, University of Florida, 1998.
[45] D. Xu and D. Erdogmuns, "Renyi's entropy, divergence and their nonparametric estimators," in *Information Theoretic Learning*, ser. Information Science and Statistics. Springer New York, 2010, pp. 47–102.
[46] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer New York, 1997, pp. 423–435.

Jayadev Acharya is an assistant professor in the School of Electrical and Computer Engineering (ECE) at Cornell University. He received the Bachelor of Technology (B. Tech) degree in Electronics and Electrical Communication Engineering from Indian Institute of Technology (IIT), Kharagpur in 2007, and the M.S. (2009) and Ph.D degree (2014) in Electrical and Computer Engineering (ECE) from University of California, San Diego (UCSD). He was a postdoctoral associate in Electrical Engineering and Computer Science at MIT from 2014-16.

Alon Orlitsky received B.Sc. degrees in Mathematics and Electrical Engineering from Ben Gurion University in 1980 and 1981, and M.Sc. and Ph.D. degrees in Electrical Engineering from Stanford University in 1982 and 1986.

From 1986 to 1996 he was with the Communications Analysis Research Department of Bell Laboratories. He spent the following year as a quantitative analyst at D.E. Shaw and Company, an investment firm in New York city. In 1997 he joined the University of California San Diego, where he is currently a professor of Electrical and Computer Engineering and of Computer Science and Engineering. His research concerns information theory, statistical modeling, and machine learning.

From 2011 to 2014 Alon directed UCSD's Center for Wireless Communications, and since 2006 he has directed the Information Theory and Applications Center. He is currently the president of the Information Theory Society. He has co-organized numerous programs on information theory, machine learning, and statistics, including the Information Theory and Applications Workshop that he started in 2006 and has helped organize since.

Alon is a recipient of the 1981 ITT International Fellowship and the 1992 IEEE W.R.G. Baker Paper Award, and co-recipient of the 2006 Information Theory Society Paper Award and the 2016 NIPS Paper Award. He co-authored two papers for which his students received student-paper awards: the 2003 Capocelli Prize and the 2010 ISIT Student Paper Award. He is a fellow of the IEEE, and holds the Qucalcomm Chair for Information Theory and its Applications at UCSD.

Ananda Theertha Suresh received the B.Tech. degree from IIT Madras, Chennai in 2006, and the M.S. and Ph.D. degrees from the University of California at San Diego in 2012 and 2016, respectively. He is currently a Research Scientist at Google Research. His research interests lie in the intersection of statistics, machine learning, and information theory. He is a co-recipient of the 2015 Neural Information Processing Systems (NIPS) Best Paper Award.

Himanshu Tyagi received the Bachelor of Technology degree in electrical engineering and the Master of Technology degree in communication and information technology, both from the Indian Institute of Technology, Delhi, India in 2007. He received the Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park. From 2013 to 2014, he was a postdoctoral researcher at the Information Theory and Applications (ITA) Center, University of California, San Diego. Since January 2015, he has been an Assistant Professor at the Indian Institute of Science in Bangalore.