Renyi Entropy Estimation Revisited*

Maciej Obremski[1] and Maciej Skorski[2]

1 **Aarhus University, Aarhus, Denmark**†
obremski@cs.au.dk
2 **IST Austria, Klosterneuburg, Austria**‡
maciej.skorski@gmail.com

Abstract

We revisit the problem of estimating entropy of discrete distributions from independent samples, studied recently by Acharya, Orlitsky, Suresh and Tyagi (SODA 2015), improving their upper and lower bounds on the necessary sample size n. For estimating Renyi entropy of order α, up to constant accuracy and error probability, we show the following

- Upper bounds $n = O(1) \cdot 2^{\left(1-\frac{1}{\alpha}\right)H_\alpha}$ for integer $\alpha > 1$, as the worst case over distributions with Renyi entropy equal to H_α.
- Lower bounds $n = \Omega(1) \cdot K^{1-\frac{1}{\alpha}}$ for any real $\alpha > 1$, with the constant being an inverse polynomial of the accuracy, as the worst case over all distributions on K elements.

Our upper bounds essentially replace the alphabet size by a factor exponential in the entropy, which offers improvements especially in low or medium entropy regimes (interesting for example in anomaly detection). As for the lower bounds, our proof explicitly shows how the complexity depends on both alphabet and accuracy, partially solving the open problem posted in previous works.

The argument for upper bounds derives a clean identity for the variance of falling-power sum of a multinomial distribution. Our approach for lower bounds utilizes convex optimization to find a distribution with possibly worse estimation performance, and may be of independent interest as a tool to work with Le Cam's two point method.

1998 ACM Subject Classification G.1.2 Approximation, G.3 Statistical Computing

Keywords and phrases Renyi entropy, entropy estimation, sample complexity, convex optimization

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2017.20

1 Introduction

1.1 Renyi Entropy

Renyi entropy [25] arises in many applications as a generalization of Shannon Entropy [27]. It is also of interests on its right, with a number of applications including unsupervised learning (like clustering) [30, 12], multiple source adaptation [17], image processing [16, 20, 26], password guessability [3, 24, 10], network anomaly detection [15], quantifying neural activity [22] or to analyze information flows in financial data [13].

In particular Renyi entropy of order 2, known also as collision entropy, is used in quality tests for random number generators [14, 29], to estimate the number of random bits

* Available at http://eprint.iacr.org/2017/588.pdf
† This project has received funding from the European research Council (ERC) under the European Unions's Horizon 2020 research and innovation programme (grant agreement No 669255).
‡ Supported by the European Research Council consolidator grant (682815-TOCNeT).

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017).
Editors: Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala; Article No. 20; pp. 20:1–20:15

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Algorithm 1: Estimation of Renyi Entropy

Input: entropy parameter $\alpha > 1$ (integer),
alphabet $\mathcal{A} = \{a_1, \ldots, a_K\}$,
samples $x_1, \ldots, x_n$ from an unknown distribution p on $\mathcal{A}$

Output: number H approximating the α-entropy of p

$I \leftarrow \{i : \exists j \quad a_i = x_j\}$ /* compute the list of occurring symbols[1] */
for $i \in I$ **do**
  $n_i \leftarrow \#\{j : x_j = a_i\}$ /* compute empirical frequencies */
end
$M \leftarrow \sum_i \frac{n_i^{\underline{\alpha}}}{n^{\underline{\alpha}}}$ /* bias-corrected power sum estimation by falling powers[2] */
$H \leftarrow \frac{1}{1-\alpha} \log M$ /* entropy from power sums */
return H

that can be extracted from a physical source [11, 7], characterizes security of certain key derivation functions [4, 8], helps testing graph expansion [9] and closeness of distributions to uniformity [6, 23] and bounds the number of reads needed to reconstruct a DNA sequence [19].

1.2 Estimation and Sample Complexity

Motivated by the discussed applications, algorithms that estimate Renyi entropy of an unknown distribution from samples were proposed for discrete [31] and also for continuous distributions [21]. For Shannon entropy, estimators with multiplicative errors were studied in [5] and follow-up works; the existence of sublinear (in terms of the alphabet size) additive estimators was proved in [22], and the optimal additive estimator was given in [28]. For the general case of Renyi entropy, the state of the art was established in [1], with upper and lower bounds on the sample complexity.

Interestingly, the estimation of Renyi entropy of integer orders $\alpha \geqslant 1$ is *sublinear* in the alphabet size. More precisely, to estimate the entropy of an integer order $\alpha > 1$ of a distribution over an alphabet of size K, with a *constant accuracy and constant error probability*, one needs

$$n = \Theta(K^{1-\frac{1}{\alpha}})$$

samples. On the other hand, the necessary sample size for non-integer $\alpha > 1$ is

$$n = \Omega(K^{1-o(1)}),$$

with the upper bound $O(K/\log K)$, for large K and the accuracy sufficiently small [1, 2].

The estimator itself is a biased-reduced adaptation of the naive "plug-in" estimator. Note that computing empirical frequencies as estimates to true probabilities and putting them straight into the entropy formula (which we refer to as naive estimation) would yield a biased estimator. To obtain better convergence properties, one needs to add some *corrections* to the formula. In the case of Renyi entropy, one replaces powers of empirical frequencies in the entropy formula by *falling powers*, obtaining better estimator with the complexity bounds discussed above [1]. See Algorithm 1 for the pseudocode.

[2] Storing and updating empirical frequencies can be implemented with different data structures, we don't discuss the optimal solution as our primary interest is in the sample complexity.

[2] Here $z^{\underline{\alpha}}$ stands for the falling α-power of the number z.

1.3 Our contribution

1.3.1 Results

We revisit the analysis of the minimal number of samples n (sample complexity) needed to estimate Renyi entropy up to certain additive accuracy, obtaining improvements upon the result in [1]. In the presentation below we consider the estimation up to constant error probability, unless stated otherwise.

(a) Better upper bounds for the sample complexity, with a simplified analysis:

$$n = O\left(2^{\left(1-\frac{1}{\alpha}\right)H_\alpha}\delta^{-2}\right), \quad \text{for integer } \alpha > 1$$

valid for Algorithm 1, any accuracy $\delta > 0$, and all distributions with Renyi entropy of order α equal to H_α

(b) Lower bounds for non-integer $\alpha > 1$, explicit w.r.t. both alphabet and accuracy:

$$n = \Omega(1) \cdot \max\left(\delta^{-\frac{1}{\alpha}}K^{1-\frac{1}{\alpha}}, \delta^{-\frac{1}{2}}K^{\frac{1}{2}}\right), \quad \text{for any non-integer } \alpha > 1$$

valid for any estimator, any accuracy $\delta \leqslant 1$ and some distribution over K elements.

(c) Refining the technique for proving lower bounds; we explain how to obtain optimal bounds for the ideas used in [1]; our construction for lower bounds is also simpler.

The first improvement essentially parameterizes the previous bound by the entropy amount, and is of interest in *medium/low entropy* regimes. Note that when the entropy is at most a half of the maximal amount ($H_\alpha \leqslant \frac{1}{2}\log K$) then the complexity drops to $n = O(K^{\frac{1}{2}})$ even for most demanding min-entropy ($\alpha = \infty$). The improvements may be relevant for anomaly detection algorithms based on evaluating entropy of data streams [15]. The precise statement, which addresses arbitrary accuracy and error probability, appears in Corollary 7.

The lower bounds given in [1] and improved in the journal version [2] depend only on the alphabet, and are valid for large K and sufficiently small δ. As opposed to that, our lower bounds apply to all regimes of K and δ and explicitly show that *large alphabets and small accuracy both contribute to the complexity*. Thus we make a progress[3] towards understanding how the sample complexity depends on δ and K, which is an open problem except for integer α [2]. In particular, our results show that the sample complexity may be much bigger than $\Omega\left(K^{1-o(1)}\right)$ for δ being small depending on K, which is not guaranteed by the previous results (e.g. Table 1 in [2]).

The technique for lower bound in [1] essentially boils down to the construction of two statistically close distributions that differ in entropy (the technique known as *Le Cam's two-point method*). The authors obtained implicitly a suboptimal pair with this property. We instead construct explicitly a simpler pair with much better properties.

1.3.2 Techniques

The original proof of the upper bounds proceeds by estimating the variance of the falling-power sum in Line 5 in Algorithm 1. This analysis is somewhat difficult because the empirical frequencies n_i in Line 3 are not independent. A workaround proposed in [1] uses *Poisson sampling* to randomize the number n in a convenient way (which doesn't hurt the convergence

[3] Our result is worse in the dependency on K, but the added value is the dependency on δ.

Table 1 Our lower bounds for estimation of Renyi entropy of order α. By K we denote the alphabet size, δ is the additive error of estimation, $\Omega(1)$ is an absolute constant.

Entropy	Accuracy	Sample Complexity
$1 < \alpha < 2$	$\delta \leqslant 1$	$\Omega(1) \cdot \min\left(\delta^{-\frac{1}{2}} K^{\frac{1}{2}}, \delta^{-\alpha} K^{1-\frac{1}{\alpha}}\right)$
	$\delta > 1$	$\Omega(1) \cdot \min\left(\left(2^{-\delta} K\right)^{\frac{1}{2}}, 2^{-\left(1-\frac{1}{\alpha}\right)\delta} K^{1-\frac{1}{\alpha}}\right)$
$2 \leqslant \alpha$	$\delta \leqslant 1$	$\Omega(1) \cdot \delta^{-\frac{1}{\alpha}} K^{1-\frac{1}{\alpha}}$
	$\delta > 1$	$\Omega(1) \cdot \left(2^{-\left(1-\frac{1}{\alpha}\right)\delta} K\right)^{1-\frac{1}{\alpha}}$

much), so that the frequencies are independent and the variance of power sums can directly computed.

We get rid of the Poisson sampling, by showing that the falling-power sum obeys a nice and clean algebraic identity, that can be further used to compute the variance (see Lemma 4). We believe that our technique may be of benefit to related problems, e.g. when estimating moments for streaming algorithms.

The argument for lower bounds in [1] starts by modifying the estimator so that it is a function of empirical frequencies (called *profiles* in [1]). Then, by certain facts on zeros of polynomials and exponential sums, one exhibits two probability distributions with certain relations between power sums. As a conclusion, again under Poisson sampling, one obtains two distributions such that their profiles differ much in entropy, yet are close in total variation. This yields a contradiction unless n is big enough.

Our approach deviates from these techniques. We share the same core idea, that estimation should be continuous in total variation, yet use it to conclude a clear bound without referring to profiles: if distributions are γ-close and the entropy differs by δ, the number n must satisfy $n = \Omega(\gamma^{-1})$ (see Corollary 9). It remains to construct two such distributions with possibly small γ and possibly big δ. By solving the related optimization task (which we do by an elegant application of *majorization theory*), we conclude that a simpler and better choice is one distribution being flat, and other being a combination of a flat distribution with a unit mass (see the proof of Lemma 11). We remark that our optimization approach not only gives better lower bounds for Renyi entropy, but may be also applied to similar estimation problems, e.g. lower bounds on the complexity for estimating functionals of a discrete distribution. The lower bounds are summarized in Table 1.

2 Preliminaries

For any natural α and real number x, by $x^{\underline{\alpha}} \stackrel{def}{=} \prod_{i=0}^{\alpha-1}(x-i)$ we denote the α-th falling power of x, with the convention $x^{\underline{0}} = 1$. If a discrete random variable X has a probability distribution p, we denote $p(x) = \Pr[X = x]$. For any distribution X by X^n we denote the n-fold product of independent copies of X. The moment of a distribution p of order α equals $p_\alpha = \sum_x p(x)^\alpha$. Through the paper, we use logarithms at base 2.

▶ **Definition 1** (Total variation (statistical closeness))**.** For two distributions p, q over the same finite alphabet the total variation equals $d_{TV} = \frac{1}{2}\sum_x |p(x) - q(x)|$. If $d_{TV}(p,q) \leqslant \epsilon$ we also say that p and q are ϵ-close.

▶ **Definition 2** (Renyi Entropy). The Renyi entropy of order α for $\alpha > 1$ equals

$$H_\alpha(p) \stackrel{def}{=} -\frac{1}{\alpha-1}\log\left(\sum_x p(x)^\alpha\right) = -\frac{1}{\alpha-1}\log p_\alpha.$$

Sometimes for shortness we simply say "α-entropy", referring to Renyi entropy of order α.

▶ **Definition 3** (Entropy Estimators). Given an alphabet $\mathcal{X}$ and a fixed number n we say that an algorithm $\hat{f}$ provides a (δ,ϵ)-approximation to α-entropy if for any distribution p over $\mathcal{X}$

$$|\hat{f}(x_1,\ldots,x_n) - H_\alpha(p)| > \delta$$

holds with probability at most ϵ over samples $x_1,\ldots,x_n$ drawn independently from p.

3 Auxiliary Facts

Define $\xi_i(x) = [X_i = x]$ and the *empirical frequency* of the symbol x by

$$n(x) = \sum_{i=1}^n \xi_i(x). \tag{1}$$

Note that the vector $(n(x))_{x\in\mathcal{X}}$ follows a multinomial distribution with sum n and probabilities $(p(x))_{x\in\mathcal{X}}$. The lemma below states that we have very simple expressions for the falling powers of $n(x)$.

▶ **Lemma 4** (Falling powers of empirical frequencies). *For every x we have*

$$n(x)^{\underline{\alpha}} = \sum_{i_1\neq i_2\neq\ldots\neq i_\alpha} \xi_{i_1}(x)\xi_{i_2}(x)\cdot\ldots\cdot\xi_{i_\alpha}(x). \tag{2}$$

In particular, we have

$$\mathbb{E}\left[\sum_x n(x)^{\underline{\alpha}}\right] = n^{\underline{\alpha}} p_\alpha. \tag{3}$$

The proof appears in Appendix A. We also obtain the following closed-form expressions for the variance of the sum of falling powers.

▶ **Lemma 5** (Variance of frequency falling powers sums). *We have*

$$\mathrm{Var}\left[\sum_x n(x)^{\underline{\alpha}}\right] = n^{\underline{\alpha}}((n-\alpha)^{\underline{\alpha}} - n^{\underline{\alpha}})p_\alpha^2 + \sum_{\ell=1}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\; p_{2\alpha-\ell}. \tag{4}$$

The proof appears in Appendix B.

4 Upper Bounds

Similarly as in [1], we observe that to estimate Renyi entropy with additive accuracy $O(\delta)$, it suffices to estimate power sums with multiplicative accuracy $O(\delta)$.

▶ **Theorem 6** (Estimator Performance). *The number of samples needed to estimate p_α up to a multiplicative error δ and error probability ϵ equals $n = O_\alpha\left(2^{\frac{\alpha-1}{\alpha}\cdot H_\alpha(p)}\delta^{-2}\log(1/\epsilon)\right)$.*

From this result one immediately obtains

▶ **Corollary 7.** *The number of samples needed to estimate $H_\alpha(p)$, up to an additive error δ and error probability ϵ, equals $n = O_\alpha\left(2^{\frac{\alpha-1}{\alpha}\cdot H_\alpha(p)}\delta^{-2}\log(1/\epsilon)\right)$. The matching estimator is Algorithm 1.*

Proof of Theorem 6. It suffices to construct an estimator with error probability $\frac{1}{3}$. We can amplify this probability to ϵ with a loss of a factor of $O(\log(1/\epsilon))$ in the sample size, by a standard argument: running the estimator in parallel on fresh samples and taking the median (as in [1]).

From Lemma 5 we conclude that the variance of the estimator equals

$$\mathrm{Var}[\mathsf{Est}] = -\Theta_\alpha(1)\cdot n^{-1}(p_\alpha)^2 + \sum_{\ell=1}^{\alpha}\Theta_\alpha(1)\cdot n^{-\ell}p_{2\alpha-\ell},$$

where $\Theta_\alpha(1)$ are constants dependent on α. Note that we have

$$p_{2\alpha-\ell} \leqslant (p_\alpha)^{\frac{2\alpha-\ell}{\alpha}}$$

by elementary inequalities[4], and therefore

$$\mathrm{Var}[\mathsf{Est}] = O_\alpha(1)\cdot p_\alpha^2\sum_{\ell=1}^{\alpha}\left(np_\alpha^{\frac{1}{\alpha}}\right)^{-\ell} = O_\alpha(1)\cdot n^{-1}p_\alpha^{2-\frac{1}{\alpha}}\sum_{\ell=0}^{\alpha-1}\left(np_\alpha^{\frac{1}{\alpha}}\right)^{-\ell}.$$

Note that the negative term $-\Theta_\alpha(1)n^{-1}(p_\alpha)^2$ we skipped is of smaller order than the term $\ell = 1$ of the sum on the right hand side, so it doesn't help to improve the bounds. For $n > 2p_\alpha^{\frac{1}{\alpha}}$ the right hand side equals $O_\alpha(1)\cdot n^{-1}p_\alpha^{1-\frac{1}{\alpha}}$. By the Chebyszev Inequality

$$\Pr_{X^n\sim p}[|\mathsf{Est}(X^n)) - p_\alpha| > \delta p_\alpha] < \frac{\mathrm{Var}[\mathsf{Est}]}{\delta^2 p_\alpha^2} = O_\alpha(1)\cdot n^{-1}p_\alpha^{-\frac{1}{\alpha}}\delta^{-2},$$

which is smaller than $\frac{1}{3}$ for some $n = O_\alpha(1)\cdot p_\alpha^{-\frac{1}{\alpha}}\delta^{-2}$. ◀

5 Lower Bounds

We will need the following lemma, stated in a slightly different way in [1]. It captures the intuition that if two distributions differ much in entropy, then they must be far away in total variation (otherwise the estimator, presumably working well, would distinguish them).

▶ **Lemma 8** (Estimation is continuous in total variation)**.** *Suppose that $\hat{f}$ is a (δ,ϵ)-estimator for H_α. Then the following is true:*

$$\forall X,Y \quad |H_\alpha(X) - H_\alpha(Y)| \geqslant 2\delta \Rightarrow d_{TV}(X^n;Y^n) \geqslant 1-2\epsilon. \tag{5}$$

The proof is illustrated on Figure 1 and appears in Appendix C. By combining Lemma 8 with a simple inequality $d_{TV}(X^n,Y^n) \leqslant n\cdot d_{TV}(X,Y)$ (which can be proved by a hybrid argument) we obtain

▶ **Corollary 9.** *Let X,Y be such that (a) $d_{TV}(X;Y) \leqslant \gamma$ and (b) $|H_\alpha(X) - H_\alpha(Y)| \geqslant 2\delta$. Then any (δ,ϵ)-estimator for H_α, where $\epsilon \leqslant \frac{1}{3}$, requires $\frac{1}{3}\gamma^{-1}$ samples.*

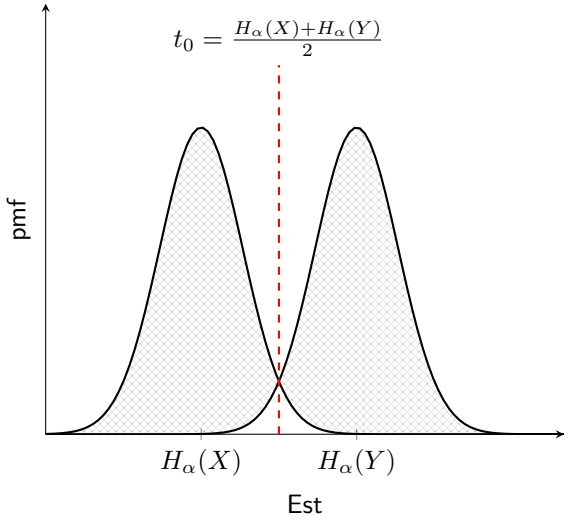

Figure 1 Turning estimators into distinguishers in total variation.

We will need the following inequalities, that refine the known Bernoulli-inequality $(1+u)^\alpha \geqslant 1+\alpha u$ by introducing higher-order terms.

▶ **Proposition 10** (Bernouli-type inequalities). *We have*

$$\forall \alpha > 1,\ \forall u > -1: \quad (1+u)^\alpha \geqslant 1+\alpha u \tag{6}$$

$$\forall \alpha \geqslant 2,\ \forall u > 0: \quad (1+u)^\alpha \geqslant 1+\alpha u + u^\alpha \tag{7}$$

$$\forall \alpha \in [1,2],\ \forall u \in [0,1]: \quad (1+u)^\alpha \geqslant 1+\alpha u + \frac{\alpha(\alpha-1)}{4}u^2 \tag{8}$$

$$\forall \alpha \in [1,2],\ \forall u > 1: \quad (1+u)^\alpha \geqslant 1+\alpha u + \frac{\alpha-1}{3}u^\alpha \tag{9}$$

Proof. To prove Equation (6) consider the function $f(u) = (1+u)^\alpha$. It is convex when $\alpha > 1$, hence its graph is above the tangent line at $u = 0$. This means that $f(u) \geqslant f(0) + \frac{\partial f}{\partial}(0)u$, and since $f(0) = 1$ and $\frac{\partial f}{\partial u}(0) = \alpha$ the inequality follows.

In order to prove Equation (7), we consider the function $f(u) = (1+u)^\alpha - 1 - \alpha u - u^\alpha$. Its derivative equals $\frac{\partial f}{\partial u}(u) = \alpha\left((1+u)^{\alpha-1} - u^{\alpha-1} - 1\right)$. If we show it is non-negative for $u \geqslant 0$, we establish the claimed inequality as then $f(u) \geqslant f(0) \geqslant 0$. We calculate the second derivative $\frac{\partial^2 f}{\partial u^2}(u) = \alpha(\alpha-1)\left((1+u)^{\alpha-2} - u^{\alpha-2}\right)$ and see it is positive when $u \geqslant 0$ (here we use the assumption that $\alpha \geqslant 2$). We conclude that $\frac{\partial f}{\partial u}(u)$ is increasing for $u \geqslant 0$ and hence $\frac{\partial f}{\partial u}(u) \geqslant \frac{\partial f}{\partial u}(0) = 0$, which finishes the proof.

To prove Equation (8) we define $f = (1+u)^\alpha - 1 - \alpha u - \frac{\alpha(\alpha-1)}{4}u^2$. We note that $\frac{\partial f}{\partial u}(u) = \alpha(1+u)^{\alpha-1} - \alpha - \frac{\alpha(\alpha-1)}{2}u$. This function is concave because $\alpha \in [1,2]$. Since $\frac{\partial f}{\partial u}(0) = 0$ and $\frac{\partial f}{\partial u}(1) = \alpha(1+1)^{\alpha-1} - \alpha - \frac{\alpha(\alpha-1)}{2} \geqslant \alpha^2 - \alpha - \frac{\alpha(\alpha-1)}{2} = \frac{1}{2}(\alpha^2-\alpha) \geqslant 0$ (we have used the Bernouli inequality $(1+1)^{\alpha-1} \geqslant 1+\alpha-1$), by concavity we conclude that the $\frac{\partial f}{\partial u}(u) \geqslant 0$ on the whole interval $u \in [0,1]$. This means that f is decreasing and $f(u) \geqslant f(0) = 0$ for $u \in [0,1]$, which establishes the claimed inequality.

[4] We use the fact that α-norms, defined by $\|p\|_\alpha = \left(\sum_i |p_i|^\alpha\right)^{\frac{1}{\alpha}}$, are decreasing in α. The same inequality is applied in [1], the proof of Lemma 2.1.

To obtain Equation (9) we consider the function $f(u) = (1+u)^\alpha - 1 - \alpha u - Cu^\alpha$. Its derivative equals $\frac{\partial f}{\partial u}(u) = \alpha\left((1+u)^{\alpha-1} - 1 - Cu^{\alpha-1}\right)$. It suffices to choose C such that $f(1) \geqslant 0$ and $\frac{\partial f}{\partial u}(u) \geqslant 0$ for $u \geqslant 1$ as then $f(u) \geqslant 1$ for $u \geqslant 1$. The second derivative equals $\frac{\partial^2 f}{\partial u^2}(u) = \alpha(\alpha-1)\left((1+u)^{\alpha-2} - Cu^{\alpha-2}\right)$, and we conclude that, for $1 \leqslant \alpha \leqslant 2$ and $u \geqslant 1$, it bigger than zero when $C \leqslant 2^{\alpha-2}$. Thus the first derivative increases and is non-negative if, in addition, $\frac{\partial f}{\partial u}(1) \geqslant 0$, that is $C \leqslant 2^{\alpha-1} - 1$. We conclude that $f(u) \geqslant 0$ with $C = \min\left(2^{\alpha-2}, 2^{\alpha-1} - 1, 2^\alpha - \alpha - 1\right)$, that is when $\frac{\partial^2 f}{\partial u^2}(1), \frac{\partial f}{\partial u}(1), f(1)$ are all non-negative. Under the assumption $\alpha \leqslant 2$ this can be simplified to $C = 2^\alpha - 1 - \alpha$. We notice further that $2^{\alpha-1} - 1 - \alpha \geqslant (\ln 4 - 1)(\alpha - 1)$ when $\alpha \in (1,2)$ which shows that we can take $C = 0.38(\alpha - 1)$. ◀

▶ **Lemma 11** (Distributions with different entropy yet close in total variation). *For any real $\alpha > 1$ and any set S of size $K \geqslant 2$ there exist distributions on S that are γ-close but with Renyi α-entropy different by at least Δ, for any parameters satisfying the following*

- *For any $\Delta \leqslant 1$, any $\alpha \in [1,2]$ and $\gamma = O\left(\max\left(\Delta^{\frac{1}{2}} K^{-\frac{1}{2}}, K^{-1+\frac{1}{\alpha}} \Delta^{\frac{1}{\alpha}}\right)\right)$*
- *For any $\Delta \leqslant 1$, any $\alpha > 2$ and $\gamma = O\left(\Delta^{\frac{1}{\alpha}} K^{-1+\frac{1}{\alpha}}\right)$*
- *For any $\Delta \geqslant 1$, any $\alpha \in [1,2]$ and $\gamma = \max\left(2^{\left(1-\frac{1}{\alpha}\Delta\right)} K^{-1+\frac{1}{\alpha}}, 2^{\frac{1}{2}\Delta} K^{-\frac{1}{2}}\right)$*
- *For any $\Delta \geqslant 1$, any $\alpha > 2$ and $\gamma = O\left(2^{\left(1-\frac{1}{\alpha}\right)\Delta} K^{-1+\frac{1}{\alpha}}\right)$*

In particular, by applying Corollary 9 to the setting in the lemma above, we obtain the lower bounds on the sample complexity.

▶ **Corollary 12** (Estimating entropy with constant additive error). *For any constant $\alpha > 1$, estimating α-entropy with additive error at most 1 requires at least $\Omega(1) \cdot \max\left(K^{\frac{1}{2}}, K^{1-\frac{1}{\alpha}}\right)$ samples. More generally bounds (for any accuracy Δ) apply as shown in Table 1.*

Proof of Lemma 11. Fix a K-element set S and a parameter $\epsilon > 0$ and consider the following pair of distributions (given the choice of X, the choice of Y is close to the "worst" choice as shown in Section D):

(a) X is uniform over S,

(b) Y puts a mass of $\frac{1}{K} + \gamma$ on one fixed point of S and $\frac{1}{K} - \frac{\gamma}{K-1}$ on the remaining points of S,

where the exact value of the parameter γ is to be optimized later. We calculate that

$$\sum_x (P_Y(x))^\alpha = \left(K^{-1} + \gamma\right)^\alpha + (K-1)\left(K^{-1} - \gamma(K-1)^{-1}\right)^\alpha$$

and

$$K^\alpha \cdot \sum_x (P_Y(x))^\alpha = (1 + K\gamma)^\alpha + (K-1)\left(1 - \gamma\frac{K}{K-1}\right)^\alpha.$$

Since $\sum_x (P_X(x))^\alpha = K^{1-\alpha}$ we get

$$\frac{\sum_x (P_Y(x))^\alpha}{\sum_x (P_X(x))^\alpha} = K^{-1}\left((1+K\gamma)^\alpha + (K-1)\left(1 - \gamma\frac{K}{K-1}\right)^\alpha\right). \tag{10}$$

Now if either $K\gamma \leqslant 1$ and $\alpha \in (1,2)$ or $\alpha \geqslant 2$, by Proposition 10 we obtain

$$(1+K\gamma)^\alpha + (K-1)\left(1 - \gamma\frac{K}{K-1}\right)^\alpha \geqslant K + \Omega_\alpha(1)\min\left((K\gamma)^2, (K\gamma)^\alpha\right) \tag{11}$$

for some constants depending on α, where we have used Equation (6) to lower-bound $\left(1-\gamma\frac{K}{K-1}\right)^{\alpha}$ and Equations (8) and (7) to lower-bound $(1+K\gamma)^{\alpha}$. More precisely, we have

$$(1+K\gamma)^{\alpha}+(K-1)\left(1-\gamma\frac{K}{K-1}\right)^{\alpha}\geqslant\begin{cases}K+\frac{\alpha-1}{3}(K\gamma)^{\alpha} & \text{if } \alpha\in(1,2)\wedge K\gamma>1\\ K+\frac{\alpha(\alpha-1)}{4}(K\gamma)^{2} & \text{if } \alpha\in(1,2)\wedge K\gamma\leqslant 1\\ K+(K\gamma)^{\alpha} & \text{if } \alpha>2\end{cases}$$

Using this bound in the right-hand side of Equation (10), we obtain

$$\left(\frac{\sum_x (P_Y(x))^{\alpha}}{\sum_x (P_X(x))^{\alpha}}\right)^{\frac{1}{\alpha-1}}\geqslant\begin{cases}1+\frac{\alpha-1}{3}K^{\alpha-1}\gamma^{\alpha} & \text{if } \alpha\in(1,2)\wedge K\gamma>1\\ 1+\frac{\alpha(\alpha-1)}{4}K\gamma^{2} & \text{if } \alpha\in(1,2)\wedge K\gamma\leqslant 1\\ 1+K^{\alpha-1}\gamma^{\alpha} & \text{if } \alpha>2\end{cases}\tag{12}$$

It remains to choose the parameter γ, remembering about the assumptions on γ and α made in Equation (11). We may choose it the following ways:

Case 1: for $\Delta\in(0,1)$ and $\alpha>2$ we will choose: $\frac{1}{\alpha-1}\cdot K^{\alpha-1}\gamma^{\alpha}<1$. By taking the logarithm of Equation (12) and dividing by $\alpha-1$ we obtain

$$H_\alpha(Y)-H_\alpha(X)\geqslant\frac{1}{\alpha-1}\log\left(1+K^{\alpha-1}\gamma^{\alpha}\right).$$

Now the elementary inequality $\log(1+u)\geqslant u$ valid for $0\leqslant u\leqslant 1$ yields

$$H_\alpha(Y)-H_\alpha(X)\geqslant\frac{1}{\alpha-1}\cdot K^{\alpha-1}\gamma^{\alpha}.$$

Thus we achieve the entropy gap $\Delta=\frac{1}{\alpha-1}\cdot K^{\alpha-1}\gamma^{\alpha}$ and the distance $\gamma=((\alpha-1)\Delta)^{\frac{1}{\alpha}}K^{-1+\frac{1}{\alpha}}$ for any Δ between 0 and 1.

Case 2: for $\Delta\leqslant 1$ and $\alpha\in(1,2)$ we choose $\min\left(K\gamma^2,K^{\alpha-1}\gamma^{\alpha}\right)<1$. Using Equation (12), taking the logarithm of both sides and dividing by $\alpha-1$ we obtain

$$H_\alpha(Y)-H_\alpha(X)>\frac{1}{\alpha-1}\log\left(1+\frac{\alpha(\alpha-1)}{4}\cdot\min\left(K\gamma^2,K^{\alpha-1}\gamma^{\alpha}\right)\right).$$

Now the elementary inequality $\log(1+u)\geqslant u$ valid for $0\leqslant u\leqslant 1$ yields

$$H_\alpha(Y)-H_\alpha(X)\geqslant\frac{\alpha}{4}\cdot\min\left(K\gamma^2,K^{\alpha-1}\gamma^{\alpha}\right).$$

Hence we can have the entropy gap $\Delta=\frac{\alpha}{4}\cdot\min\left(K\gamma^2,K^{\alpha-1}\gamma^{\alpha}\right)$ and the distance $\gamma=\max\left(K^{-1+\frac{1}{\alpha}}\left(\frac{4\Delta}{\alpha}\right)^{\frac{1}{\alpha}},K^{-\frac{1}{2}}\left(\frac{4\Delta}{\alpha}\right)^{\frac{1}{2}}\right)$. The number Δ can be arbitrary between 0 and 1.

Case 3: for $\Delta>1$ and $\alpha\geqslant 2$ we choose $\frac{1}{\alpha-1}\cdot K^{\alpha-1}\gamma^{\alpha}>1$. Under this assumption, Equation (12) holds with the term $K^{\alpha-1}\gamma^{\alpha}$ on the right-hand side. By taking the logarithm in Equation (12) and dividing by $\alpha-1$ we obtain

$$H_\alpha(Y)-H_\alpha(X)>\frac{1}{\alpha-1}\cdot\log\left(1+K^{\alpha-1}\gamma^{\alpha}\right).$$

Now the inequality $\log(1+u) > \log u$ implies

$$H_\alpha(Y) - H_\alpha(X) > \frac{1}{\alpha-1}\log\left(K^{\alpha-1}\gamma^\alpha\right).$$

Thus, we can have the entropy gap $\Delta = \frac{1}{\alpha-1}\log\left(K^{\alpha-1}\gamma^\alpha\right)$ and the distance $\gamma = 2^{\Delta\left(1-\frac{1}{\alpha}\right)}K^{-1+\frac{1}{\alpha}}$, for any $1 \leqslant \Delta \leqslant \log K - O(1)$ (the upper bound follows by substituting $\gamma = \frac{K-1}{K}$ which is the maximal value).

Case 4: for $\Delta > 1$ and $\alpha \in (1,2)$ we choose $\min\left(K\gamma^2, K^{\alpha-1}\gamma^\alpha\right) > 1$. Recall, as for Case 2, that for $\alpha < 2$ we have $K^{\alpha-1}\gamma^\alpha > K\gamma^2$ when $K\gamma > 1$. Using this in Equation (12), taking the logarithm of both sides and dividing by $\alpha - 1$ we obtain

$$H_\alpha(Y) - H_\alpha(X) > \frac{1}{\alpha-1}\log\left(1 + \frac{\alpha(\alpha-1)}{4}\cdot\min\left(K\gamma^2, K^{\alpha-1}\gamma^\alpha\right)\right).$$

Now the inequality $\log(1+u) > \log u$ implies

$$H_\alpha(Y) - H_\alpha(X) > \frac{1}{\alpha-1}\log\left(\frac{\alpha(\alpha-1)}{4}\cdot\min\left(K\gamma^2, K^{\alpha-1}\gamma^\alpha\right)\right).$$

Thus, for the entropy gap $\Delta = \frac{1}{\alpha-1}\log\left(\frac{\alpha(\alpha-1)}{4}\cdot\min\left(K\gamma^2, K^{\alpha-1}\gamma^\alpha\right)\right)$ we get the distance $\gamma = \frac{4}{\alpha(\alpha-1)}\cdot\max\left(2^{\Delta\left(1-\frac{1}{\alpha}\right)}K^{-1+\frac{1}{\alpha}}, 2^{\frac{1}{2}\Delta}K^{-\frac{1}{2}}\right)$, for for any $1 \leqslant \Delta \leqslant \frac{1}{\alpha-1}\log K - O(1)$ (the upper bound follows by substituting $\gamma = \frac{K-1}{K}$ which is the maximal value). ◀

6 Conclusion

This paper offers stronger upper and lower bounds on the complexity of estimating Renyi entropy. Except quantitative improvements, it also provides simplifies the analysis, and provides more insight into the technique used to prove lower bounds.

Applying this technique to related problems, e.g. estimating different properties of discrete distributions besides entropy, is an interesting problem for future research.

We also emphasize that our construction for the lower bounds can be somewhat improved in two aspects: firstly, in Lemma 11 the choice of Y is optimal but X may be not - we assumed for simplicity that it is flat; secondly, there may be need for a more carefull bound on the variational distance between n-fold product distributions Lemma 8.

As for upper bounds, it remains an intriguing question if we can obtain improvements also for Shannon entropy estimation in low or medium entropy regimes.

References

1 Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1855–1869, 2015. doi:10.1137/1.9781611973730.124.

2 Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Information Theory*, 63(1):38–56, 2017. doi:10.1109/TIT.2016.2620435.

3 Erdal Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Trans. Information Theory*, 42(1):99–105, 1996. doi:10.1109/18.481781.

4 Boaz Barak, Yevgeniy Dodis, Hugo Krawczyk, Olivier Pereira, Krzysztof Pietrzak, François-Xavier Standaert, and Yu Yu. Leftover hash lemma, revisited. In *Advances in Cryptology – CRYPTO 2011 – 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings*, pages 1–20, 2011. doi:10.1007/978-3-642-22792-9_1.

5 Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating entropy. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 678–687, 2002. doi:10.1145/509907.510005.

6 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, 2013. doi:10.1145/2432622.2432626.

7 Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Ueli M. Maurer. Generalized privacy amplification. *IEEE Trans. Information Theory*, 41(6):1915–1923, 1995. doi:10.1109/18.476316.

8 Yevgeniy Dodis and Yu Yu. Overcoming weak expectations. In *Theory of Cryptography – 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, 2013. Proceedings*, pages 1–22, 2013. doi:10.1007/978-3-642-36594-2_1.

9 Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation – In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75. 2011. doi:10.1007/978-3-642-22670-0_9.

10 Manjesh Kumar Hanawal and Rajesh Sundaresan. Guessing revisited: A large deviations approach. *IEEE Trans. Information Theory*, 57(1):70–78, 2011. doi:10.1109/TIT.2010.2090221.

11 Russell Impagliazzo and David Zuckerman. How to recycle random bits. In *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October – 1 November 1989*, pages 248–253, 1989. doi:10.1109/SFCS.1989.63486.

12 R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft. Clustering using renyi's entropy. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 1, pages 523–528 vol.1, July 2003. doi:10.1109/IJCNN.2003.1223401.

13 Petr Jizba, Hagen Kleinert, and Mohammad Shefaat. Rényi's information transfer between financial time series. *Physica A: Statistical Mechanics and its Applications*, 391(10):2971–2989, 2012. doi:10.1016/j.physa.2011.12.064.

14 Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.

15 Ke Li, Wanlei Zhou, Shui Yu, and Bo Dai. Effective ddos attacks detection using generalized entropy metric. In *Algorithms and Architectures for Parallel Processing, 9th International Conference, ICA3PP 2009, Taipei, Taiwan, June 8-11, 2009. Proceedings*, pages 266–280, 2009. doi:10.1007/978-3-642-03095-6_27.

16 Bing Ma, Alfred O. Hero III, John D. Gorman, and Olivier J. J. Michel. Image registration with minimum spanning tree algorithm. In *Proceedings of the 2000 International Conference on Image Processing, ICIP 2000, Vancouver, BC, Canada, September 10-13, 2000*, pages 481–484, 2000. doi:10.1109/ICIP.2000.901000.

17 Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 367–374,

2009. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1600&proceeding_id=25.

18 Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities : Theory of Majorization and its Applications*. Springer Science+Business Media, LLC, New York, 2011.

19 Abolfazl S. Motahari, Guy Bresler, and David N. C. Tse. Information theory of DNA shotgun sequencing. *IEEE Trans. Information Theory*, 59(10):6273–6289, 2013. doi:10.1109/TIT.2013.2270273.

20 Huzefa Neemuchwala, Alfred O. Hero III, Sakina Zabuawala, and Paul L. Carson. Image registration methods in high-dimensional space. *Int. J. Imaging Systems and Technology*, 16(5):130–145, 2006. doi:10.1002/ima.20079.

21 Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, pages 1849–1857, USA, 2010. Curran Associates Inc. URL: http://dl.acm.org/citation.cfm?id=2997046.2997102.

22 Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, June 2003. doi:10.1162/089976603321780272.

23 Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Information Theory*, 54(10):4750–4755, 2008. doi:10.1109/TIT.2008.928987.

24 C. E. Pfister and W. G. Sullivan. Rényi entropy, guesswork moments, and large deviations. *IEEE Trans. Information Theory*, 50(11):2794–2800, 2004. doi:10.1109/TIT.2004.836665.

25 A. Renyi. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960. URL: http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf.

26 Prasanna K. Sahoo and Gurdial Arora. A thresholding method based on two-dimensional renyi's entropy. *Pattern Recognition*, 37(6):1149–1161, 2004. doi:10.1016/j.patcog.2003.10.008.

27 C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001. doi:10.1145/584091.584093.

28 Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 685–694, 2011. doi:10.1145/1993636.1993727.

29 Paul C. van Oorschot and Michael J. Wiener. Parallel collision search with cryptanalytic applications. *J. Cryptology*, 12(1):1–28, 1999. doi:10.1007/PL00003816.

30 Dongxin Xu. *Energy, Entropy and Information Potential for Neural Computation*. PhD thesis, University of Florida, Gainesville, FL, USA, 1998. AAI9935317.

31 Dongxin Xu and Deniz Erdogmuns. *Renyi's Entropy, Divergence and Their Nonparametric Estimators*, pages 47–102. Springer New York, New York, NY, 2010. doi:10.1007/978-1-4419-1570-2_2.

A Proof of Lemma 4

Proof. The proof of Equation (2) goes by induction. It is clearly valid for $\alpha = 1$. Assuming that it is valid for some $\alpha \geqslant 1$, we obtain

$$\begin{aligned}
n(x)^{\underline{\alpha+1}} &= n(x)^{\underline{\alpha}} \cdot (n(x) - \alpha) \\
&= \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x) \cdot \sum_{i_{\alpha+1}} (\xi_{i_{\alpha+1}}(x) - \alpha) \\
&= -\alpha \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x) + \\
&\quad + \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha \neq i_{\alpha+1}} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x) \\
&\quad + \sum_{\substack{i_1 \neq i_2 \neq \ldots \neq i_\alpha \\ i_{\alpha+1} \in \{i_1, \ldots, i_\alpha\}}} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x)\xi_{i_{\alpha+1}}(x).
\end{aligned}$$

Since ξ_i are boolean we have

$$\sum_{\substack{i_1 \neq i_2 \neq \ldots \neq i_\alpha \\ i_{\alpha+1} \in \{i_1, \ldots, i_\alpha\}}} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x)\xi_{i_{\alpha+1}}(x) = \alpha \cdot \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha} \xi_{i_1}(x)\xi_{i_2}(x) \cdot \ldots \cdot \xi_{i_\alpha}(x)$$

By putting together the last two equations we end the proof of Equation (2). To get Equation (3) we simply take the expectation and use independence. ◀

B Proof of Lemma 5

Proof. Note that

$$\begin{aligned}
\left(\sum_x n(x)^{\underline{\alpha}}\right)^2 &= \sum_{x,y} \sum_{\substack{i_1 \neq i_2 \neq \ldots \neq i_\alpha \\ j_1 \neq j_2 \neq \ldots \neq j_\alpha}} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(y) \\
&= \sum_{x \neq y} \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha \neq j_1 \neq j_2 \neq \ldots \neq j_\alpha} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(y) + \\
&\quad + \sum_x \sum_{\substack{i_1 \neq i_2 \neq \ldots \neq i_\alpha \\ j_1 \neq j_2 \neq \ldots \neq j_\alpha}} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(x).
\end{aligned}$$

Now we have

$$\begin{aligned}
I_1 &= \mathbb{E}\left[\sum_{x \neq y} \sum_{i_1 \neq i_2 \neq \ldots \neq i_\alpha \neq j_1 \neq j_2 \neq \ldots \neq j_\alpha} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(y)\right] \\
&= n^{\underline{2\alpha}} \sum_{x \neq y} p(x)^\alpha p(y)^\alpha \\
&= n^{\underline{2\alpha}} \left((p_\alpha)^2 - p_{2\alpha}\right).
\end{aligned}$$

Also

$$
\begin{aligned}
I_2 &= \mathbb{E}\left[\sum_x \sum_{\substack{i_1\neq i_2\neq\ldots\neq i_\alpha \\ j_1\neq j_2\neq\ldots\neq j_\alpha}} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(x)\right] \\
&= \mathbb{E}\left[\sum_{x\in\mathcal{X}}\sum_{\ell=0}^{\alpha} \sum_{\substack{i_1\neq i_2\neq\ldots\neq i_\alpha \\ j_1\neq j_2\neq\ldots\neq j_\alpha \\ |\{i_1\neq i_2\neq\ldots\neq i_\alpha\}\cap\{j_1\neq j_2\neq\ldots\neq j_\alpha\}|=\ell}} \prod_{r=1}^{\alpha} \xi_{i_r}(x)\xi_{j_r}(x)\right] \\
&= \sum_{x\in\mathcal{X}}\sum_{\ell=0}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\cdot p(x)^{2\alpha-\ell} \\
&= \sum_{\ell=0}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\cdot p_{2\alpha-\ell} \\
&= n^{\underline{2\alpha}}p_{2\alpha} + \sum_{\ell=1}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\cdot p_{2\alpha-\ell},
\end{aligned}
$$

where we observed that if the sets $\{i_1,\ldots,i_\alpha\}$ and $\{j_1,\ldots,j_\alpha\}$ have exactly ℓ common elements then $\mathbb{E}\prod_{r=1}^{\alpha}\xi_{i_r}(x)\xi_{j_r}(x) = p(x)^{2\alpha-\ell}$, and that there are $n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!$ choices for the such sets $\{i_1,\ldots,i_\alpha\}$ and $\{j_1,\ldots,j_\alpha\}$[5]. Putting this all together we obtain

$$
\begin{aligned}
\mathrm{Var}\left[\sum_x n(x)^{\underline{\alpha}}\right] &= n^{\underline{2\alpha}}(p_\alpha)^2 + \sum_{\ell=1}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\cdot p_{2\alpha-\ell} - (n^{\underline{\alpha}}p_\alpha)^2 \\
&= n^{\underline{\alpha}}((n-\alpha)^{\underline{\alpha}} - n^{\underline{\alpha}})(p_\alpha)^2 + \sum_{\ell=1}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!\cdot p_{2\alpha-\ell}
\end{aligned}
$$

which completes the proof. ◀

C Proof of Lemma 8

Proof. We will use the fact that if two distributions are ϵ-close (i.e. $d_{TV}(X',Y') < \epsilon$) then no distinguisher can distinguish between them with advantage greater then $\frac{\epsilon}{2}$. Let us assume that $|H_\alpha(X) - H_\alpha(Y)| \geqslant 2\delta$, then by using estimator $\hat{f}$ as part of the distinguisher i.e. if $|\hat{f}(.) - H_\alpha(X)| \leq \delta$ then distinguisher "guesses" that initial distribution was X^n, else "guesses" Y^n. Now we notice that initial distribution was X^n distinguisher will "guess" correctly with probability $1-\epsilon$, and if the initial distribution was Y^n then estimator with probability $1-\epsilon$ will output value in $[H_\alpha(Y)-\delta\ ;\ H_\alpha(Y)+\delta]$ thus distinguisher will guess correctly again. Our distinguisher achieves $1/2-\epsilon$ advantage thus we deduce that $d_{TV}(X^n;Y^n) > 1-2\epsilon$. ◀

[5] For a quick sanity check of this formula, note that when $p_i = 1$ (a constant random variable) then we should get $(n^{\underline{\alpha}})^2 = \sum_{\ell=0}^{\alpha} n^{\underline{\alpha}}(n-\alpha)^{\underline{\alpha-\ell}}\binom{\alpha}{\ell}^2 l!$. For $\alpha = 2$ this reduces to the identity $n(n-1) = (n-2)(n-3) + 4(n-2) + 2$.

D Maximizing entropy gap within variational distance constraints

▶ **Theorem 13.** *Let q be a fixed distribution over k elements, and $\alpha > 1$, $\epsilon \in (0,1)$ be fixed. Suppose that $q_1 \geqslant q_2 \geqslant \ldots \geqslant q_k$. Then the distribution p which is ϵ-close to q and has minimal possible α-entropy is given by*

$$q_i = \begin{cases} p_1 + \epsilon & i = 1 \\ p_i & 1 < i < i_0 \\ p_{i_0} - \sum_{j \geqslant i_0} p_j & i = i_0 \\ 0 & i > i_0 \end{cases} \tag{13}$$

where i_0 is the biggest number such that $\sum_{i \geqslant i_0} p_i \geqslant \epsilon$, for some x_0 such that $p(x_0)$ is the biggest mass, and for some $\epsilon' < \epsilon$.

Proof. We will apply majorization techniques [18]. Let q be optimal. Suppose that $q(x_1) > p(x_1)$ and $q(x_2) > p(x_2)$ where $x_1 \neq x_2$. Since q has the biggest possible power sum $S(q) = \sum_x q(x)^\alpha$ we see that $p(x_1)$ and $p(x_2)$ are two biggest probability masses. Assume, without loss of generality, that $q(x_1) \geqslant q(x_2)$. For some small $\delta > 0$ we perturb q into q' such that $q'(x_1) = q(x_1) + \delta$ and $q'(x_1) = q(x_1) - \delta$ and $q'(x) = q(x)$. Note that for small δ the distance between q' and p is at most as between p and q, and that q' majorizes q (considered as vectors) and the power sum $S(q)$ is Schur convex, hence $S(q) > S(q')$. The contradiction means that $q(x) > p(x)$ for only one $x = x_0$.

Consider now the smallest values $q(x_1), q(x_2)$ such that $0 < q(x_1) < p(x_1), 0 < q(x_2) < p(x_2)$ for $x_1 \neq x_2$ that are strictly bigger than zero. For some small $\delta > 0$ we perturb q into q' such that $q'(x_1) = q(x_1) + \delta$ and $q'(x_1) = q(x_1) - \delta$ and $q'(x) = q(x)$. We see that for δ small enough the distance from q' to p is at most as from q to p and that q' majorizes q which means $S(q') > S(q)$. The contradiction means that $0 < q(x) < p(x)$ for at most one $x = x_0$. ◀