The Complexity of Estimating Rényi Entropy

Jayadev Acharya*[1], Alon Orlitsky†[2], Ananda Theertha Suresh‡[2], and Himanshu Tyagi§[2]

[1]Massachusetts Institute of Technology
[2]University of California, San Diego

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 super-linear, roughly $k^{1/\alpha}$ samples, noninteger $\alpha > 1$ requires near-linear, roughly k samples, but integer $\alpha > 1$ requires only $\Theta(k^{1-1/\alpha})$ samples. In particular, estimating $H_2(\mathrm{p})$, which arises in security, DNA reconstruction, closeness testing, and other applications, requires only $\Theta(\sqrt{k})$ samples. 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 The most commonly used measure of randomness of a distribution p over a 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. However, it was recently shown that estimating Shannon entropy of a k-element distribution p to a given additive accuracy requires $\Theta(k/\log k)$ 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

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

As shown in its introductory paper [Rén61], 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.

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]. Yet no systematic study of the sample complexity of estimating Rényi entropy is available. For example, it was hitherto unknown if the number of samples needed to estimate

*jayadev@csail.mit.edu
†alon@ucsd.edu
‡asuresh@ucsd.edu
§htyagi@eng.ucsd.edu

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 in three different ranges. For $\alpha < 1$ it grows superlinearly with k, for $1 < \alpha \notin \mathbb{N}$ it grows roughly linearly with k, and for orders $1 < \alpha \in \mathbb{N}$ it grows as $\Theta(k^{1-1/\alpha})$, much slower than the corresponding growth rate for 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. Given independent samples $X^n = X_1, \ldots, X_n$ from p, define $S^f_\alpha(k, \delta, \epsilon)$ to be the minimum number of samples an estimator f needs to approximate $H_\alpha(\mathrm{p})$ of any k-symbol distribution p to a given additive accuracy δ with probability greater than $1 - \epsilon$, namely

$$S^f_\alpha(k, \delta, \epsilon) \stackrel{\text{def}}{=} \min \{n : \mathrm{p}\left(|H_\alpha(\mathrm{p}) - f\left(X^n\right)| > \delta\right) < \epsilon, \text{ for all } \mathrm{p} = (\mathrm{p}_1, ..., \mathrm{p}_k)\}$$

The *sample complexity* of estimating $H_\alpha(\mathrm{p})$ is then

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

the least number of samples any estimator needs to estimate the order-α Rényi entropy of all k-symbol distributions to additive accuracy δ with probability greater than $1 - \epsilon$.

We are mostly interested in the dependence of $S_\alpha(k, \delta, \epsilon)$ on the alphabet size k and typically omit δ and ϵ to write $S_\alpha(k)$. Additionally, to focus on the essential growth rate of $S_\alpha(k)$, we use standard asymptotic notation where $S_\alpha(k) = O(k^\beta)$ indicates that for some constant c that may depend on α, δ, and ϵ, for all sufficiently large k, $S_\alpha(k) \le c \cdot k^\beta$. Similarly, $S_\alpha(k) = \Theta(k^\beta)$ adds the corresponding $\Omega(k^\beta)$ lower bound for sufficiently small δ and ϵ. Finally, we let $S_\alpha(k) = \tilde{\tilde{\Omega}}\left(k^\beta\right)$ indicate that for all sufficiently small δ and ϵ, and for all $\eta > 0$, for all sufficiently large k, $S_\alpha(k) > k^{\beta-\eta}$, namely that if δ and ϵ are small enough then $S_\alpha(k)$ grows polynomially in k with exponent not less than β.

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

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

namely the sample complexity grows superlinearly in k and estimating the Rényi entropy of these orders is even more difficult than estimating Shannon entropy. As shown in Subsection 1.3, the upper bound follows from a result in [JVW14b], and in Theorem 3.2 we show that the simple empirical-frequency estimator requires only a little more, $O\left(k^{1/\alpha}\right)$ samples. The lower bound is proved in Theorem 4.4.

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

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

namely as with Shannon entropy, the sample complexity grows roughly linearly in the alphabet size. The lower bound is proved in Theorem 4.3 and the upper bound in Theorem 3.1 using the empirical-frequency estimator.

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

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

namely the sample complexity is sublinear in the alphabet size. The upper and lower bounds are shown in Theorems 3.3 and 4.2, respectively.

It was recently brought to our attention that [BKS01] considered the related problem of estimating integer moments of frequencies in a sequence. While their setting and proofs are different, their analysis implies our results for integer α, though not for non-integer α.

Of the three ranges, the most frequently used is the last, $\alpha = 2, 3, \ldots$. Some elaboration is therefore in order.

First, for all orders in this range, $H_\alpha(\mathrm{p})$ 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 and genetic analysis.

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

Finally, note that Rényi entropy is continuous in its 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. However, there

is no contradiction. Rényi entropy, while continuous in α, is not uniformly continuous. In fact, as shown in Example 1.2, the difference between say $H_2(\mathrm{p})$ and $H_{2.001}(\mathrm{p})$ may increase to infinity when the distribution-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 Relation to power sum estimation The *power sum* of order α for a distribution p over $\mathcal{X}$ is

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

and it is related to Rényi entropy 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 $2^{\pm\delta\cdot(1-\alpha)}$. Since the dependence on δ is absorbed in the asymptotic notation, letting $S_\alpha^{P\times}(k)$ denote the number of samples needed to estimate $P_\alpha(\mathrm{p})$ to a fixed multiplicative accuracy, it follows that

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

and consequently the results outlined in Susbection 1.2 for the additive estimation of $H_\alpha(\mathrm{p})$ also apply to the multiplicative estimation of $P_\alpha(\mathrm{p})$.

Clearly, the power sums too measure the randomness of a distribution [Goo89], and starting with [AMS96], estimating the empirical power sums of a data stream using minimum space has generated considerable interest, with the order-optimal space complexity for $\alpha \geq 2$ determined in [IW05].

Let $S_\alpha^{P+}(k)$ denote the number of samples needed to estimate $P_\alpha(\mathrm{p})$ to a given additive accuracy. Results derived in [BKS01] for the frequency-estimation setting of this problem imply that for $1 < \alpha \in \mathbb{N}$, $S_\alpha^{P+}(k)$ is a constant independent of k. Recently, [JVW14b] showed that for $\alpha < 1$,

$$\Omega\left(\frac{k^{1/\alpha}}{\log^{3/2} k}\right) \leq S_\alpha^{P+}(k) \leq O\left(\frac{k^{1/\alpha}}{\log k}\right), \tag{1.1}$$

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

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

In the Appendix we show that for for all $\alpha > 1$, $S_\alpha^{P+}(k)$ is a constant independent of k. Similar results were concurrently obtained and appeared 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

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

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

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

where the first inequality follows from Theorem 4.4 and the last follows from the upper-bound (1.1) derived in [JVW14b]. 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 3.1 and 4.3 imply that for non integer α, $\tilde{\tilde{\Omega}}(k) \leq S_\alpha^{P\times}(k) \leq O(k)$, 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.

1.4 The estimators As suggested by the above discussion, we construct multiplicative-accuracy power sum estimators and use them to derive additive-accuracy estimators for Rényi entropy. We use two simple and practical estimators, one for integer, and the other for noninteger, values of α.

To simplify the analysis we use *Poisson sampling*, described further in Section 2. Instead of generating exactly n independent samples from p, we generate $N \sim \mathrm{Poi}(n)$ samples, where $\mathrm{Poi}(n)$ is the Poisson distribution with parameter n. Let $X_1, \ldots, X_N$, be the samples, and let

$$N_x \stackrel{\text{def}}{=} |\{1 \leq i \leq N : X_i = x\}|$$

be the number of times a symbol x appears. Note that under Poisson sampling $N_x \sim \mathrm{Poi}(np_x)$ independent of other symbols, and hence the *empirical frequency* N_x/n is an unbiased estimator for p_x.

The following estimators are used for different ranges of α.

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

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

While $\widehat{P}_\alpha^{\mathrm{e}}$ is biased, it gives a reasonably good performance for all $\alpha \neq 1$. Theorem 3.2 shows that for $\alpha < 1$ its sample complexity is $O(k^{1/\alpha})$, and Theorem 3.1 shows that for $\alpha > 1$ it is $O(k)$. The lower bounds in Section 4 show that these sample complexities have the optimal exponent of k for all noninteger α.

1.4.2 Bias-corrected estimator 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. The resulting *bias-corrected* estimator for $P_\alpha(\mathrm{p})$ is

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

as

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

Theorem 3.3 show that for $1 < \alpha \in \mathbb{N}$, $\widehat{P}_\alpha^{\mathrm{u}}$ estimates $P_\alpha(\mathrm{p})$ using $O(k^{1-1/\alpha})$ samples, and Theorem 4.2 shows that this number is optimal up to a constant factor.

We relate $\widehat{P}_\alpha^{\mathrm{u}}$ to another simple power sum estimator considered in [BKS01]. For $\alpha \in \mathbb{N}$, $P_\alpha(\mathrm{p})$ is the probability that α independent samples from p are all identical. This suggests taking n samples, and estimating $P_\alpha(\mathrm{p})$ by the fraction $\widehat{P}_\alpha^{\mathrm{u}'}$ of α-element subsets that consist of a single value.

More formally, given n independent samples $X^n = X_1, \ldots, X_n$ from p, for $S \subseteq [n]$, let

$$\mathbb{1}_S(X^n) = \begin{cases} 1 & X_i = X_j \text{ for all } i, j \in S, \\ 0 & X_i \neq X_j \text{ for some } i, j \in S \end{cases}$$

indicate whether the X_i are identical for all $i \in S$, and let $\binom{[n]}{\alpha}$ denote the collection of all α-element subsets of $[n]$. Then

$$\widehat{P}_\alpha^{\mathrm{u}'} = \frac{1}{\binom{n}{\alpha}} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n).$$

Note that $\widehat{P}_\alpha^{\mathrm{u}'}$ is unbiased because for all $S \subseteq [n]$,

$$\mathbb{E}[\mathbb{1}_S(X^n)] = P_{|S|}(\mathrm{p}),$$

and hence,

$$\mathbb{E}\left[\widehat{P}_\alpha^{\mathrm{u}'}\right] = \mathbb{E}\left[\frac{1}{\binom{n}{\alpha}} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n)\right] = P_\alpha(\mathrm{p}).$$

To relate $\widehat{P}_\alpha^{\mathrm{u}'}$ and $\widehat{P}_\alpha^{\mathrm{u}}$, observe that

$$\mathbb{1}_S(X^n) = \sum_{x \in \mathcal{X}} \mathbb{1}(X_i = x \quad \forall i \in S)$$

and let N'_x denote the number of $1 \leq i \leq n$ such that $X_i = x$. Then

$$\begin{aligned} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n) &= \sum_{S \in \binom{[n]}{\alpha}} \sum_{x \in \mathcal{X}} \mathbb{1}(X_i = x \quad \forall i \in S) \\ &= \sum_{x \in \mathcal{X}} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}(X_i = x \quad \forall i \in S) \\ &= \sum_{x \in \mathcal{X}} \binom{N'_x}{\alpha}. \end{aligned}$$

Hence

$$\begin{aligned} \widehat{P}_\alpha^{\mathrm{u}'} &= \frac{1}{\binom{n}{\alpha}} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n) \\ &= \frac{1}{\binom{n}{\alpha}} \sum_{x \in \mathcal{X}} \binom{N'_x}{\alpha} \\ &= \sum_{x \in \mathcal{X}} \frac{(N'_x)^{\underline{\alpha}}}{n^{\underline{\alpha}}}, \end{aligned}$$

namely $\widehat{P}_\alpha^{\mathrm{u}'}$ can be viewed as a fixed-sampling equivalent of the Poisson-sampling $\widehat{P}_\alpha^{\mathrm{u}}$.

The special case of $\widehat{P}_\alpha^{\mathrm{u}'}$ where $\alpha = 2$ was considered in [GR00, BFR+13] for testing whether a distribution is close to uniform. They also analyzed its variance and their calculations along with Lemma 2.1 and Lemma 3.2 can provide an alternative derivation of an order-optimal estimator for $H_2(\mathrm{p})$.

1.5 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 1.1. *The* uniform distribution U_k *over* $[k] = \{1, \ldots, k\}$ *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.

EXAMPLE 1.2. *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*

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

We illustrate that the continuity of $H_\alpha(Z_{\beta,k})$ *in* α *is not uniform in* k*. Specifically, we note that the difference between* $H_2(Z_{\beta,k})$ *and* $H_{2+\epsilon}(Z_{\beta,k})$ *may increase to infinity when the distribution-size increases. This implies that one cannot approximate, for instance,* $H_{2.001}(\mathrm{p})$ *using* $H_2(\mathrm{p})$ *when the underlying distribution* p *is unknown.*

Table 1 summarizes the leading term $g(k)$ *in the approximation*[1] $H_\alpha(Z_{\beta,k}) \sim g(k)$*.*

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

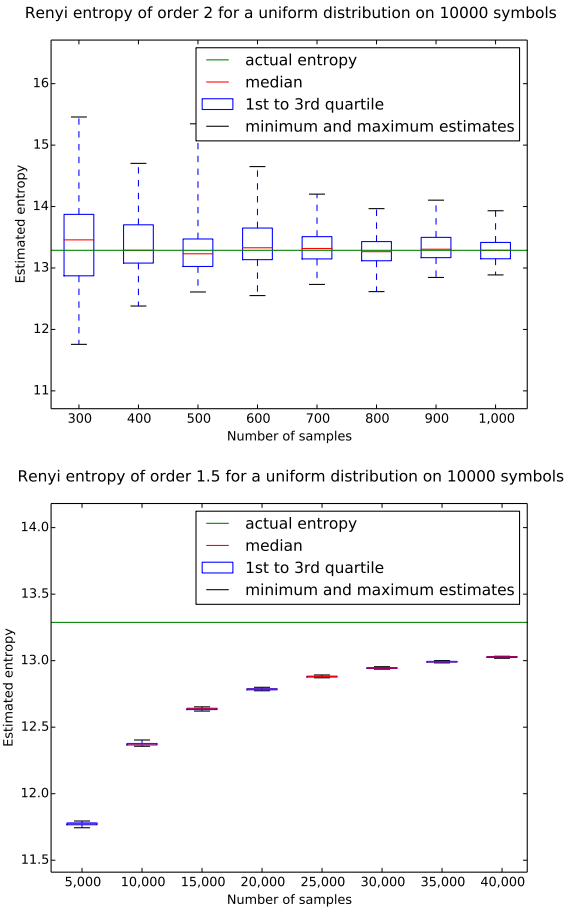

Figure 1: Estimation of Rényi entropy of order 2 and order 1.5 using the bias-corrected estimator and empirical estimator, respectively, for samples drawn from a uniform distribution. The boxplots display the estimated values for 100 independent experiments.

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

Table 1: 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. Figure 2 shows the performance of our estimators for samples drawn from $Z_{1,k}$.

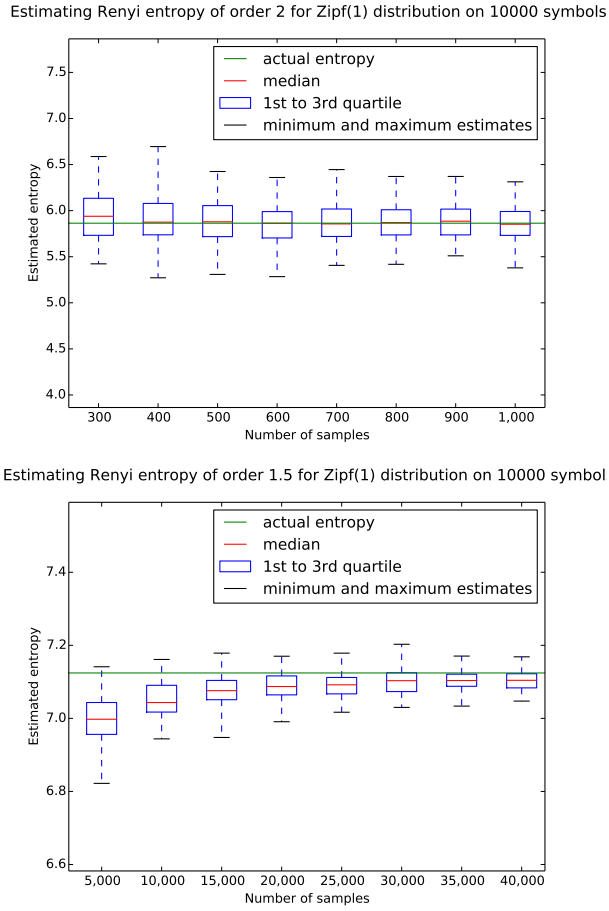

Figure 2: Estimation of Rényi entropy of order 2 and order 1.5 using the bias-corrected estimator and empirical estimator, respectively, for samples drawn from $Z_{1,k}$. The boxplots display the estimated values for 100 independent experiments.

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.

1.6 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:

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

or equivalently

$$1 \le P_\alpha(\mathrm{p}) \le k^{1-\alpha}, \quad \text{for } \alpha < 1 \tag{2.2}$$

$$k^{1-\alpha} \le P_\alpha(\mathrm{p}) \le 1, \quad \text{for } \alpha > 1. \tag{2.3}$$

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 2.1. *For every $0 \le \alpha$,*

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

Further, for $\alpha > 1$ and $0 \le \beta \le \alpha$,

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

and

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

Proof. By the monotonicity of norms,

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

which gives

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

The first inequality follows upon choosing $\beta = \alpha$. For $1 < \alpha$ and $0 \le \beta \le \alpha$, we get the second by (2.2). For

the final inequality, note that by the monotonicity of Hölder means, we have

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

The final inequality follows upon rearranging the terms and using (2.2).

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 2.2. *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

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

where the inequality follows from

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

Therefore,

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

■

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

LEMMA 2.3. *Let $X \sim \mathrm{Poi}(\lambda)$ and let β be a positive real number. Then,*

$$\mathbb{E}\left[X^{\beta}\right] \le c(\beta) \max\{\lambda, \lambda^{\beta}\},$$

where the constant $c(\beta)$ does not depend on λ.

Proof. For the case when $\lambda > 1$, we have

$$\begin{aligned}
\mathbb{E}\left[X^{\beta}\right] &\le \sum_{i \le 2\lambda} \mathrm{Poi}(\lambda, i)\, i^{\beta} + \sum_{i > 2\lambda} \mathrm{Poi}(\lambda, i)\, i^{\beta} \\
&\le 2^{\beta}\lambda^{\beta} + \sum_{i > 2\lambda} \mathrm{Poi}(\lambda, i)\, i^{\beta}.
\end{aligned}$$

Using the standard tail bound for Poisson random variables, if $\lambda > 1$, for all $i > 2\lambda$

$$\mathrm{Poi}(\lambda, i) \le \mathbb{P}\left(X \ge i\right) \le \exp\left(-\frac{i-\lambda}{8\lambda}\right),$$

where the second inequality follows upon bounding the cumulant-generating function of X for $i > 2\lambda$ and $\lambda > 1$. Therefore,

$$\mathbb{E}\big[X^\beta\big] \le \left(2^\beta + e^{1/8}\int_0^\infty e^{-x/8}x^\beta\,dx\right)\lambda^\beta, \quad \lambda > 1.$$

For $\lambda \le 1$, since $\lambda^i \le \lambda$ for all $i \ge 1$

$$\mathbb{E}\big[X^\beta\big] \le \lambda\sum_{i=1}^\infty \frac{i^\beta}{i!}, \tag{2.4}$$

and the lemma follows upon choosing

$$c(\beta) = \max\left\{\sum_{i=1}^\infty \frac{i^\beta}{i!}, \left(2^\beta + e^{1/8}\int_0^\infty e^{-x/8}x^\beta\,dx\right)\right\},$$

which is a finite quantity. ∎

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

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

where the constant c is given by $\sqrt{c(2\alpha-2)}$ with $c(2\alpha-2)$ as in Lemma 2.3.

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

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

Taking expectations on both sides,

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

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,

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

where the last-but-one inequality is by Lemma 2.3. ∎

3 Upper bounds on sample complexity

In this section, we analyze the performances of the estimators we proposed in Section 1.4. 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:

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

The following reduction to Poisson sampling is well-known.

LEMMA 3.1. **(Poisson approximation 1)** *For $n \ge 8\log(2/\epsilon)$ and $N \sim \mathrm{Poi}(n/2)$,*

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

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 3.2. *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* $O(n\log(1/\epsilon))$ *samples and ensures*

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

Proof. By Chebychev'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. 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^{e}_α and f^{u}_α, respectively, the empirical estimator $\frac{1}{1-\alpha}\log\widehat{P}^{\mathrm{e}}_\alpha$ and the bias-corrected estimator $\frac{1}{1-\alpha}\log\widehat{P}^{\mathrm{u}}_\alpha$.

3.1 Performance of empirical estimator First, we present upper bounds for the sample complexity of the empirical estimator separately for $\alpha > 1$ and $\alpha < 1$.

THEOREM 3.1. *For* $\alpha > 1$, $\delta > 0$, *and* $0 < \epsilon < 1$, *the estimator* f^e_α *satisfies*

$$S^{f^e_\alpha}_\alpha(k, \delta, \epsilon) \le O\left(\frac{k}{\delta^{\max\{4,\, 1/(\alpha-1)\}}}\log\frac{1}{\epsilon}\right).$$

Proof. Denote $\lambda_x \stackrel{\mathrm{def}}{=} n\mathrm{p}_x$. For $\alpha > 1$, the bias of the power sum estimator is bounded using Lemma 2.4 as follows:

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right|\\ &\le \frac{1}{n^\alpha}\sum_x |\mathbb{E}[N_x^\alpha] - \lambda_x^\alpha|\\ &\le \frac{\alpha}{n^\alpha}\sum_x \left(c\lambda_x + (c+1)\lambda_x^{\alpha-1/2}\right)\\ &\le \frac{\alpha c}{n^{\alpha-1}} + \frac{\alpha(c+1)}{\sqrt{n}} P_{\alpha-1/2}(\mathrm{p})\\ &\le \alpha\left(c\left(\frac{k}{n}\right)^{\alpha-1} + (c+1)\sqrt{\frac{k}{n}}\right) P_\alpha(\mathrm{p}) \qquad (3.5)\end{aligned}$$

where the previous inequality is by Lemma 2.1 and (2.2).

Similarly, for bounding the variance, using independence of multiplicities, the following inequalities ensue:

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right]\\ &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{V}\mathrm{ar}[N_x^\alpha]\\ &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - [\mathbb{E}N_x^\alpha]^2\\ &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha} + \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2\\ &\le \frac{1}{n^{2\alpha}}\sum_x \left|\mathbb{E}\left[N_x^{2\alpha}\right] - \lambda_x^{2\alpha}\right|\\ &\le 2\alpha\left(c\left(\frac{k}{n}\right)^{2\alpha-1} + (c+1)\sqrt{\frac{k}{n}}\right) P_\alpha(\mathrm{p})^2 \qquad (3.6)\end{aligned}$$

where the last-but-one inequality holds by Jensen's inequality since z^α is a convex function; the final inequality is by[2] (3.5) and Lemma 2.1. Therefore, the bias and variance are small when $n = O(k)$ and theorem follows by Lemma 3.2. ■

THEOREM 3.2. *For* $\alpha < 1$, $\delta > 0$, *and* $0 < \epsilon < 1$, *the estimator* f^e_α *satisfies*

$$S^{f^e_\alpha}_\alpha(k, \delta, \epsilon) \le O\left(\frac{k^{1/\alpha}}{\delta^{\max\{4,\, 2/\alpha\}}}\log\frac{1}{\epsilon}\right).$$

Proof. For $\alpha < 1$, once again we take a recourse to

[2] For brevity, the constants in (3.5) and (3.6), albeit different, are both denoted by c.

Lemma 2.4 to bound the bias as follows:

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

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

$$\left|\mathbb{E}\left[\frac{\sum_x N_x^\alpha}{n^\alpha}\right] - P_\alpha(\mathrm{p})\right| \le 2\left(\frac{k^{1/\alpha}}{n}\right)^\alpha \le 2P_\alpha(\mathrm{p})\left(\frac{k^{1/\alpha}}{n}\right)^\alpha, \tag{3.7}$$

where the last inequality uses (2.2). For bounding the variance, note that

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

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 (3.6) applies to give

$$\frac{1}{n^{2\alpha}}\sum_x \mathbb{E}\big[N_x^{2\alpha}\big] - \lambda_x^{2\alpha} \le 2\alpha\left(\frac{c}{n^{2\alpha-1}} + (c+1)\sqrt{\frac{k}{n}}\right)P_\alpha(\mathrm{p})^2. \tag{3.8}$$

For the second term, we have

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

where the last-but-one inequality is by (3.7) and the last inequality uses the concavity of z^α in z. The proof is completed by combining the two bounds above and using Lemma 3.2. ■

3.2 Performance of bias-corrected estimator for integral α Next, we bound the number of samples needed for the bias-corrected estimator, thereby establishing its optimality for integer $\alpha > 1$.

THEOREM 3.3. *For an integer $\alpha > 1$, any $\delta > 0$, and $0 < \epsilon < 1$, the estimator f_α^u satisfies*

$$S_\alpha^{f_\alpha^u}(k, \delta, \epsilon) \le O\left(\frac{k^{(\alpha-1)/\alpha}}{\delta^2}\log\frac{1}{\epsilon}\right).$$

Proof. For bounding the variance of g_α, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\frac{\sum_x N_x^{\underline{\alpha}}}{n^\alpha}\right] &= \frac{1}{n^{2\alpha}}\sum_x \mathbb{V}\mathrm{ar}[N_x^{\underline{\alpha}}]\\ &\le \frac{1}{n^{2\alpha}}\sum_x \left(\lambda_x^\alpha(\lambda_x+\alpha)^\alpha - \lambda_x^{2\alpha}\right)\\ &= \frac{1}{n^{2\alpha}}\sum_{r=0}^{\alpha-1}\sum_x \binom{\alpha}{r}\alpha^{\alpha-r}{\lambda_x}^{\alpha+r}\\ &= \frac{1}{n^{2\alpha}}\sum_{r=0}^{\alpha-1} n^{\alpha+r}\binom{\alpha}{r}\alpha^{\alpha-r}P_{\alpha+r}(\mathrm{p}),\end{aligned} \tag{3.9}$$

where the inequality uses Lemma 2.2. It follows from Lemma 2.1 that

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

Furthermore, by Lemma 2.2 the estimator is unbiased under Poisson sampling, which completes the proof by Lemma 3.2. ■

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 $H_\alpha(\mathrm{p}) \ne H_\alpha(\mathrm{q})$ for which similar multiplicities appear if fewer samples than the claimed lower bound are available.

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

LEMMA 4.1. **(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 4.2. **(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 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 4.3. *If for two distributions* p *and* q *on $\mathcal{X}$ the variational distance $\|\mathrm{p} - \mathrm{q}\| < \epsilon$, then one of the following holds for every function $\hat{f}$:*

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

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

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

$$|H_\alpha(\mathrm{p}) - H_\alpha(\mathrm{q})| > \delta; \tag{4.10}$$

(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

$$\|\mathrm{p}_\Phi - \mathrm{q}_\Phi\| < \epsilon, \tag{4.11}$$

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 (4.11) in terms of the differences of power sums $|P_a(\mathrm{p}) - P_a(\mathrm{q})|$.

THEOREM 4.1. *[Val08] 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 (4.10) and (4.11) above. By Theorem 4.1, 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

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

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

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

A constructive proof of Lemma 4.4 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 4.2. *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

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

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 4.4. In view of the foregoing discussion, we need to verify (4.10) and (4.11) to prove the theorem. Therefore, (4.10) holds by Lemma 4.4 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 (4.11) follows from Theorem 4.1 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 4.3. *Given a nonintegral $\alpha > 1$, for every $0 < \epsilon < 1/2$, we have*

$$S_\alpha(k, \delta, \epsilon) \geq \widetilde{\widetilde{\Omega}}\left(k\right).$$

Proof. For a fixed d, let distributions p and q be as in the previous proof. Then, as in the proof of Theorem 4.3, inequality (4.10) holds by Lemma 4.4 and (4.11) holds by Theorem 4.1 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 4.4. *Given $\alpha < 1$, for every $0 < \epsilon < 1/2$, we have*

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

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

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

where the vectors $\mathbf{x}$ and $\mathbf{y}$ are given by Lemma 4.4 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

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

and

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

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, (4.10) holds by Lemma 4.4 since $\alpha(1 + \beta) < 1$, and for $n < C_2 k^{(1+\beta-1/d)}$ we get (4.11) by Theorem 4.1 as

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

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

We close with a proof of Lemma 4.4.

Proof of Lemma 4.4. 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

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

and that

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

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

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

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.

[AMS96] N. Alon, T. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *STOC*, 1996.

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

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

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[BKS01] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: lower bounds and applications. In *STOC*, pages 266–275, 2001.

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

[Goo89] I. J. Good. Surprise indexes and p-values. *Statistical Computation and Simulation*, 32:90–92, 1989.

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

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

[HNO08] N. J. A. Harvey, J. Nelson, and K. Onak. Sketching and streaming entropy via approximation theory. In *FOCS*, pages 489–498, 2008.

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

[IW05] P. Indyk and D. Woodruff. Optimal approximations of the frequency moments of data streams. In *STOC*, 2005.

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

[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.

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

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

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

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

[LSO+06] A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang. Data streaming algorithms for estimating entropy of network traffic. *SIGMETRICS Perform. Eval. Rev.*, 34(1):145–156, June 2006.

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

[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.

[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.

[Mil55] G. A. Miller. Note on the bias of information estimates. *Information theory in psychology: Problems*

and methods, 2:95–100, 1955.

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

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

[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.

[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.

[OSVZ04] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *UAI*, 2004.

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

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

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

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

[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.

[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.

[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.

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

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

[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.

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

[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.

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

Appendix: Estimating power sums

As in Section 1.3, let $S_\alpha^{P+}(k)$ denote the number of samples needed to estimate the power sum $P_\alpha(\mathrm{p})$ to a given additive accuracy. We 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 3.2, it suffices to bound the bias and variance of the empirical estimator.

As before, we comsider 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 4.5. *For an appropriately chosen constant $c > 0$, the bias and the variance of the empirical estimator are bounded above as*

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

for all $n \ge 1$.

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

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

where the last inequality holds by Lemma 2.4 and (2.4) since x^α is convex in x. Noting $\sum_i \lambda_x = n$, we get

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

Similarly, proceeding as in the proof of Theorem 3.1,

the variance of the empirical estimator is bounded as

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

The proof is completed upon showing that

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

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

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

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

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

and completes the proof. ■