The Complexity of Estimating Rényi Entropy

Jayadev Acharya
EECS, MIT
jayadev@csail.mit.edu

Alon Orlitksy
ECE & CSE, UCSD
alon@ucsd.edu

Ananda Theertha Suresh
ECE, UCSD
asuresh@ucsd.edu

Himanshu Tyagi
ECE, IISc
htyagi@iisc.ernet.in

Abstract

It was recently shown 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 noninteger 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.

1 Introduction

1.1 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 [SEM91], quantifying neural activity [Pan03, NBdRvS04], network anomaly detection [LSO+06], 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[1] $\Theta(k/\log k)$ independent samples from p [Pan04, VV11]; see [JVW14b, WY14] 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 statistical 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 [Rén61] 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.* [IS13]).

Rényi entropy too has many applications. It is often used as a bound on Shannon entropy [Mok89, NBdRvS04, HNO08], and in many applications it replaces Shannon entropy as a measure of randomness [Csi95, Mas94, Ari96]. It is also of interest in its own right, with diverse applications to unsupervised learning [Xu98, JHE+03], source adaptation [MMR12], image registration [MIGM00, NHZC06], and password guessability [Ari96, PS04, HS11] among others. In particular, the Rényi entropy of order 2, $H_2(\mathrm{p})$, measures the quality of random number generators [Knu73, OW99], determines the number of unbiased bits that can be extracted from a physical source of randomness [IZ89, BBCM95], helps test graph expansion [GR00] and closeness of distributions [BFR+13, Pan08], and characterizes the number of reads needed to reconstruct a DNA sequence [MBT13].

Motivated by these applications, asymptotically consistent and normal estimates of Rényi entropy were proposed [XE10, KLS11]. 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.

1.2 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 superlinearly with k, for $1 < \alpha \notin \mathbb{Z}$ it grows almost

[1] $f(k) = \Theta(g(k))$ if there exist constants c and C such that $cg(k) \leq f(k) \leq Cg(k)$.

linearly with k, and most interestingly, for the popular orders $1 < \alpha \in \mathbb{Z}$ 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

$$S_\alpha^f(k,\delta,\epsilon) \overset{\text{def}}{=} \min \max_p \left\{ n : \mathrm{p}\left(|H_\alpha(\mathrm{p}) - f\left(X^n\right)| > \delta \right) < \epsilon \right\},$$

the minumum 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

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

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 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 $S_\alpha(k)$ in k for k large. 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 superlinearly in k and estimating the Rényi entropy of these orders is even more difficult than estimating Shannon entropy. In fact, the upper bound follows from a corresponding result on estimation of power sums considered in [JVW14b] (see Section 3.3 for further discussion). For completeness, we show in Theorem 10 that the empirical estimator requires $O(k^{1/\alpha})$ samples and in Theorem 13 prove the improvement by a factor of $\log k$. The lower bound is proved in Theorem 21.

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. The lower bound is proved in Theorem 20. In the conference version of this paper, a weaker $O(k)$ upper bound was established using the empirical-frequency estimator. For the sake of completeness, we include this result as Theorem 9. The tighter upper bound reported here uses the best polynomial approximation based estimator of [JVW14b, WY14] and is proved in Theorem 12.

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

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

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

and in particular, the sample complexity is *strictly sublinear* in the alphabet size. The upper and lower bounds are shown in Theorems 11 and 19, respectively.

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 therefore 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 using just $\Theta(\sqrt{k})$ 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.

Second, when estimating Shannon entropy using $\Theta(k/\log k)$ samples, the implicit constant factors are fairly high (in the orders of 10^6). For Rényi entropy of orders $\alpha = 2, 3, ...$, the constants implied by $\Theta(k^{1-1/\alpha})$ are shown to be small in Theorem 11. Furthermore, the experiments described below suggest that they may be even lower.

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

It should also be noted that the estimators achieving the upper bounds 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.

1.3 The estimators

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

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

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 $2^{\pm\delta\cdot(1-\alpha)}$.

We construct estimators for the power-sums of distributions with multiplicative-accuracy guarantees, and hence obtain additive-accuracy estimators for Rényi entropy. We consider the following three different estimators for different ranges of α and with varying performance guarantees.

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

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

	Empirical	**Bias-corrected**	**Polynomial**	**Lower bounds**
$\alpha < 1$	$O(k^{1/\alpha})$		$O(k^{1/\alpha}/\log k)$	$\Omega(k^{1/\alpha-\eta})$, for all $\eta > 0$
$\alpha > 1,\, \alpha \notin \mathbb{N}$	$O(k)$		$O(k/\log k)$	$\Omega(k^{1-\eta})$, for all $\eta > 0$
$\alpha > 1,\, \alpha \in \mathbb{N}$	$O(k)$	$O(k^{1-1/\alpha})$		$\Omega(k^{1-1/\alpha})$

Table 1: Performance of estimators and lower bounds for estimating Rényi entropy

For $\alpha \neq 1$, $\widehat{P}^{\mathrm{e}}_\alpha$ is a not an unbiased estimator of $P_\alpha(\mathrm{p})$. However, we prove in Theorem 10 that for $\alpha < 1$ the sample complexity of the empirical estimator is $O(k^{1/\alpha})$, and in Theorem 9 that for $\alpha > 1$ the complexity is $O(k)$.

Using the lower bounds in Section 4, we prove that the empirical estimator achieves the optimal exponent of k for all $\alpha \notin \mathbb{N}$.

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

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

where for integers N and $r > 0$, $N^{\underline{r}} \overset{\text{def}}{=} N(N-1)\ldots(N-r+1)$. A variation of this estimator was proposed first in [BKS01] for estimating moments of frequencies in a sequence using random samples drawn from it.

Theorem 11 show that for $1 < \alpha \in \mathbb{Z}$, $\widehat{P}^{\mathrm{u}}_\alpha$ estimates $P_\alpha(\mathrm{p})$ using $O(k^{1-1/\alpha})$ samples, and Theorem 19 shows that this number is optimal up to a constant factor.

Polynomial approximation estimator To obtain a logarithmic improvement in $S_\alpha(k)$, we consider the polynomial approximation estimator proposed in [WY14, JVW14b] for different problems, concurrently to an earlier version of this paper. The polynomial approximation estimator first considers the *best polynomial approximation* of degree d to y^α for the interval $y \in [0,1]$ [Tim63]. 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. Suppose N'_x and N_x be the multiplicities of x in the first and second parts respectively. The polynomial approximation estimator uses a polynomial for small N'_x and the empirical estimate for large N'_x.

The estimator is rougly of the form

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

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 $\widehat{P}^{d,\tau}_\alpha$ is $O(k/\log k)$ and $O(k^{\frac{1}{\alpha}}/\log k)$, resulting in a reduction in sample complexity of $O(\log k)$ over the empirical estimator.

Table 1 summarizes the performance of these estimators in terms of their sample complexity. The last column denote the lower bounds from Section 4.

1.4 Examples and experiments

We demonstrate the performance of the estimators for two popular distributions, uniform and Zipf. For each, we determine the Rényi entropy of any order and illustrate the performance for integer and noninteger orders by showing that estimating Rényi entropy of order 2 requires only a small multiple of $\sqrt{k}$ samples, while for order 1.5 the estimators require nearly k samples.

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

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

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

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

Figure 1 shows the performance of the bias-corrected and the empirical estimators for samples drawn from a uniform distribution. ■

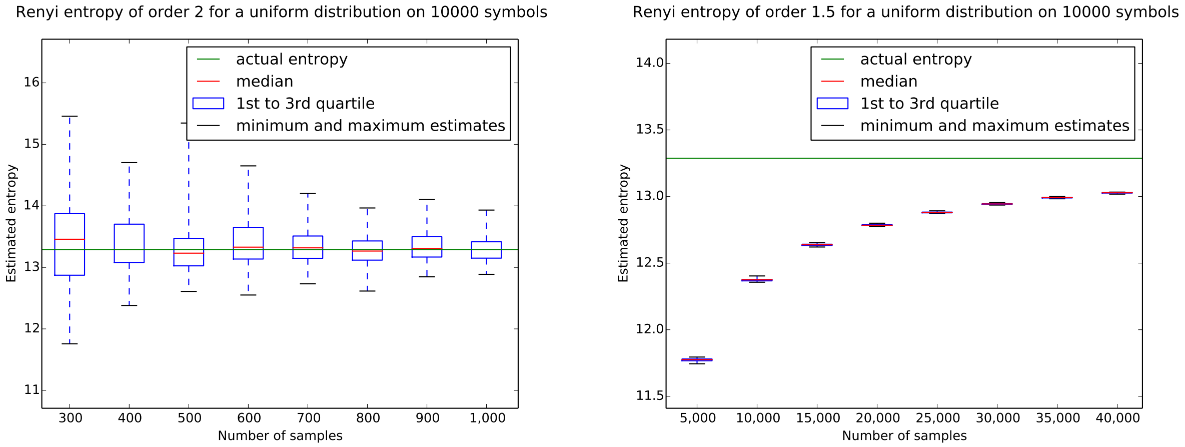

Figure 1: Estimation of Rényi entropy of order 2 and order 1.5 using the bias-corrected estimator and empirical estimator, respectively, for samples drawn from a uniform distribution. The boxplots display the estimated values for 100 independent experiments.

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

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

Its Rényi entropy of order $\alpha \neq 1$ 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 2 summarizes the leading term $g(k)$ in the approximation[3] $H_\alpha(Z_{\beta,k}) \sim g(k)$.

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

Table 2: 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 $O\left(\epsilon \log \log k\right)$. Therefore, even for very small ϵ this difference is unbounded and approaches infinity in the limit as k goes to infinity. Figure 2 shows the performance of our estimators for samples drawn from $Z_{1,k}$. ■

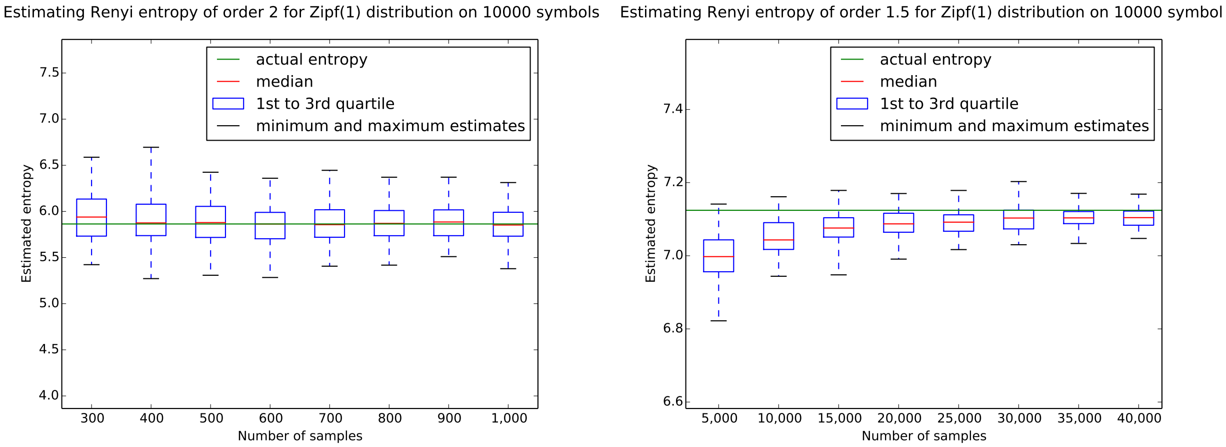

Figure 2: Estimation of Rényi entropy of order 2 and order 1.5 using the bias-corrected estimator and empirical estimator, respectively, for samples drawn from $Z_{1,k}$. The boxplots display the estimated values for 100 independent experiments.

Figures 1 and 2 above illustrate the estimation of Rényi entropy for $\alpha = 2$ and $\alpha = 1.5$ using the empirical and the bias-corrected estimators, respectively. As expected, for $\alpha = 2$ the estimation works quite well for $n = \sqrt{k}$ and requires roughly k samples to work well for $\alpha = 1.5$. Note that the empirical estimator is negatively biased for $\alpha > 1$ and the figures above confirm this. Our goal in this work is to find the exponent of k in $S_\alpha(k)$, and as our results show, for noninteger α the empirical estimator attains the optimal exponent; we do not consider the possible improvement in performance by reducing the bias in the empirical estimator.

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

1.5 Organization

The rest of the paper is organized as follows. Section 2 presents basic properties of power sums of distributions and moments of Poisson random variables, which may be of independent interest. The estimation algorithms are analyzed in Section 3, in Section 3.1 we show results on the empirical or plug-in estimate, in Section 3.2 we provide optimal results for integral α and finally we provide an improved estimator for non-integral $\alpha > 1$. Finally, the lower bounds on the sample complexity of estimating Rényi entropy are established in Section 4.

2 Technical preliminaries

2.1 Bounds on power sums

Consider a distribution p over $[k] = \{1, \ldots, k\}$. Since Rényi entropy is a measure of randomness (see [Rén61] 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

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

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, [HLP52]).

Lemma 1. *For every* $0 \leq \alpha$,

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

Further, for $\alpha > 1$ *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. By the monotonicity of norms,

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

which gives

$$\frac{P_{\alpha+\beta}(\mathrm{p})}{P_\alpha(\mathrm{p})^2} \leq P_\alpha(\mathrm{p})^{\frac{\beta}{\alpha}-1}.$$

The first inequality follows upon choosing $\beta = \alpha$. For $1 < \alpha$ and $0 \leq \beta \leq \alpha$, we get the second by (4). For the final inequality, note that by the monotonicity of Hölder means, we have

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

The final inequality follows upon rearranging the terms and using (4). ■

2.2 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 $r \in \mathbb{N}$*

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

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

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

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,

$$\mathbb{V}\mathrm{ar}[X^{\underline{r}}] = \mathbb{E}\big[(X^{\underline{r}})^2\big] - [\mathbb{E}\,X^{\underline{r}}]^2 \le \lambda^r \cdot ((\lambda+r)^r - \lambda^r)\,. \qquad \blacksquare$$

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

Lemma 3. *Let $X \sim \mathrm{Poi}(\lambda)$ and let β be a positive real number. Then,*

$$\mathbb{E}\Big[X^\beta\Big] \le 2^{\beta+2}\max\{\lambda, \lambda^\beta\}.$$

Proof. Let $Z = \max\{\lambda^{1/\beta}, \lambda\}$.

$$\begin{aligned}
\mathbb{E}\left[\frac{X^\beta}{Z^\beta}\right] &\le \mathbb{E}\left[\left(\frac{X}{Z}\right)^{\lceil\beta\rceil} + \left(\frac{X}{Z}\right)^{\lfloor\beta\rfloor}\right] \\
&= \sum_{i=1}^{\lceil\beta\rceil}\left(\frac{\lambda}{Z}\right)^{\lceil\beta\rceil}\binom{\lceil\beta\rceil}{i} + \sum_{i=1}^{\lfloor\beta\rfloor}\left(\frac{\lambda}{Z}\right)^{\lfloor\beta\rfloor}\binom{\lfloor\beta\rfloor}{i} \\
&\le \sum_{i=1}^{\lceil\beta\rceil}\binom{\lceil\beta\rceil}{i} + \sum_{i=1}^{\lfloor\beta\rfloor}\binom{\lfloor\beta\rfloor}{i} \\
&\le 2^{\beta+2}.
\end{aligned}$$

The first inequality follows from the fact that either $X/Z > 1$ or ≤ 1. The second inequality uses the property that $\lambda/Z \le 1$. Multiplying both sides by Z^β results in the lemma. $\blacksquare$

We close this section with bounds on $|\mathbb{E}[X^\alpha] - \lambda^\alpha|$, which will be used in the next section to bound the bias of the empirical estimator.

Lemma 4. *For $X \sim \mathrm{Poi}(\lambda)$,*

$$|\mathbb{E}[X^\alpha] - \lambda^\alpha| \le \begin{cases} \alpha\left(2^\alpha\lambda + (2^\alpha+1)\lambda^{\alpha-1/2}\right) & \alpha > 1 \\ \min(\lambda^\alpha, \lambda^{\alpha-1}) & \alpha \le 1. \end{cases}$$

Proof. 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,

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

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

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

For $\alpha > 1$,

$$|x^\alpha - y^\alpha| \le \alpha |x-y| \left(x^{\alpha-1} + y^{\alpha-1}\right),$$

hence by the Cauchy-Schwarz Inequality,

$$\begin{aligned}\mathbb{E}[|X^\alpha - \lambda^\alpha|\,] &\le \alpha \mathbb{E}\big[|X-\lambda|\left(X^{\alpha-1} + \lambda^{\alpha-1}\right)\big] \\ &\le \alpha\sqrt{\mathbb{E}[(X-\lambda)^2\,]}\sqrt{\mathbb{E}[(X^{2\alpha-2} + \lambda^{2\alpha-2})\,]} \\ &\le \alpha\sqrt{\lambda}\sqrt{\mathbb{E}[(X^{2\alpha-2} + \lambda^{2\alpha-2})\,]} \\ &\le \alpha\sqrt{2^{2\alpha}\max\{\lambda^2, \lambda^{2\alpha-1}\} + \lambda^{2\alpha-1}} \\ &\le \alpha\left(2^\alpha \max\{\lambda, \lambda^{\alpha-1/2}\} + \lambda^{\alpha-1/2}\right),\end{aligned}$$

where the last-but-one inequality is by Lemma 3. ■

2.3 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

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

Lemma 5 ([Tim63])**.** *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

$$\begin{aligned}\max_{x\in[0,1]} |q_\alpha(x) - x^\alpha| &\le |q_0| + \max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &= |q'_\alpha(0) - 0^\alpha| + \max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &\le 2\max_{x\in[0,1]} |q'_\alpha(x) - x^\alpha| \\ &= \frac{2c'_\alpha}{d^{2\alpha}} \\ &\overset{\text{def}}{=} \frac{c_\alpha}{d^{2\alpha}}. \end{aligned} \tag{5}$$

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 ([Mar92])**.** *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 $|x^\alpha| \leq 1$ for $x \in [0,1]$, the approximation bound (5) implies $|q_\alpha(x)| < 1 + \frac{c_\alpha}{d^{2\alpha}}$ for all $x \in [0,1]$. It follows from Lemma 6 that

$$\max_m |a_m| < \left(1 + \frac{c_\alpha}{d^{2\alpha}}\right)(\sqrt{2}+1)^d. \tag{6}$$

3 Upper bounds on sample complexity

In this section, we analyze the performances of the estimators we proposed in Section 1.3. 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 $X_1, ..., X_n$ 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 Tand the sample size. Here $\widehat{\mathrm{P}}_\alpha(n, X^n)$ is the corresponding estimate of $P_\alpha(\mathrm{p})$ – as discussed in Section 1, 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:

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

The following reduction to Poisson sampling is well-known.

Lemma 7. (Poisson approximation 1) *For $n \geq 8 \log(2/\epsilon)$ and $N \sim \mathrm{Poi}(n/2)$,*

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

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*

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

Then, there exists an estimator $\widehat{\mathrm{P}}'_\alpha$ that uses $18n \log(1/\epsilon)$ samples and ensures

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

Proof. By Chebychev's Inequality

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

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.

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 $\mathbb{E}[\mathbb{1}_{\mathcal{E}_i}] \le 1/3$ and hence by Hoeffding's inequality

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

The claimed bound follows 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. ■

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_α^{e}, f_α^{u}, and $f_\alpha^{d,\tau}$, respectively, the empirical estimator $\frac{1}{1-\alpha}\log \widehat{P}_\alpha^{\mathrm{e}}$, the bias-corrected estimator $\frac{1}{1-\alpha}\log \widehat{P}_\alpha^{\mathrm{u}}$, and the polynomial approximation estimator $\frac{1}{1-\alpha}\log \widehat{P}_\alpha^{d,\tau}$. We begin by analyzing the performances of f_α^{e} and f_α^{u} and build-up on these steps to analyze $f_\alpha^{d,\tau}$.

3.1 Performance of empirical estimator

The empirical estimator was presented in (1). We follow the recipe above of Poisson-sampling and then bound the expected value and variance of the estimator.

We bound the sample complexity of the empirical estimator for $\alpha > 1$ in Theorem 9 and $\alpha < 1$ in Theorem 10.

Theorem 9. *For $\alpha > 1$, $0 < \delta < 1/2$, and $0 < \epsilon < 1$, the estimator f_α^e satisfies*

$$S_\alpha^{f_\alpha^e}(k, \delta, \epsilon) \le O_\alpha\left(\frac{k}{\min(\delta^{1/(\alpha-1)}, \delta^2)}\log\frac{1}{\epsilon}\right),$$

for all k sufficiently large.

Proof. Denote $\lambda_x \overset{\text{def}}{=} n\mathrm{p}_x$. For $\alpha > 1$, we bound the bias of the power sum estimator using:

$$\begin{aligned}
\left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right| &\overset{(a)}{\le} \frac{1}{n^\alpha}\sum_x |\mathbb{E}[N_x^\alpha] - \lambda_x^\alpha| \\
&\overset{(b)}{\le} \frac{\alpha}{n^\alpha}\sum_x \left(2^\alpha \lambda_x + (2^\alpha+1)\lambda_x^{\alpha-1/2}\right) \\
&= \frac{\alpha 2^\alpha}{n^{\alpha-1}} + \frac{\alpha(2^\alpha+1)}{\sqrt{n}} P_{\alpha-1/2}(\mathrm{p}) \\
&\overset{(c)}{\le} \alpha\left(2^\alpha\left(\frac{k}{n}\right)^{\alpha-1} + (2^\alpha+1)\sqrt{\frac{k}{n}}\right) P_\alpha(\mathrm{p}) \\
&\le 2\alpha 2^\alpha \left[\left(\frac{k}{n}\right)^{\alpha-1} + \left(\frac{k}{n}\right)^{1/2}\right] P_\alpha(\mathrm{p}), \qquad (7)
\end{aligned}$$

where (a) is from the triangle inequality, (b) from Lemma 4, and (c) follows from Lemma 1 and (4). Thus, the bias of the estimator is less than $\delta(\alpha-1)P_\alpha(\mathrm{p})/2$ when

$$n \ge k \cdot \left(\frac{8\alpha 2^\alpha}{\delta(\alpha-1)}\right)^{\max(2,1/(\alpha-1))}.$$

Similarly, to bound the variance, using independence of multiplicities:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}} \sum_x \mathbb{V}\mathrm{ar}[N_x^\alpha] \\
&= \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - [\mathbb{E}N_x^\alpha]^2 \\
&\overset{(a)}{\leq} \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha} \\
&\leq \frac{1}{n^{2\alpha}} \sum_x \left|\mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha}\right| \\
&\leq \frac{2\alpha}{n^{2\alpha}} \sum_x \left(2^{2\alpha}\lambda_x + (2^{2\alpha}+1)\lambda_x^{2\alpha-1/2}\right) \qquad (8) \\
&= \frac{2\alpha 2^{2\alpha}}{n^{2\alpha-1}} + \frac{2\alpha(2^{2\alpha}+1)}{\sqrt{n}} P_{2\alpha-1/2}(\mathrm{p}) \\
&\overset{(c)}{\leq} 2\alpha 2^{2\alpha} \left(\frac{k}{n}\right)^{2\alpha-1} P_\alpha(\mathrm{p})^2 + 2\alpha(2^{2\alpha}+1) \left(\frac{k^{\frac{\alpha-1}{\alpha}}}{n}\right)^{1/2} P_\alpha(\mathrm{p})^2
\end{aligned}$$

(a) is from Jensen's inequality since z^α is convex and $\mathbb{E}[N_x] = \lambda_x$, (c) follows from Lemma 1. Thus, the variance is less than $\delta^2(\alpha-1)^2 P_\alpha(\mathrm{p})^2/12$ when

$$n \geq k \cdot \max\left(\left(\frac{48\alpha 2^{2\alpha}}{\delta^2(\alpha-1)^2}\right)^{1/(2\alpha-1)}, \left(\frac{96\alpha 2^{2\alpha}}{k^{1/2\alpha}\delta^2(\alpha-1)^2}\right)^2\right) = k \cdot \left(\frac{48\alpha 2^{2\alpha}}{\delta^2(\alpha-1)^2}\right)^{1/(2\alpha-1)},$$

where the equality holds for k sufficiently large. The theorem follows by using Lemma 8. ■

Theorem 10. *For $\alpha < 1$, $\delta > 0$, and $0 < \epsilon < 1$, the estimator f_α^e satisfies*

$$S_\alpha^{f_\alpha^e}(k, \delta, \epsilon) \leq O\left(\frac{k^{1/\alpha}}{\delta^{\max\{4, 2/\alpha\}}} \log\frac{1}{\epsilon}\right).$$

Proof. For $\alpha < 1$, once again we take a recourse to Lemma 4 to bound the bias as follows:

$$\begin{aligned}
\left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right| &\leq \frac{1}{n^\alpha} \sum_x \left|\mathbb{E}[N_x^\alpha] - \lambda_x^\alpha\right| \\
&\leq \frac{1}{n^\alpha} \sum_x \min\left(\lambda_x^\alpha, \lambda_x^{\alpha-1}\right) \\
&\leq \frac{1}{n^\alpha}\left[\sum_{x \notin A} \lambda_x^\alpha + \sum_{x \in A} \lambda_x^{\alpha-1}\right],
\end{aligned}$$

for every subset $A \subset [k]$. Upon choosing $A = \{x : \lambda_x \geq 1\}$, we get

$$\begin{aligned}
\left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right| &\leq 2\left(\frac{k^{1/\alpha}}{n}\right)^\alpha \\
&\leq 2P_\alpha(\mathrm{p})\left(\frac{k^{1/\alpha}}{n}\right)^\alpha, \qquad (9)
\end{aligned}$$

where the last inequality uses (4). For bounding the variance, note that

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{V}\text{ar}[N_x^\alpha] \\
&= \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\big[N_x^{2\alpha}\big] - [\mathbb{E}N_x^\alpha]^2 \\
&\le \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\big[N_x^{2\alpha}\big] - \lambda_x^{2\alpha} + \frac{1}{n^{2\alpha}}\sum_x \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2 . \qquad (10)
\end{aligned}$$

Consider the first term on the right-side. For $\alpha \le 1/2$, it is bounded above by 0 since $z^{2\alpha}$ is concave in z, and for $\alpha > 1/2$ the bound in (8) and Lemma 1 applies to give

$$\frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\big[N_x^{2\alpha}\big] - \lambda_x^{2\alpha} \le 2\alpha\left(\frac{c}{n^{2\alpha-1}} + (c+1)\sqrt{\frac{k}{n}}\right) P_\alpha(\mathrm{p})^2. \qquad (11)$$

For the second term, we have

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

where (a) is from (9) and (b) from the concavity of z^α in z. The proof is completed by combining the two bounds above and using Lemma 8. ■

3.2 Performance of bias-corrected estimator for integral α

To reduce the sample complexity for integer orders $\alpha > 1$ to below k we follow the path of the development of Shannon entropy estimators. Traditionally, Shannon entropy was estimated via an empirical estimator, analyzed in, for instance, [AK01]. However, with $o(k)$ samples, the bias of the empirical estimator remains high [Pan04]. This bias is reduced by the Miller-Madow correction [Mil55, Pan04], but even then, $O(k)$ samples are needed for a reliable Shannon-entropy estimation [Pan04].

We similarly 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 $1 < \alpha \in \mathbb{Z}$, $\widehat{P}_\alpha^{\mathrm{u}}$ estimates $P_\alpha(\mathrm{p})$ using $O(k^{1-1/\alpha})$ samples. Theorem 19 in Section 4 shows that this number is optimal up to constant factors.

Bias-corrected estimator Consider the unbiased esitmator for $P_\alpha(\mathrm{p})$ given by

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

which is unbiased since by Lemma 2,

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

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

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

Upper bound on sample complexity We bound the number of samples needed for the bias-corrected estimator. In particular:

Theorem 11. *For an integer $\alpha > 1$, any $\delta > 0$, and $0 < \epsilon < 1$, the estimator f^u_α satisfies*

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

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

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\frac{\sum_x N_x^{\underline{\alpha}}}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}} \sum_x \mathbb{V}\mathrm{ar}[N_x^{\underline{\alpha}}] \\
&\le \frac{1}{n^{2\alpha}} \sum_x \left(\lambda_x^\alpha (\lambda_x + \alpha)^\alpha - \lambda_x^{2\alpha}\right) \\
&= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} \sum_x \binom{\alpha}{r} \alpha^{\alpha-r} \lambda_x{}^{\alpha+r} \\
&= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} n^{\alpha+r} \binom{\alpha}{r} \alpha^{\alpha-r} P_{\alpha+r}(\mathrm{p}), \qquad (12)
\end{aligned}$$

where the inequality uses Lemma 2. It follows from Lemma 1 that

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

Applying Lemma 8 completes the proof. ■

3.3 The polynomial approximation estimator

Concurrently with an earlier version of this paper, a polynomial approximation based approach was proposed in [JVW14b] and [WY14] 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, [JVW14b] 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 $P_\alpha(\mathrm{p}) \geq 1$ 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 $P_\alpha(\mathrm{p}) \leq 1$ 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 [WY14] 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 $\alpha < 1$.

Polynomial approximation estimator 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 $(N_x/n)^\alpha$.

$N'_x \leq \tau/n$: Suppose $q(x) = \sum_{m=0}^{d} a_m x^m$ is the polynomial satisfying Lemma 5. We estimate p_x^α using an unbiased estimate of $(\tau/n)^\alpha q(n\mathrm{p}_x/\tau)$, 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

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

Denoting by $\hat{\mathrm{p}}_x$ the estimated probability of the symbol x, note that the polynomial approximation estimator relies on the empirical estimator for $\hat{\mathrm{p}}_x > \tau/n$ and the bias-corrected estimator for $\hat{\mathrm{p}}_x \leq \tau/n$.

Sample complexity of polynomial estimator We derive upper bounds for the sample complexity of the polynomial approximation estimator.

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*

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

Proof. We follow the approach in [WY14] 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.*,

$$\max_{x\in(0,1)} |q(x) - x^\alpha| < c_\alpha / d^{2\alpha} \tag{13}$$

To bound the bias of $\widehat{P}_\alpha^{d,\tau}$, note first that for $N'_x < \tau$

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{m=0}^{d} \frac{a_m (2\tau)^{\alpha-m} N_x^{\underline{m}}}{n^\alpha}\right] - \mathrm{p}_x^\alpha\right| &= \left|\sum_{m=0}^{d} a_m \left(\frac{2\tau}{n}\right)^{\alpha-m} \mathrm{p}_x^m - \mathrm{p}_x^\alpha\right| \\ &= \frac{(2\tau)^\alpha}{n^\alpha}\left|\sum_{m=0}^{d} a_m \left(\frac{n\mathrm{p}_x}{2\tau}\right)^m - \left(\frac{n\mathrm{p}_x}{2\tau}\right)^\alpha\right| \\ &= \frac{(2\tau)^\alpha}{n^\alpha}\left|q\left(\frac{n\mathrm{p}_x}{2\tau}\right) - \left(\frac{n\mathrm{p}_x}{2\tau}\right)^\alpha\right| \\ &< \frac{(2\tau)^\alpha c_\alpha}{(nd^2)^\alpha},\end{aligned} \tag{14}$$

where the last inequality uses (13) and $n\mathrm{p}_x/(2\tau) \le 1$, which holds under the assumption $N'_x < \tau$.

For $N'_x > \tau$, the bias of empirical part of the power sum is bounded as

$$\begin{aligned}\left|\mathbb{E}\left[\left(\frac{N_x}{n}\right)^\alpha\right] - \mathrm{p}_x^\alpha\right| &\overset{(a)}{\le} \alpha c \frac{\mathrm{p}_x}{n^{\alpha-1}} + \alpha(c+1)\frac{\mathrm{p}_x^{\alpha-\frac{1}{2}}}{\sqrt{n}} \\ &\overset{(b)}{\le} \alpha c \frac{\mathrm{p}_x^\alpha}{(\tau/2)^{\alpha-1}} + \alpha(c+1)\frac{\mathrm{p}_x^\alpha}{\sqrt{\tau/2}},\end{aligned}$$

and (a) is from Lemma 4 and (b) from $\mathrm{p}_x > \tau/(2n)$, which holds when $N'_x > \tau$. Thus, we obtain the following bound on the bias of $\widehat{P}_\alpha^{d,\tau}$:

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\mathrm{P}}_\alpha\right] - P_\alpha(\mathrm{p})\right| &\overset{(a)}{\le} \frac{k(2\tau)^\alpha c_\alpha}{(nd^2)^\alpha} + \alpha P_\alpha(\mathrm{p})\left[\frac{c}{(\tau/2)^{\alpha-1}} + \frac{c+1}{\sqrt{\tau/2}}\right] \\ &\overset{(b)}{\le} P_\alpha(\mathrm{p})\left[c_\alpha\left(\frac{k\cdot 2\tau}{nd^2}\right)^\alpha + \frac{\alpha c}{(\tau/2)^{\alpha-1}} + \frac{\alpha(c+1)}{\sqrt{\tau/2}}\right],\end{aligned} \tag{15}$$

where (a) is from the triangle inequality and (b) from (4).

For variance, independence of multiplicities under Poisson sampling gives

$$\mathbb{V}\mathrm{ar}\left[\widehat{\mathrm{P}}_\alpha\right] = \sum_{x:N'_x\le\tau} \mathbb{V}\mathrm{ar}\left(\sum_{m=0}^{d} \frac{a_m (2\tau)^{\alpha-m} N_x^{\underline{m}}}{n^\alpha}\right) + \sum_{x:N_x>\tau} \mathbb{V}\mathrm{ar}\left(\frac{N_x}{n}\right)^\alpha. \tag{16}$$

Let $a = \max_m |a_m|$. By Lemma 2, for any x with $\mathrm{p}_x \le 2\tau/n$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\sum_{m=0}^{d}\frac{a_m(2\tau)^{\alpha-m}N_x^{\underline{m}}}{n^\alpha}\right) &\le a^2d^2 \max_{1\le m\le d}\left\{\frac{(2\tau)^{2\alpha-2m}}{n^{2\alpha}}\mathbb{V}\mathrm{ar}N_x^{\underline{m}}\right\}\\ &\overset{(a)}{\le} a^2d^2 \max_{1\le m\le d}\left\{\frac{(2\tau)^{2\alpha-2m}}{n^{2\alpha}}(n\mathrm{p}_x)^m((n\mathrm{p}_x+m)^m - n\mathrm{p}_x^m)\right\}\\ &\overset{(b)}{\le} \frac{a^2d^2(2\tau+d)^{2\alpha}}{n^{2\alpha}},\end{aligned}\tag{17}$$

where (a) is from Lemma 2, and (b) from plugging $n\mathrm{p}_x \le 2\tau$. Furthermore, using similar steps as (8) together with Lemma 4, for x with $\mathrm{p}_x > \tau/(2n)$ we get

$$\mathbb{V}\mathrm{ar}\left[\left(\frac{N_x}{n}\right)^\alpha\right] \le 2\alpha c\frac{\mathrm{p}_x^{2\alpha}}{(\tau/2)^{2\alpha-1}} + 2\alpha(c+1)\frac{\mathrm{p}_x^{2\alpha}}{\sqrt{\tau/2}}.$$

The two bounds above along with Lemma 1 and (4) yield

$$\mathbb{V}\mathrm{ar}\left[\widehat{\mathrm{P}}_\alpha\right] \le P_\alpha(\mathrm{p})^2\left[\frac{a^2d^2(2\tau+d)^{2\alpha}}{n}\left(\frac{k}{n}\right)^{2\alpha-1} + \frac{2\alpha c}{(\tau/2)^{2\alpha-1}} + \frac{2\alpha(c+1)}{\sqrt{\tau/2}}\right].\tag{18}$$

For $d = \tau/8 = \frac{1}{2}\log n$, the last terms in (15) are $o(1)$ which gives

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

Recall from (6) 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 (18) we get

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

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 (14). For a symbol x with $N'_x > \tau$, the bias contribution of the empirical estimator is bounded as

$$\begin{aligned}\left|\mathbb{E}\left[\left(\frac{N_x}{n}\right)^\alpha\right] - \mathrm{p}_x^\alpha\right| &\overset{(a)}{\le} \frac{\mathrm{p}_x^{\alpha-1}}{n}\\ &\overset{(b)}{\le} \frac{2\mathrm{p}_x^\alpha}{\tau}\end{aligned}$$

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

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

where the last inequaliy is by (4).

To bound the variance, first note that bound (17) still holds for $\mathrm{p}_x \leq 2\tau/n$. To bound the contribution to the variance from the terms with $n\mathrm{p}_x > \tau/2$, we borrow steps from the proof of Theorem 10. In particular, (10) gives

$$\mathbb{V}\mathrm{ar}\left[\sum_{x:N'_x>\tau}\frac{N_x^\alpha}{n^\alpha}\right] \leq \frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha} + \frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\lambda_x^{2\alpha} - \left[\mathbb{E}N_x^\alpha\right]^2. \tag{19}$$

The first term can be bounded in the manner of (11) as

$$\frac{1}{n^{2\alpha}}\sum_{x:N'_x>\tau}\mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha} \leq 2\alpha\left(\frac{c}{n^{2\alpha-1}} + (c+1)\frac{1}{\sqrt{\tau/2}}\right)P_\alpha(\mathrm{p})^2,$$

For the second term, we have

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

where (a) follows from Lemma 4 and concavity of z^α in z and (b) 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 (17), the following bound for the variance:

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

Here a^2 is the largest squared coefficient of the approximating polynomial and, by (6), 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. ■

4 Lower bounds on sample complexity

We now establish lower bounds on $S_\alpha(k)$. The proof relies on the approach in [Val08] and 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.

As before, there is no loss in considering Poisson sampling.

Lemma 14. (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 $n > 4\log(2/\epsilon)$.

Next, denote by $\Phi = \Phi(X^N)$ the *profile* of X^N [OSVZ04], 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 15. (Sufficiency of profiles). *Consider an estimator $\hat{f}$ such that*

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

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

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

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. The result below facilitates such a contradiction.

Lemma 16. *If for two distributions p and q on $\mathcal{X}$ the variational distance over profiles satisfy $\|\mathrm{p} - \mathrm{q}\| < \epsilon$, then one of the following holds for every function $\hat{f}$:*

$$\begin{aligned} \mathrm{p}\left(|H_\alpha(\mathrm{p}) - \hat{f}(X)| \geq \frac{|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})|}{2}\right) &\geq \frac{1-\epsilon}{2}, \\ \textit{or} \quad \mathrm{q}\left(|H_\alpha(\mathrm{q}) - \hat{f}(X)| \geq \frac{|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})|}{2}\right) &\geq \frac{1-\epsilon}{2}. \end{aligned}$$

We omit the simple proof. Therefore, the required contradiction, and consequently the lower bound

$$S_\alpha(k) > k^{\,c(\alpha)},$$

will follow upon showing that there are distributions p and q of support-size k such that the following hold:

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

$$|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| > \delta; \tag{20}$$

(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

$$\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\| < \epsilon, \tag{21}$$

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

Therefore, we need to find two distributions p and q with different Rényi entropies and with small variation distance between the distributions of their profiles, when n is sufficiently small. For the latter requirement, we recall a result of [Val08] that allows us to bound the variation distance in (21) in terms of the differences of power sums $|P_a(\mathrm{p}) - P_a(\mathrm{q})|$.

Theorem 17. *[Val08] Given distributions* p *and* q *such that*

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

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

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

It remains to construct the required distributions p and q, satisfying (20) and (21) above. By Theorem 17, the 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

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

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 18. *For every* $d \in \mathbb{N}$ *and* α *not integer, there exist positive vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *such that*

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

A constructive proof of Lemma 18 will be given at the end of this section. We are now in a position to prove our converse results.

We first prove the lower bound for an integer $\alpha > 1$.

Theorem 19. *Given an integer $\alpha > 1$ and any estimator f of $H_\alpha(\mathrm{p})$, for every $0 < \epsilon < 1$ there exits a distribution p with support of size k, $\delta > 0$ and a constant $C > 0$ such that for $n < Ck^{(\alpha-1)/\alpha}$ we have*

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

In particular, for every $0 < \epsilon < 1/2$ there exists $\delta > 0$ such that

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

Proof. For $d = \alpha$, let p and q, respectively, be the distributions $\mathrm{p}^{\mathbf{x}}$ and $\mathrm{p}^{\mathbf{y}}$, where the vectors $\mathbf{x}$ and $\mathbf{y}$ are given by Lemma 18. In view of the foregoing discussion, we need to verify (20) and (21) to prove the theorem. Therefore, (20) holds by Lemma 18 since

$$|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| = \frac{\alpha}{1-\alpha}\left|\log\frac{\|\mathbf{x}\|_\alpha}{\|\mathbf{y}\|_\alpha}\right| > 0,$$

and for $n < C_2 k^{(d-1)/d}$ and $5C_2^d/(1-C_2) < \epsilon/2$. inequality (21) follows from Theorem 17 as

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

Next, we lower bound $S_\alpha(k)$ for noninteger $\alpha > 1$ and show that it must be almost linear in k.

Theorem 20. *Given a nonintegral $\alpha > 1$, for every $0 < \epsilon < 1/2$, we have*

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

Proof. For a fixed d, let distributions p and q be as in the previous proof. Then, as in the proof of Theorem 20, inequality (20) holds by Lemma 18 and (21) holds by Theorem 17 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 21. *Given $\alpha < 1$, for every $0 < \epsilon < 1/2$, we have*

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

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

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

where the vectors $\mathbf{x}$ and $\mathbf{y}$ are given by Lemma 18 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

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

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

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, (20) holds by Lemma 18 since $\alpha(1+\beta) < 1$, and for $n < C_2 k^{(1+\beta-1/d)}$ we get (21) by Theorem 17 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. ■

We close with a proof of Lemma 18.

Proof of Lemma 18. 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

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

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, [Tos06]). Therefore, $h(\alpha) \neq 0$ for all $\alpha \notin \{1, ..., d-1\}$; in particular, $\|\mathbf{x}\|_\alpha \neq \|\mathbf{y}\|_\alpha$ for all Δ sufficiently small. ■

Acknowledgements

The authors thank Chinmay Hegde and Piotr Indyk for helpful discussions and suggestions.

References

[AK01] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms*, 19(3-4):163–193, October 2001. 3.2

[Ari96] Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996. 1.1

[BBCM95] C.H. Bennett, G. Brassard, C. Crepeau, and U.M. Maurer. Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6), Nov 1995. 1.1

[BFR+13] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013. 1.1

[BKS01] Ziv Bar-Yossef, Ravi 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*, pages 266–275, 2001. 1.3

[Csi95] I. Csiszár. Generalized cutoff rates and renyi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, January 1995. 1.1

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000. 1.1

[HLP52] G. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities. 2nd edition.* Cambridge University Press, 1952. 2.1

[HNO08] Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof 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*, pages 489–498, 2008. 1.1

[HS11] Manjesh Kumar Hanawal and Rajesh Sundaresan. Guessing revisited: A large deviations approach. *IEEE Transactions on Information Theory*, 57(1):70–78, 2011. 1.1

[IS13] Velimir M. Ilic and Miomir S. Stankovic. A unified characterization of generalized information and certainty measures. *CoRR*, abs/1310.4896, 2013. 1.1

[IZ89] R. Impagliazzo and D. Zuckerman. How to recycle random bits. In *FOCS*, 1989. 1.1

[JHE$^+$03] R. Jenssen, KE Hild, D. Erdogmus, J.C. Principe, and T. Eltoft. Clustering using Renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE, 2003. 1.1

[JVW14a] J. Jiao, K. Venkat, and T. Weissman. Maximum likelihood estimation of functionals of discrete distributions. *CoRR*, abs/1406.6959, 2014. 4, 4

[JVW14b] J. Jiao, K. Venkat, and T. Weissman. Order-optimal estimation of functionals of discrete distributions. *CoRR*, abs/1406.6956, 2014. 1.1, 1.2, 1.3, 3.3, 4, 4

[KLS11] D. Källberg, N. Leonenko, and O. Seleznjev. Statistical inference for rényi entropy functionals. *CoRR*, abs/1103.4977, 2011. 1.1

[Knu73] Donald E. Knuth. *The Art of Computer Programming, Volume III: Sorting and Searching*. Addison-Wesley, 1973. 1.1

[LSO$^+$06] Ashwin Lall, Vyas Sekar, Mitsunori Ogihara, Jun Xu, and Hui Zhang. Data streaming algorithms for estimating entropy of network traffic. *SIGMETRICS Perform. Eval. Rev.*, 34(1):145–156, June 2006. 1.1

[Mar92] VA Markov. On functions deviating least from zero in a given interval. *Izdat. Imp. Akad. Nauk, St. Petersburg*, pages 218–258, 1892. 6

[Mas94] J.L. Massey. Guessing and entropy. In *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, pages 204–, Jun 1994. 1.1

[MBT13] A.S. Motahari, G. Bresler, and D.N.C. Tse. Information theory of dna shotgun sequencing. *Information Theory, IEEE Transactions on*, 59(10):6273–6289, Oct 2013. 1.1

[MIGM00] B. Ma, A. O. Hero III, J. D. Gorman, and O. J. J. Michel. Image registration with minimum spanning tree algorithm. In *ICIP*, pages 481–484, 2000. 1.1

[Mil55] G. A. Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100, 1955. 3.2

[MMR12] Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Renyi divergence. *CoRR*, abs/1205.2628, 2012. 1.1

[Mok89] A. Mokkadem. Estimation of the entropy and information of absolutely continuous random variables. *IEEE Transactions on Information Theory*, 35(1):193–196, 1989. 1.1

[NBdRvS04] 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*, 69:056111–056111, 2004. 1.1

[NHZC06] H. Neemuchwala, A. O. Hero, S. Z., and P. L. Carson. Image registration methods in high-dimensional space. *Int. J. Imaging Systems and Technology*, 16(5):130–145, 2006. 1.1

[OSVZ04] Alon Orlitsky, Narayana P. Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *UAI*, 2004. 4

[OW99] Paul C. Van Oorschot and Michael J. Wiener. Parallel collision search with cryptanalytic applications. *Journal of Cryptology*, 12:1–28, 1999. 1.1

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. 1.1

[Pan04] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004. 1.1, 3.2

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1.1

[PS04] C.-E. Pfister and W.G. Sullivan. Renyi entropy, guesswork moments, and large deviations. *IEEE Transactions on Information Theory*, 50(11):2794–2800, Nov 2004. 1.1

[Rén61] 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*, pages 547–561, 1961. 1.1, 2.1

[SEM91] P. S. Shenkin, B. Erman, and L. D. Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins*, 11(4):297–313, 1991. 1.1

[Tim63] A. F. Timan. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, 1963. 1.3, 5

[Tos06] T. Tossavainen. On the zeros of finite sums of exponential functions. *Australian Mathematical Society Gazette*, 33(1):47–50, 2006. 4

[Val08] P. Valiant. Testing symmetric properties of distributions. In *STOC*, 2008. 4, 4, 17

[VV11] G. Valiant and P. Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *STOC*, 2011. 1.1, 4

[WY14] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *CoRR*, abs/1407.0381v1, 2014. 1.1, 1.2, 1.3, 3.3, 3.3

[XE10] D. Xu and D. Erdogmuns. Renyi's entropy, divergence and their nonparametric estimators. In *Information Theoretic Learning*, Information Science and Statistics, pages 47–102. Springer New York, 2010. 1.1

[Xu98] D. Xu. *Energy, Entropy and Information Potential for Neural Computation*. PhD thesis, University of Florida, 1998. 1.1

Appendix: Estimating power sums

The broader problem of estimating smooth functionals of distributions was considered in [VV11]. Independently and concurrently with this work, [JVW14b] 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, [JVW14b] showed that for $\alpha < 1$,

$$\Omega\left(\frac{k^{1/\alpha}}{\log^{3/2} k}\right) \le S_\alpha^{P+}(k) \le O\left(\frac{k^{1/\alpha}}{\log k}\right), \tag{22}$$

and [JVW14a] showed that for $1 < \alpha < 2$,

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

In fact, using techniques similar to multiplicative guarangess on $P_\alpha(\mathrm{p})$ we show that for $S_\alpha^{P+}(k)$ is a constant independent of k for all $k > 1$. Concurrently with this work, similar results were obtained in an updated version of [JVW14b].

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

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

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

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

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 [JVW14b, JVW14a] and the Rényi-entropy estimation results in this paper complement each other in several ways. For example, for $\alpha < 1$,

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

where the first inequality follows from Theorem 21 and the last follows from the upper-bound (22) derived in [JVW14b] 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 20 imply that for non integer α, $\tilde{\tilde{\Omega}}\,(k) \le S_\alpha^{P\times}(k) \le O\,(k)$, 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 [JVW14b].

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

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

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

Lemma 22. *For an appropriately chosen constant $c > 0$, the bias and the variance of the empirical estimator are bounded above as*

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

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:

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

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

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

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

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

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. ■