The Complexity of Estimating Rényi Entropy

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

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

Abstract

It was recently shown that estimating the Shannon entropy \( H(p) \) 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 \( H(p) \) can be replaced by the more general Rényi entropy of order \( \alpha \), \( H_\alpha(p) \). We determine the number of samples needed to estimate \( H_\alpha(p) \) for all \( \alpha \), 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. In particular, estimating \( H_2(p) \), 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.

∗jayadev@csail.mit.edu
†alon@ucsd.edu
‡asuresh@ucsd.edu
§htyagi@eng.ucsd.edu
1 Introduction

1.1 Shannon and Rényi entropies

The most commonly used measure of randomness of a distribution $p$ over a set $X$ is its Shannon entropy

$$H(p) \overset{\text{def}}{=} \sum_{x \in X} p_x \log \frac{1}{p_x}.$$ 

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

$$H_\alpha(p) \overset{\text{def}}{=} \frac{1}{1-\alpha} \log \sum_{x \in X} p_x^\alpha$$

and for $\alpha = 1$ by

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

As shown in its introductory paper [Rén61], Rényi entropy of order 1 is the Shannon entropy, namely $H_1(p) = H(p)$, 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 the Shannon entropy [Mok89, NBdRvS04], 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, $H_2(p)$, 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 the Rényi entropy of a given order $\alpha$ differs from that required for Shannon entropy, or whether it varies with the order $\alpha$, 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 $H_\alpha(p)$ falls in three different ranges. For $\alpha < 1$ it grows superlinearly with $k$, for $1 < \alpha \notin \mathbb{Z}$ it grows roughly
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 \( f : \mathcal{X}^* \rightarrow \mathbb{R} \) 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

\[
S^f_\alpha(k, \delta, \epsilon) \overset{\text{def}}{=} \min \{ n : p (|H_\alpha(p) - f(X^n)| > \delta) < \epsilon \}
\]

to be the minimum number of samples an estimator \( f \) needs to approximate \( H_\alpha(p) \) of any \( k \)-symbol distribution \( p \) to a given additive accuracy \( \delta \) with probability greater than \( 1 - \epsilon \). The sample complexity of estimating \( H_\alpha(p) \) is then

\[
S_\alpha(k, \delta, \epsilon) \overset{\text{def}}{=} \min_f S^f_\alpha(k, \delta, \epsilon),
\]

the least number of samples any estimator needs to estimate the order-\( \alpha \) Rényi entropy of all \( k \)-symbol distributions to additive accuracy \( \delta \) 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 \( \delta \) and \( \epsilon \) 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 \( \alpha \), \( \delta \), and \( \epsilon \), for all sufficiently large \( k \), \( S_\alpha(k) \leq c \cdot k^{\beta} \). Similarly \( S_\alpha(k) = \Theta(k^{\beta}) \) adds the corresponding \( \Omega(k^{\beta}) \) lower bound for sufficiently small \( \delta \) and \( \epsilon \). Finally, extending the \( \Omega \) notation, we let \( S_\alpha(k) = \tilde{\Omega}(k^{\beta}) \) indicate that for all sufficiently small \( \delta \) and \( \epsilon \), and for all \( \eta > 0 \), for all sufficiently large \( k \), \( S_\alpha(k) > k^{\beta - \eta} \), namely that if \( \delta \) and \( \epsilon \) are small enough then \( S_\alpha(k) \) grows polynomially in \( k \) with exponent not less than \( \beta \).

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

\[
\tilde{\Omega} \left( k^{1/\alpha} \right) \leq S_\alpha(k) \leq O \left( k^{1/\alpha} / \log k \right),
\]

namely the sample complexity grows superlinearly in \( k \) and estimating the Rényi entropy of these orders is even more difficult than estimating the Shannon entropy. The upper bound was proved in [JVW14b], see also further discussion in Subsection 1.4, and a subsequent result that the simple empirical-frequency estimator requires \( O(k^{1/\alpha}) \) samples is shown in Theorem 8. The lower bound is proved in Theorem 17.

For \( 1 < \alpha \notin \mathbb{N} \),

\[
\tilde{\Omega}(k) \leq S_\alpha(k) \leq O(k),
\]

namely as with Shannon entropy, the sample complexity grows roughly linearly in the alphabet size. The lower bound is proved in Theorem 16 and the upper bound in Theorem 7 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 lower and upper bounds are shown in Theorems 15 and 9, 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 orders in this range, \( H_\alpha(p) \) can be estimated with a sublinear number of samples. The most commonly used, \( H_2(p) \), 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 the Shannon entropy using $\Theta(k/\log k)$ samples, the constant factors implied by the $\Theta$ notation are fairly high. For Rényi entropy of orders $\alpha = 2, 3, ...,\,$ the constants implied by $\Theta(k^{1-1/\alpha})$ are small and shown in Theorem 9 to be below $24\alpha^2$. Furthermore, the experiments described below suggest that they may be even lower.

Finally, note that Rényi entropy is continuous in its order $\alpha$. 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 $H_{2.001}(p)$ one could approximate $H_2(p)$ using significantly fewer samples. However, there is no contradiction. Rényi entropy, while continuous in $\alpha$, is not uniformly continuous. In fact, as shown in Example 2, the difference between say $H_2(p)$ and $H_{2.001}(p)$ 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 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 $[k] = \{1, \ldots , k\}$ is defined by

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

Its Rényi entropy for every order $1 \neq \alpha > 0$, and hence for all $\alpha > 0$, 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 our estimators for samples drawn from a uniform distribution.

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

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

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

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

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

<table>
<thead>
<tr>
<th>$\alpha\beta$</th>
<th>$\beta &lt; 1$</th>
<th>$\beta = 1$</th>
<th>$\beta &gt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha &lt; 1$</td>
<td>$\log k$</td>
<td>$\frac{1}{1-\alpha} \log k$</td>
<td>$\frac{1}{1-\alpha} \log k$</td>
</tr>
<tr>
<td>$\alpha = 1$</td>
<td>$\frac{\alpha-\alpha^2}{\alpha-1} \log k$</td>
<td>$\frac{1}{2} \log k$</td>
<td>$\frac{1}{\alpha} \log \log k$</td>
</tr>
<tr>
<td>$\alpha &gt; 1$</td>
<td>$\frac{\alpha-\alpha^2}{\alpha-1} \log k$</td>
<td>$\frac{1}{\alpha-1} \log \log k$</td>
<td>constant</td>
</tr>
</tbody>
</table>

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 $\beta$. 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 $|H_2(p) - H_{2+\epsilon}(p)|$ is $O(\epsilon \log \log k)$. Therefore, even for very small $\epsilon$ 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 $Z_{1,k}$.

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 $\alpha$ the empirical estimator attains the optimal exponent; we do not consider the possible improvement in performance by reducing the bias in the empirical estimator.

---

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

4
1.4 Relation to moment estimation

The $\alpha$th moment of a distribution$^2$ $p$ over $\mathcal{X}$ is

$$M_\alpha(p) = \sum_{x \in \mathcal{X}} p_x^\alpha,$$

and it is related to Rényi entropy via

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

Hence estimating $H_\alpha(p)$ to an additive accuracy of $\pm \delta$ is equivalent to estimating $M_\alpha(p)$ to a multiplicative accuracy of $2^{\pm \delta/(1 - \alpha)}$. Since the dependence on $\delta$ is absorbed in the asymptotic notation, letting $S_{\alpha}^{M_X}(k)$ denote the number of samples needed to estimate $M_\alpha(p)$ to a fixed multiplicative accuracy, it follows that

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

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

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

The broader problem of estimating smooth functionals of distributions was considered in [VV11]. Recently, [JVW14b] considered estimating more general functionals and applied their technique to estimating the moments of a distribution to a given additive accuracy. Letting $S_{\alpha}^{M^+}(k)$ denote the number of samples needed to estimate $M_\alpha(p)$ to a given additive accuracy, independently and

$^2$With a slight abuse of notation, we distinguish the moment of a distribution $p$, $M_\alpha(p)$, from the expected moment of a random variable $X$, $E[X^\alpha]$. 

---

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.
around the same time as this work [JVW14b] showed that for $\alpha < 1$,
\[
\Omega \left( \frac{k^{1/\alpha}}{\log^{3/2} k} \right) \leq S_\alpha^M(k) \leq O \left( \frac{k^{1/\alpha}}{\log k} \right),
\]
(1)
and [JVW14a] showed that for $1 < \alpha < 2$,
\[
S_\alpha^M(k) \leq O \left( k^{2/\alpha - 1} \right).
\]
(2)

Since $M_\alpha(p) > 1$ for $\alpha < 1$, moment estimation to a fixed additive accuracy implies also a fixed multiplicative accuracy, and therefore
\[
S_\alpha(k) = \Theta(S_\alpha^M(k)) \leq O(S_\alpha^M(k)),
\]
namely for estimation to an additive accuracy, Rényi entropy requires fewer samples than moments. Similarly, $M_\alpha(p) < 1$ for $\alpha > 1$, and therefore
\[
S_\alpha(k) = \Theta(S_\alpha^M(k)) \geq \Omega(S_\alpha^M(k)),
\]
namely for an additive accuracy in this range, Rényi entropy requires more samples than moments.

It follows that the moment-estimation results in [JVW14b, JVW14a] and the Rényi-entropy estimation results in this paper complement each other in several ways. For example, for $\alpha < 1$,
\[
\tilde{\omega} \left( k^{1/\alpha} \right) \leq S_\alpha(k) = \Theta(S_\alpha^M(k)) \leq O(S_\alpha^M(k)) \leq O \left( \frac{k^{1/\alpha}}{\log k} \right),
\]
where the first inequality follows from Theorem 17 and the last follows from the upper-bound proof of (1) in [JVW14b]. Hence, for $\alpha < 1$, estimating moments to additive and multiplicative accuracy require a comparable number of samples.

On the other hand, for $\alpha > 1$, Theorems 16 and 7 imply that for non integer $\alpha$, $\tilde{\omega} \left( k^{1/\alpha} \right) \leq S_\alpha^M(k) \leq O(k)$, while (2) proved in [JVW14a] shows that for $1 < \alpha < 2$, $S_\alpha^M(k) \leq O \left( k^{2/\alpha - 1} \right)$, and for higher $\alpha$, $S_\alpha^M(k)$ is constant. Hence in this range, moment estimation to a multiplicative accuracy requires considerably more samples than to an additive accuracy.

1.5 The estimators

As suggested by the above discussion, we construct multiplicative-accuracy moment estimators and use them to derive additive-accuracy estimators for Rényi entropy. We use two simple and practical estimators, one for integer $\alpha$ and the other for noninteger.

To simplify the analysis we use Poisson sampling, which is described further in Section 2. Instead of generating exactly $n$ independent samples from $p$, we generate $N \sim P_n$ samples, where $P_n$ is the Poisson distribution with parameter $n$. Let $X_1, \ldots, X_N$, be the samples, and let
\[
N_x = \left| \{ 1 \leq i \leq N : X_i = x \} \right|
\]
be the number of times a symbol $x$ appears. As is well known, $N_x \sim P_{n p_x}$, hence the empirical frequency $N_x/n$ is an unbiased estimator for $p_x$. The empirical, or plug-in, estimator of $M_\alpha(p)$ is therefore
\[
\hat{M}_\alpha^{\text{emp}} \overset{\text{def}}{=} \sum_x \left( \frac{N_x}{n} \right)^\alpha.
\]
While $\hat{M}_\alpha^u$ is biased, Theorem 8 shows that for $\alpha < 1$ its sample complexity is $O(k^{1/\alpha})$, and Theorem 7 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 $\alpha$.

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. For $\alpha \in \mathbb{Z}_+$, let $n^\alpha \overset{def}{=} n \cdot (n - 1) \cdot \ldots \cdot (n - \alpha + 1)$ denote the $\alpha$th falling power of $n$. As shown in Section 2, under Poisson sampling, $N_x^\alpha/n^\alpha$ is an unbiased estimator for $p_x^\alpha$. The bias-corrected estimator for $\hat{M}_\alpha^u(p)$ is therefore

$$\hat{M}_\alpha^u = \sum_x \frac{N_x^\alpha}{n^\alpha},$$

as

$$E\left[\hat{M}_\alpha^u\right] = \sum_x E\left[\frac{N_x^\alpha}{n^\alpha}\right] = \sum_x p_x^\alpha = M_\alpha(p).$$

Theorem 9 show that for $1 < \alpha \in \mathbb{Z}$, $\hat{M}_\alpha^u$ estimates $M_\alpha(p)$ using $O(k^{1-1/\alpha})$ samples, and Theorem 15 shows that this number is optimal up to a constant factor.

$\hat{M}_\alpha^u$ is closely related to another simple moment estimator. For $\alpha \in \mathbb{Z}_+$, $M_\alpha(p)$ is the probability that $\alpha$ independent samples from $p$ are all identical. This suggests taking $n$ samples, and estimating $M_\alpha(p)$ by the fraction $\hat{M}_\alpha^u'$ of $\alpha$-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

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

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

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

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

$$E[1_S(X^n)] = M_{|S|}(p),$$

and hence,

$$E\left[\hat{M}_\alpha^u'\right] = E\left[\frac{1}{\binom{n}{\alpha}} \sum_{S \in \binom{[n]}{\alpha}} 1_S(X^n)\right] = M_\alpha(p).$$

To relate $\hat{M}_\alpha^u'$ and $\hat{M}_\alpha^u$, observe that

$$1_S(X^n) = \sum_{x \in X} 1(x_i = x \ \forall i \in S)$$
and let $N'_x$ denote the number of $1 \leq i \leq n$ such that $X_i = x$. Then

$$\sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n) = \sum_{S \in \binom{[n]}{\alpha}} \sum_{x \in \mathcal{X}} \mathbb{1}(X_i = x) \forall i \in S = \sum_{x \in \mathcal{X}} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}(X_i = x) \forall i \in S = \sum_{x \in \mathcal{X}} \left(\frac{N'_x}{\alpha}\right).$$

Hence

$$\widehat{M}_\alpha' = \frac{1}{(n/\alpha)} \sum_{S \in \binom{[n]}{\alpha}} \mathbb{1}_S(X^n) = \frac{1}{(n/\alpha)} \sum_{x \in \mathcal{X}} \left(\frac{N'_x}{\alpha}\right) = \sum_{x \in \mathcal{X}} \frac{(N'_x)^\alpha}{n^\alpha},$$

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

The special case of $\widehat{M}_\alpha'$ 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, which along with Lemma 1 and Lemma 6 can provide an alternative derivation of an order-optimal estimator for $H_2(p)$.

1.6 Organization

The rest of the paper is organized as follows. Section 2 presents basic properties of moments of distributions and expected moments of Poisson random variables, which may be of independent interest. The estimation algorithms are analyzed in Section 3, while lower bounds on the sample complexity of estimating Rényi entropy are established in Section 4. Concluding remarks are given in the final section.

2 Technical preliminaries

2.1 Bounds on moments of a distribution

Consider a distribution $p$ over $[k] = \{1, \ldots, k\}$. Since Rényi entropy is a measure of randomness (see [Rény61] for a detailed discussion), it is maximized by the uniform distribution and the following inequalities hold:

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

or equivalently

$$1 \leq M_\alpha(p) \leq k^{1-\alpha}, \quad \alpha < 1 \quad \text{and} \quad k^{1-\alpha} \leq M_\alpha(p) \leq 1, \quad \alpha > 1. \quad (3)$$

Furthermore, for $\alpha > 1$, $M_{\alpha+\beta}(p)$ and $M_{\alpha-\beta}(p)$ can be bounded in terms of $M_\alpha(p)$, using the monotonicity of norms and of Hölder means (see, for instance, [HLP52]).

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

$$M_{2\alpha}(p) \leq M_\alpha(p)^2$$

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

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

and

$$M_{\alpha-\beta}(p) \leq k^\beta M_\alpha(p).$$
Proof. By the monotonicity of norms,

\[ M_{\alpha+\beta}(p) \leq M_\alpha(p)^{\frac{\alpha+\beta}{\alpha}} , \]

which gives

\[ \frac{M_{\alpha+\beta}(p)}{M_\alpha(p)^2} \leq M_\alpha(p)^{\frac{\beta}{\alpha} - 1} . \]

The first inequality follows upon choosing \( \beta = \alpha \). For \( 1 < \alpha \) and \( 0 \leq \beta \leq \alpha \), we get the second by (3). For the final inequality, note that by the monotonicity of H"older means, we have

\[ \left( \frac{1}{k} \sum_x p_x^{\alpha-\beta} \right)^{\frac{1}{\alpha-\beta}} \leq \left( \frac{1}{k} \sum_x p_x^\alpha \right)^{\frac{1}{\alpha}} . \]

The final inequality follows upon rearranging the terms and using (3). \( \blacksquare \)

2.2 Bounds on expected moments of a Poisson random variable

Let \( P_\lambda \) be the Poisson distribution with parameter \( \lambda \). We consider Poisson sampling where \( N \sim P_n \) 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 \( P_{np_x} \) and are all independent, leading to simpler analysis. To facilitate our analysis under Poisson sampling, we note a few properties of the expected 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 \( X \sim P_\lambda \). Then, for all \( r \in \mathbb{N} \)

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

and

\[ \text{Var}[X^r] \leq \lambda^r ((\lambda + r)^r - \lambda^r) . \]

**Proof.** The expectation is

\[ \mathbb{E}[X^r] = \sum_{i=0}^{\infty} P_{\lambda,i} \cdot i^r = \sum_{i=r}^{\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \cdot \frac{i!}{(i-r)!} = \lambda^r \sum_{i=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} = \lambda^r . \]

The variance satisfies

\[ \mathbb{E}[(X^r)^2] = \sum_{i=0}^{\infty} P_{\lambda,i} \cdot (i^r)^2 \]

\[ = \sum_{i=r}^{\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \cdot \frac{i!}{(i-r)!} \]

\[ = \lambda^r \sum_{i=0}^{\infty} e^{-\lambda} \cdot \frac{\lambda^i}{i!} \cdot (i + r)^r \]

\[ = \lambda^r \cdot \mathbb{E}[(X + r)^r] . \]
\[
\leq \lambda^r \cdot \mathbb{E} \left[ \sum_{j=0}^{r} \binom{r}{j} X_j \cdot r^{r-j} \right] \\
= \lambda^r \cdot \sum_{j=0}^{r} \binom{r}{j} \cdot \lambda^j \cdot r^{r-j} \\
= \lambda^r (\lambda + r)^r,
\]

where the inequality follows from

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

Therefore,

\[
\text{Var}[X^r] = \mathbb{E}[(X^r)^2] - [\mathbb{E}X^r]^2 \leq \lambda^r \cdot ((\lambda + r)^r - \lambda^r).
\]

The next result establishes a bound on the expected moments of a Poisson random variable. The proof is technical and is relegated to the Appendix.

**Lemma 3.** Let \( X \sim \mathcal{P}_\lambda \) and let \( \beta \) be a positive real number. Then,

\[
\mathbb{E}[X^\beta] \leq c(\beta) \max\{\lambda, \lambda^{\beta}\},
\]

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

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 \( X \sim \mathcal{P}_\lambda \),

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

where the constant \( c \) is given by \( \sqrt{c(2\alpha-2)} \) with \( c(2\alpha-2) \) as in Lemma 3.

**Proof.** For \( \alpha \leq 1 \), \((1 + y)^\alpha \geq 1 + \alpha y - y^2\) for all \( y \in [-1, \infty] \), hence,

\[
X^\alpha = \lambda^\alpha \left(1 + \left(\frac{X}{\lambda} - 1\right)\right)^\alpha \\
\geq \lambda^\alpha \left(1 + \alpha \left(\frac{X}{\lambda} - 1\right) - \left(\frac{X}{\lambda} - 1\right)^2\right).
\]

Taking expectations on both sides,

\[
\mathbb{E}[X^\alpha] \geq \lambda^\alpha \left(1 + \alpha \mathbb{E}\left[\left(\frac{X}{\lambda} - 1\right)\right] - \mathbb{E}\left[\left(\frac{X}{\lambda} - 1\right)^2\right]\right) \\
= \lambda^\alpha \left(1 - \frac{1}{\lambda}\right).
\]

Since \( x^\alpha \) is a concave function and \( X \) is nonnegative, the previous bound yields

\[
|\mathbb{E}[X^\alpha] - \lambda^\alpha| = \lambda^\alpha - \mathbb{E}[X^\alpha] \\
\leq \min(\lambda^\alpha, \lambda^{\alpha-1}).
\]
For $\alpha > 1$, 
\[ |x^\alpha - y^\alpha| \leq \alpha |x - y| (x^{\alpha-1} + y^{\alpha-1}), \]
hence by the Cauchy-Schwarz Inequality,
\[ \mathbb{E}[|X^\alpha - \lambda^\alpha|] \leq \alpha \sqrt{\mathbb{E}[(X - \lambda)^2] \mathbb{E}[(X^{2\alpha-2} + \lambda^{2\alpha-2})]} \]
\[ \leq \alpha \sqrt{\lambda \mathbb{E}[(X^{2\alpha-2} + \lambda^{2\alpha-2})]} \]
\[ \leq \alpha \sqrt{c(2\beta - 2) \max\{\lambda^2, \lambda^{2\alpha-1}\} + \lambda^{2\alpha-1}} \]
where the last-but-one inequality is by Lemma 3.

3 Upper bounds on sample complexity

In this section, we analyze the performance of the empirical estimator $f_\alpha$ and the bias-corrected estimator $g_\alpha$ described in Section 1, and prove the upper bounds on $S_\alpha(k)$ for $\alpha < 1$, noninteger $\alpha > 1$, and integer $\alpha > 1$ in Theorem 8, 7, and 9, respectively. Our proofs are based on bounding the bias and the variance of the two estimators under Poisson sampling. We first describe our general recipe and then analyze the performance of the two estimators separately.

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

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

The following reduction to Poisson sampling is well-known.

Lemma 5. *(Poisson approximation 1)* For $n \geq 8 \log(2/\epsilon)$ and $N \sim \text{Poisson}(n/2)$,

\[ \mathbb{P}\left(|H_\alpha(p) - \tilde{f}_\alpha(X^n)| > \epsilon\right) \leq \mathbb{P}\left(|H_\alpha(p) - \frac{1}{1 - \alpha} \log \tilde{M}_\alpha(n/2, X^n)| > \epsilon\right) + \frac{\epsilon}{2}. \]

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 6. For $N \sim \text{Poisson}(n)$, let the moment estimator $\tilde{M}_\alpha = \tilde{M}_\alpha(n, X^n)$ have bias and variance satisfying

\[ \left| \mathbb{E}[\tilde{M}_\alpha] - M_\alpha(p) \right| \leq \frac{\delta}{2} M_\alpha(p), \]
\[ \text{Var}[\tilde{M}_\alpha] \leq \frac{\delta^2}{12} M_\alpha(p)^2. \]
Then, there exists an estimator $\hat{M}_\alpha'$ that uses $O(n \log(1/\epsilon))$ samples and ensures
\[ \mathbb{P}\left( \left| \hat{M}_\alpha' - M_\alpha(p) \right| > \delta M_\alpha(p) \right) \leq \epsilon. \]

**Proof.** By Chebychev’s Inequality
\[ \mathbb{P}\left( \left| \hat{M}_\alpha - M_\alpha(p) \right| > \delta M_\alpha(p) \right) \leq \mathbb{P}\left( \left| \hat{M}_\alpha - \mathbb{E}[\hat{M}_\alpha] \right| > \delta \frac{1}{2} M_\alpha(p) \right) \leq \frac{1}{3}. \]

To reduce the probability of error to $\epsilon$, we use the estimate $\hat{M}_\alpha$ repeatedly for $O(\log(1/\epsilon))$ independent samples $X^N$ and take the estimate $\hat{M}_\alpha'$ to be the sample median of the resulting estimates. Specifically, let $\hat{M}_1, ..., \hat{M}_t$ denote $t$-estimates of $M_\alpha(p)$ obtained by applying $\hat{M}_\alpha$ to independent sequences $X^N$, and let $\mathbb{I}_{E_i}$ be the indicator function of the event $E_i = \{ |\hat{M}_i - M_\alpha(p)| > \delta M_\alpha(p) \}$.

By the analysis above we have $\mathbb{E}[\mathbb{I}_{E_i}] \leq 1/3$ and hence by Hoeffding’s inequality
\[ \mathbb{P}\left( \sum_{i=1}^{t} \mathbb{I}_{E_i} > \frac{t}{2} \right) \leq \exp(-t/18). \]

The claimed bound follows on choosing $t = 18 \log(1/\epsilon)$ and noting that if more than half of $\hat{M}_1, ..., \hat{M}_t$ satisfy $|\hat{M}_i - M_\alpha(p)| \leq \delta M_\alpha(p)$, 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 $f_\alpha$ and $g_\alpha$ when the number of samples $n$ are of the appropriate order.

**Theorem 7.** For $\alpha > 1$, $\delta > 0$, and $0 < \epsilon < 1$, the estimator $f_\alpha$ satisfies
\[ S^f_\alpha(k, \epsilon, \delta) \leq O\left( \frac{k}{\delta \max\{4, 1/|\alpha - 1|\}} \log \frac{1}{\epsilon} \right). \]

**Proof.** Denote $\lambda_x \overset{\text{def}}{=} np_x$. For $\alpha > 1$, the bias of the moment estimator is bounded using Lemma 4 as follows:
\[
\left| \mathbb{E}\left[ \sum_{x} N_x^\alpha \right]^{\alpha} - M_\alpha(p) \right| \leq \frac{1}{n^\alpha} \sum_{x} \left| \mathbb{E}[N_x^\alpha] - \lambda_x^\alpha \right| \\
\leq \frac{\alpha}{n^{\alpha - 1}} \sum_{x} \left( c\lambda_x + (c + 1)\lambda_x^{\alpha - 1}/2 \right) \\
\leq \frac{\alpha c}{n^{\alpha - 1}} + \frac{\alpha(c + 1)\sqrt{n}}{\sqrt{2}} M_{\alpha - 1/2}(p) \\
\leq \alpha \left( c \left( \frac{k}{n} \right)^{\alpha - 1} + (c + 1)\sqrt{\frac{k}{n}} \right) M_\alpha(p) \tag{4}
\]
where the previous inequality is by Lemma 1 and (3).
Similarly, for bounding the variance, using independence of multiplicities, the following inequalities ensue:

\[
\text{Var} \left[ \sum_x N_x^\alpha \right] = \frac{1}{n^{2\alpha}} \sum_x \text{Var}[N_x^\alpha] \\
= \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - [\mathbb{E}N_x^\alpha]^2 \\
= \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha} + 2\lambda_x^{\alpha} - [\mathbb{E}N_x^\alpha]^2 \\
\leq \frac{1}{n^{2\alpha}} \sum_x |\mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha}| \\
\leq 2\alpha \left( c \left( \frac{k}{n} \right)^{2\alpha - 1} + (c + 1) \sqrt{\frac{k}{n}} \right) M_\alpha(p)^2
\]

(5)

where the last-but-one inequality holds by Jensen’s inequality since \( z^\alpha \) is a convex function; the final inequality is by (4) and Lemma 1. Therefore, the bias and variance are small when \( n = O(k) \) and theorem follows by Lemma 6.

\textbf{Theorem 8.} For \( \alpha < 1, \delta > 0, \) and \( 0 < \epsilon < 1, \) the estimator \( f_\alpha \) satisfies

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

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

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

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

\[
\left| \mathbb{E} \left[ \sum_x N_x^\alpha \frac{N_x}{n^\alpha} \right] - M_\alpha(p) \right| \leq 2 \left( \frac{k^{1/\alpha}}{n} \right)^\alpha \\
\leq 2M_\alpha(p) \left( \frac{k^{1/\alpha}}{n} \right)^\alpha,
\]

(6)

\footnote{For brevity, the constants in (4) and (5), albeit different, are both denoted by \( c. \)}
where the last inequality uses (3). For bounding the variance, note that

\[
\text{Var}\left[\sum_x \frac{N_x^\alpha}{n^\alpha}\right] = \frac{1}{n^{2\alpha}} \sum_x \text{Var}[N_x^\alpha] \\
= \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - [\mathbb{E}N_x^\alpha]^2 \\
\leq \frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha} + \frac{1}{n^{2\alpha}} \sum_x \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2.
\]

Consider the first term on the right-side. For \(\alpha \leq 1/2\), it is bounded above by 0 since \(z^{2\alpha}\) is concave in \(z\), and for \(\alpha > 1/2\) the bound in (5) applies to give

\[
\frac{1}{n^{2\alpha}} \sum_x \mathbb{E}[N_x^{2\alpha}] - \lambda_x^{2\alpha} \leq 2\alpha \left(\frac{c}{n^{2\alpha-1}} + (c + 1)\sqrt{\frac{k}{n}}\right) M_\alpha(p)^2.
\]

For the second term, we have

\[
\sum_x \lambda_x^{2\alpha} - [\mathbb{E}N_x^\alpha]^2 = \sum_x (\lambda_x^{\alpha} - \mathbb{E}[N_x^\alpha]) (\lambda_x^{\alpha} + \mathbb{E}[N_x^\alpha]) \\
\leq 2n^\alpha M_\alpha(p) \left(\frac{k^{1/\alpha}}{n}\right)^\alpha \sum_x (\lambda_x^{\alpha} + \mathbb{E}[N_x^\alpha]) \\
\leq 4n^{2\alpha} M_\alpha(p)^2 \left(\frac{k^{1/\alpha}}{n}\right)^\alpha,
\]

where the last-but-one inequality is by (6) and the last inequality uses the concavity of \(z^\alpha\) in \(z\).

The proof is completed by combining the two bounds above and using Lemma 6. □

**Theorem 9.** For an integer \(\alpha > 1\), any \(\delta > 0\), and \(0 < \epsilon < 1\), the estimator \(g_\alpha\) satisfies

\[
S_{g_\alpha}(k, \delta, \epsilon) \leq O\left(\frac{k(\alpha-1)^{1/\alpha}}{\delta^2 \log \frac{1}{\epsilon}}\right).
\]

**Remark.** Since we use Poisson sampling to simplify the analysis, we chose to remove the bias of \(f_\alpha\) under Poisson sampling to obtain \(g_\alpha\). Alternatively, we can remove the bias under the usual sampling and define

\[
g_\alpha(X^n) = \frac{1}{1 - \alpha} \log \sum_x \frac{(N_x)^\alpha}{n)^\alpha};
\]

a similar result as Theorem 9 can be obtained for this estimator, albeit with a slightly different proof.
Proof. For bounding the variance of \( g_\alpha \), we have
\[
\operatorname{Var}
\left[ \frac{\sum_x N_x^\alpha}{n^\alpha} \right] = \frac{1}{n^{2\alpha}} \sum_x \operatorname{Var}[N_x^\alpha]
\]
\[
\leq \frac{1}{n^{2\alpha}} \sum_x \left( \lambda_x^\alpha (\lambda_x + \alpha)^\alpha - \lambda_x^{2\alpha} \right)
\]
\[
= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} \sum_x \binom{\alpha}{r} \alpha^{\alpha-r} \lambda_x^{\alpha+r}
\]
\[
= \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} n^{\alpha+r} \binom{\alpha}{r} \alpha^{\alpha-r} M_{\alpha+r}(p),
\]
where the inequality uses Lemma 2. It follows from Lemma 1 that
\[
\frac{1}{n^{2\alpha}} \frac{\operatorname{Var}\left[ \sum_x N_x^\alpha \right]}{M_\alpha(p)^2} \leq \frac{1}{n^{2\alpha}} \sum_{r=0}^{\alpha-1} n^{\alpha+r} \binom{\alpha}{r} \alpha^{\alpha-r} M_{\alpha+r}(p) M_\alpha(p)^2
\]
\[
\leq \sum_{r=0}^{\alpha-1} n^{\alpha-r} \binom{\alpha}{r} \alpha^{\alpha-r} k^{\alpha-1}(\alpha-1)/(\alpha)
\]
\[
\leq \sum_{r=0}^{\alpha-1} \left( \frac{\alpha^2 k^{\alpha-1}(\alpha-1)/\alpha}{n} \right)^{\alpha-r}.
\]
Furthermore, by Lemma 2 the estimator is unbiased under Poisson sampling, which completes the proof by Lemma 6.

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 \( H_\alpha(p) \neq H_\alpha(q) \) 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 10. (Poisson approximation 2) Suppose there exist \( \delta, \epsilon > 0 \) such that, with \( N \sim \text{poi}(2n) \), for all estimators \( \hat{f} \) we have
\[
\max_{p \in \mathcal{P}} \mathbb{P} \left( |H_\alpha(p) - \hat{f}_{\alpha}(X^N)| > \delta \right) > \epsilon,
\]
where \( \mathcal{P} \) is a fixed family of distributions. Then, for all fixed length estimators \( \tilde{f} \)
\[
\max_{p \in \mathcal{P}} \mathbb{P} \left( |H_\alpha(p) - \tilde{f}_{\alpha}(X^n)| > \delta \right) > \frac{\epsilon}{2},
\]
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 \( \Phi_l \) is the number of elements \( x \) that appear \( l \) times in the sequence \( X^N \). The following well-known result says that for estimating \( H_\alpha(p) \), it suffices to consider only the functions of the profile.
Lemma 11. (Sufficiency of profiles). Consider an estimator $\hat{f}$ such that
\[ P\left(|H_\alpha(p) - \hat{f}(X^N)| > \delta\right) \leq \epsilon, \quad \text{for all } p. \]

Then, there exists an estimator $\tilde{f}(X^N) = \tilde{f}(\Phi)$ such that
\[ P\left(|H_\alpha(p) - \tilde{f}(\Phi)| > \delta\right) \leq \epsilon, \quad \text{for all } p. \]

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 12. If for two distributions $p$ and $q$ on $\mathcal{X}$ the variational distance $\|p - q\| < \epsilon$, then one of the following holds for every function $\hat{f}$:
\[
p\left(|H_\alpha(p) - \hat{f}(X)| \geq \frac{|H_\alpha(p) - H_\alpha(q)|}{2}\right) \geq \frac{1 - \epsilon}{2},
\]
\[\text{or} \quad q\left(|H_\alpha(q) - \hat{f}(X)| \geq \frac{|H_\alpha(p) - H_\alpha(q)|}{2}\right) \geq \frac{1 - \epsilon}{2}.\]

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 exist two distributions $p$ and $q$ of support-size $k$ such that the following hold:

(i) There exists $\delta > 0$ such that
\[ |H_\alpha(p) - H_\alpha(q)| > \delta; \quad (7) \]

(ii) denoting by $p_\Phi$ and $q_\Phi$, respectively, the distributions on the profiles under Poisson sampling corresponding to underlying distributions $p$ and $q$, there exist $\epsilon > 0$ such that
\[ \|p_\Phi - q_\Phi\| < \epsilon, \quad (8) \]
\[ \text{if } n < k^{c(\alpha)}. \]

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 (8) in terms of the differences of moments $|M_\alpha(p) - M_\alpha(q)|$.

Theorem 13. [Val08] Given distributions $p$ and $q$ such that
\[ \max_x \max \{p_x; q_x\} \leq \frac{\epsilon}{40n}, \]
for Poisson sampling with $N \sim P_n$, it holds that
\[ \|p_\Phi - q_\Phi\| \leq \frac{\epsilon}{2} + 5 \sum_a n^n |M_\alpha(p) - M_\alpha(q)|. \]
It remains to construct the required distributions $p$ and $q$, satisfying (7) and (8) above. By Theorem 13, the variation distance $\|p_\Phi - q_\Phi\|$ can be made small by ensuring that the moments of distributions $p$ and $q$ are matched, that is, we need distributions $p$ and $q$ with different Rényi entropies and identical moments for as large an order as possible. To that end, for every positive integer $d$ and every vector $x = (x_1, ..., x_d) \in \mathbb{R}^d$, associate with $x$ a distribution $p^x$ of support-size $dk$ such that

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

Note that

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

and for all $a$

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

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

**Lemma 14.** For every $d \in \mathbb{N}$ and $\alpha$ not integer, there exist positive vectors $x, y \in \mathbb{R}^d$ such that

- $\|x\|_r = \|y\|_r$, $1 \leq r \leq d - 1$,
- $\|x\|_d \neq \|y\|_d$,
- $\|x\|_\alpha \neq \|y\|_\alpha$.

A constructive proof of Lemma 14 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 15.** Given an integer $\alpha > 1$ and any estimator $f$ of $H_\alpha(p)$, 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

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

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 $p^x$ and $p^y$, where the vectors $x$ and $y$ are given by Lemma 14. In view of the foregoing discussion, we need to verify (7) and (8) to prove the theorem. Therefore, (7) holds by Lemma 14 since

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

and for $n < C_2k^{(d-1)/d}$ and $5C_2^d/(1 - C_2) < \epsilon/2$, inequality (8) follows from Theorem 13 as

$$\|p_\Phi - q_\Phi\| \leq \frac{\epsilon}{2} + 5 \sum_{a \geq d} \left( \frac{n}{k^{(\alpha - 1)/\alpha}} \right)^a \leq \epsilon.$$  

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

**Theorem 16.** Given a nonintegral $\alpha > 1$, for every $0 < \epsilon < 1/2$, we have

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

**Remark.** We can obtain a slightly stronger lower bound on $S_\alpha(k, \delta, \epsilon)$ for noninteger $\alpha > 1$, with a minor modification of our proof, upon using [Val08, Corollary 1]. In particular, the statement of the theorem above holds with

$$S_\alpha(k, \delta, \epsilon) \geq \frac{Ck}{2\epsilon \sqrt{\log k}},$$

for appropriately chosen constants $C$ and $c$.

**Proof.** For a fixed $d$, let distributions $p$ and $q$ be as in the previous proof. Then, as in the proof of Theorem 16, inequality (7) holds by Lemma 14 and (8) holds by Theorem 13 if $n < C_2k^{(d-1)/d}$. The theorem follows since $d$ can be arbitrary large. $lacksquare$

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

**Theorem 17.** Given $\alpha < 1$, for every $0 < \epsilon < 1/2$, we have

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

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

$$p_{ij} = \frac{p_{ij}}{k^\beta} \quad \text{and} \quad q_{ij} = \frac{q_{ij}}{k^\beta}, \quad 1 \leq i \leq d, \ 1 \leq j \leq k,$$

where the vectors $x$ and $y$ are given by Lemma 14 and $\beta$ satisfies $\alpha(1 + \beta) < 1$, and

$$p_0 = q_0 = 1 - \frac{1}{k^\beta}.$$

For this choice of $p$ and $q$, we have

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

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

and similarly for $q$, which further yields

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

Therefore, for sufficiently large $k$, (7) holds by Lemma 14 since $\alpha(1 + \beta) < 1$, and for $n < C_2k^{(1+\beta-1)/d}$ we get (8) by Theorem 13 as

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

The theorem follows since $d$ and $\beta < 1/\alpha - 1$ are arbitrary. $lacksquare$
We close with a proof of Lemma 14.

**Proof of Lemma 14.** Let \( x = (1, ..., d) \). Consider the polynomial

\[
p(z) = (z - x_1)...(z - x_d),
\]

and \( q(z) = p(z) - \Delta \), where \( \Delta \) 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 \( d \)th 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 \( \Delta \) 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, \( \|x\|_\alpha \neq \|y\|_\alpha \) for all \( \Delta \) sufficiently small. 

\( \blacksquare \)
5 Concluding remarks

In this paper, we studied the complexity of estimating Rényi entropy in terms of the support size $k$ of the underlying unknown distribution. In a PAC framework, for vanishing estimation error, we characterize the dependence of sample complexity on the support-size $k$ of the underlying distribution. The proposed algorithms are efficient and have time complexity linear in the number of samples $n$. The general tradeoff between $k$ and $n$, and the estimation error $\delta$ remains open. In particular, it remains unclear if our lower bounds still hold when $\delta$ is fixed arbitrarily.

A natural extension of this work entails considering the estimation of general additive functions of a distribution within a small multiplicative loss. Of particular interest is the case when the observed samples consist of pairs $(Y_i, Z_i)$ generated independently from distributions $p$ and $q$, respectively. For this case, our lower bounds already suggest that the problem of estimating the Rényi divergence [Rén61] between $p$ and $q$ by observing the samples is a difficult one. In fact, even if one of the distribution is known, reliable estimates cannot be formed by using less than $O(k^{1-\eta})$ samples for all $\eta > 0$.

Acknowledgements

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

References


21


Lemma 3. Let $X \sim P_\lambda$ and let $\beta$ be a positive real number. Then,

$$
E[X^\beta] \leq c(\beta) \max\{\lambda, \lambda^\beta\},
$$

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

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

$$
E[X^\beta] \leq \sum_{i \leq 2\lambda} P_{\lambda,i} i^\beta + \sum_{i > 2\lambda} P_{\lambda,i} i^\beta
\leq 2^\beta \lambda^\beta + \sum_{i > 2\lambda} P_{\lambda,i} i^\beta.
$$

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

$$
P_{\lambda,i} \leq \mathbb{P}(X \geq i) \leq \exp\left(-\frac{i - \lambda}{8\lambda}\right),
$$

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

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

For $\lambda \leq 1$, since $\lambda^i \leq \lambda$ for all $i \geq 1$

$$
E[X^\beta] \leq \lambda \sum_{i=1}^\infty \frac{i^\beta}{i!}, \quad (9)
$$

and the lemma follows upon choosing

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

which is a finite quantity.