Optimal Bounds for Estimating Entropy with PMF Queries

Cafer Caferov[1], Barış Kaya[1], Ryan O'Donnell[2⋆], and A. C. Cem Say[1]

[1] Boğaziçi University Computer Engineering Department
{cafer.caferov,baris.kaya,say}@boun.edu.tr
[2] Department of Computer Science, Carnegie Mellon University
odonnell@cs.cmu.edu

Abstract. Let $\boldsymbol{p}$ be an unknown probability distribution on $[n] \coloneqq \{1, 2, \ldots n\}$ that we can access via two kinds of queries: A SAMP query takes no input and returns $x \in [n]$ with probability $\boldsymbol{p}[x]$; a PMF query takes as input $x \in [n]$ and returns the value $\boldsymbol{p}[x]$. We consider the task of estimating the entropy of $\boldsymbol{p}$ to within $\pm\Delta$ (with high probability). For the usual Shannon entropy $H(\boldsymbol{p})$, we show that $\Omega(\log^2 n/\Delta^2)$ queries are necessary, matching a recent upper bound of Canonne and Rubinfeld. For the Rényi entropy $H_\alpha(\boldsymbol{p})$, where $\alpha > 1$, we show that $\Theta(n^{1-1/\alpha}/2^\Delta)$ queries are necessary and sufficient. This complements recent work of Acharya et al. in the SAMP-only model that showed $O(n^{1-1/\alpha})$ queries suffice when α is an integer, but $\widetilde{\Omega}(n)$ queries are necessary when α is a noninteger. All of our lower bounds also easily extend to the model where CDF queries (given x, return $\sum_{y \le x} \boldsymbol{p}[y]$) are allowed.

1 Introduction

The field of statistics is to a large extent concerned with questions of the following sort: How many samples from an unknown probability distribution $\boldsymbol{p}$ are needed in order to accurately estimate various properties of the distribution? These sorts of questions have also been studied more recently within the theoretical computer science framework of *property testing*. In this framework, one typically makes no assumptions about $\boldsymbol{p}$ other than that it is a discrete distribution supported on, say, $[n] \coloneqq \{1, 2, \ldots, n\}$. There is a vast literature on testing and estimating properties of unknown distributions; for a survey with pointers to the literature, see Rubinfeld [Rub12] and Canonne [Can15].

One of the most important properties of a probability distribution $\boldsymbol{p}$ is its *Shannon entropy*, $H(\boldsymbol{p}) = \sum_x \boldsymbol{p}[x] \log \frac{1}{\boldsymbol{p}[x]}$.[3] Shannon entropy is a measure of the "amount of randomness" in $\boldsymbol{p}$. In this work we will be concerned with the associated task of estimating the entropy of an unknown $\boldsymbol{p}$ within a confidence

[⋆] Work performed while the author was at the Boğaziçi University Computer Engineering Department, supported by Marie Curie International Incoming Fellowship project number 626373.

[3] In this paper, log denotes $\log_2$.

interval of $\pm\Delta$ with probability at least $1-\delta$. (Typically $\Delta = 1$ and $\delta = 1/3$.) We also remark that if $\boldsymbol{p}$ is a distribution on $[n] \times [n]$ representing the joint pmf of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $H(\boldsymbol{p})$ is related to the *mutual information* $I(\boldsymbol{X};\boldsymbol{Y})$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ via $H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y})$. Thus additively estimating mutual information easily reduces to additively estimating entropy. For an extended survey and results on the fundamental task of estimating entropy, see Paninski [Pan03]; this survey includes justification of discretization, as well as discussion of applications to neuroscience (e.g., estimating the information capacity of a synapse).

It is known that in the basic "samples-only model" — in which the only access to $\boldsymbol{p}$ is via independent samples — estimation of entropy is a very expensive task. From the works [Pan03, Pan04, Val11, VV11b, VV11a] we know that estimating $H(\boldsymbol{p})$ to within ± 1 with confidence 2/3 requires roughly a linear number of samples; more precisely, $\Theta(n/\log n)$ samples are necessary (and sufficient). Unfortunately, for many applications this quantity is too large. E.g., for practical biological experiments it may be infeasible to obtain that many samples; or, for the enormous data sets now available in computer science applications, it may simply take too long to process $\widetilde{\Theta}(n)$ samples.

To combat this difficulty, researchers have considered an extension to the samples-only model, called the "Generation+Evaluation" model in [KMR$^+$94] and the "combined model" in [BDKR05]. We will refer to it as the SAMP+PMF model because it allows the estimation algorithm two kinds of "queries" to the unknown distribution $\boldsymbol{p}$: a SAMP query, which takes no input and returns $x \in [n]$ with probability $\boldsymbol{p}[x]$; and a PMF query, which takes as input $x \in [n]$ and returns the value $\boldsymbol{p}[x]$. As we will see, in this model entropy can be accurately estimated with just $\mathrm{polylog}(n)$ queries, dramatically smaller than the $\Omega(n/\log n)$ queries needed in the samples-only model.

Regarding the relevance of the SAMP+PMF model, an example scenario in which it might occur is the Google n-gram database; the frequency of each n-gram is published, and it is easy to obtain a random n-gram from the underlying text corpus. Another motivation for SAMP+PMF access comes from the *streaming* model of computation [AMS99], where entropy estimation has been well studied [GMV06, LSO$^+$06, CDM06, BG06, CCM07, HNO08]. The SAMP+PMF testing model and the streaming model are closely related. Roughly speaking, any q-query estimation algorithm in the SAMP+PMF model can be converted to a $q \cdot \mathrm{polylog}(n)$-space streaming algorithm with one or two passes (with precise details depending on the model for how the items in the stream are ordered). More motivation and results for the SAMP+PMF model can be found in Canonne and Rubinfeld [CR14].

1.1 Our Results, and Comparison with Prior Work

The first works [BDKR05, GMV06] on entropy estimation in the SAMP+PMF model were concerned with *multiplicative* estimates of $H(\boldsymbol{p})$. Together they show relatively tight bounds for this problem: estimating (with high probability) $H(\boldsymbol{p})$

to within a multiplicative factor of $1+\gamma$ requires

$$\text{between} \quad \Omega\left(\frac{\log n}{\max(\gamma,\gamma^2)}\right)\cdot\frac{1}{H(\boldsymbol{p})} \quad \text{and} \quad O\left(\frac{\log n}{\gamma^2}\right)\cdot\frac{1}{H(\boldsymbol{p})} \tag{1}$$

queries. Unfortunately, these bounds depend quantitatively on the entropy $H(\boldsymbol{p})$ itself; the number of queries necessarily scales as $1/H(\boldsymbol{p})$. Further, whereas additive estimation of entropy can be used to obtain additive estimates of mutual information, multiplicative estimates are insufficient for this purpose. Thus in this paper we consider only the problem of additive approximation.

For this problem, Canonne and Rubinfeld [CR14] recently observed that $O(\log^2 n)$ SAMP+PMF queries are sufficient to estimate $H(\boldsymbol{p})$ to ±1 with high probability, and $\Omega(\log n)$ queries are necessary. The first main result in this work is an improved, optimal lower bound:

First main theorem. *In the SAMP+PMF model, $\Omega(\log^2 n)$ queries are necessary to estimate (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within ± 1.*

Remark 1. Our lower bound and the lower bound from (1) hold even under the promise that $H(\boldsymbol{p}) = \Theta(\log n)$. The lower bound in (1) yields a lower bound for our additive approximation problem by taking $\gamma = \frac{1}{\log n}$, but only a nonoptimal one: $\Omega(\log n)$.

More generally, Canonne and Rubinfeld showed that $O(\frac{\log^2 n}{\Delta^2})$ queries suffice for estimating Shannon entropy to within $\pm\Delta$.[4] Note that this result is trivial once $\Delta \le \frac{\log n}{\sqrt{n}}$ because of course the entire distribution $\boldsymbol{p}$ can be determined precisely with n PMF queries. In fact, our first main theorem is stated to give a matching lower bound for essentially the full range of Δ: for any $\frac{1}{n^{.4999}} \le \Delta \le \log n$ we show that $\Omega(\frac{\log^2 n}{\Delta^2})$ queries are necessary in the SAMP+PMF model.

Our second main theorem is concerned with estimation of the *Rényi entropy* $H_\alpha(\boldsymbol{p})$ for various parameters $0 \le \alpha \le \infty$. Here

$$H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha}\log\left(\sum_x \boldsymbol{p}[x]^\alpha\right),$$

interpreted in the limit when $\alpha = 0, 1, \infty$. The meaning for $\boldsymbol{p}$ is as follows: when $\alpha = 0$ it's the (log of the) support size; when $\alpha = 1$ it's the usual Shannon entropy; when $\alpha = \infty$ it's the min-entropy; when $\alpha = 2$ it's the (negative-log of the) collision probability; and for general integer $\alpha \ge 2$ it's related to higher-order collision probabilities.

A recent work of Acharya, Orlitsky, Suresh, and Tyagi [AOST15] showed that for estimating $H_\alpha(\boldsymbol{p})$ in the samples-only model to within ±1, $\Theta(n^{1-1/\alpha})$

[4] They actually state $O(\frac{\log^2(n/\Delta)}{\Delta^2})$, but this is the same as $O(\frac{\log^2 n}{\Delta^2})$ because the range of interest is $\frac{1}{\sqrt{n}} \le \Delta \le \log n$.

samples are necessary and sufficient when α is an integer greater than 1, and $\widetilde{\Omega}(n)$ queries are necessary when α is a noninteger greater than 1. Our second main result is a tight characterization of the number of SAMP+PMF queries necessary and sufficient for estimating $H_\alpha(\boldsymbol{p})$ for all $\alpha > 1$. It turns out that PMF queries do not help in estimating these more general Rényi entropies for integer α, whereas they *are* helpful for noninteger α.

Second main theorem. *Let $\alpha > 1$ be a real number. In the SAMP+PMF model, $\Theta(n^{1-1/\alpha}/2^{\Delta})$ queries are necessary and sufficient to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Finally, we mention that our two lower bounds easily go through even when the more liberal "CDF" queries introduced in [CR14] are allowed. These queries take as input $x \in [n]$ and return the value $\sum_{y \leq x} \boldsymbol{p}[y]$.[5] We will also show that the Canonne–Rubinfeld SAMP+PMF lower bound of $\Omega(1/\epsilon^2)$ for estimating support size to within $\pm\epsilon n$ can be extended to the more general SAMP+CDF model.

2 Lower Bound for Estimating Shannon Entropy

Theorem 2. *In the SAMP+PMF model, $\Omega\left(\frac{\log^2 n}{\Delta^2}\right)$ queries are necessary to estimate (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$, where $\frac{1}{n^{.4999}} \leq \Delta \leq \log n$.*

Proof. We will show that a hypothetical SAMP+PMF algorithm $\mathcal{E}$ that can estimate the entropy of an unknown distribution on $[n]$ to within $\pm\Delta$ using $o\left(\frac{\log^2 n}{\Delta^2}\right)$ queries would contradict the well-known fact that $\Omega(1/\lambda^2)$ coin tosses are necessary to determine whether a given coin is fair, or comes up heads with probability $1/2 + \lambda$.

The idea is to use the given coin to realize the probability distribution that $\mathcal{E}$ will work on. Let n be the smallest one millionth power of a natural number that satisfies $\frac{3 \cdot 10^6 \Delta}{\log n} \leq \lambda$. Partition the domain $[n]$ into $M = n^{.999999}$ consecutive blocks $I_1, \ldots, I_M$, each containing $K = \frac{n}{M} = n^{.000001}$ elements. Each block will be labeled either as a tails or a heads block. The internal distribution of each heads block is uniform, i.e. each element has probability mass $\frac{1}{MK} = \frac{1}{n}$. In each tails block, the first element has probability mass $\frac{1}{n^{.999999}}$, while the rest of the elements have probability mass 0. Note that the total probability mass of each block is $K \cdot \frac{1}{MK} = \frac{1}{M} = \frac{1}{n^{.999999}}$, regardless of its label.

We will now describe a costly method of constructing a probability distribution $\boldsymbol{p}$ of this kind, using a coin that comes up heads with probability d:
- Throw the coin M times to obtain the outcomes $\boldsymbol{X_1}, \ldots, \boldsymbol{X_M}$,
- Set the label of block I_m to $\boldsymbol{X_m}$, for all $m \in [M]$.

[5] Note that a PMF query can be simulated by two CDF queries.

Let $\boldsymbol{X}$ be the number of heads blocks in $\boldsymbol{p}$. Then $\mu = \mathbf{E}[\boldsymbol{X}] = Md$. Let $\overline{\boldsymbol{X}} = \frac{\boldsymbol{X}}{M}$ denote the proportion of heads blocks in $\boldsymbol{p}$. Then we can calculate the entropy $H[\boldsymbol{p}]$ by calculating the individual entropies of the blocks. For a heads block, the entropy is $K \cdot \frac{1}{MK} \cdot \log(MK) = \frac{1}{M}\log n$. The entropy of a tails block is $\frac{1}{n^{.999999}}\log(n^{.999999}) = \frac{.999999}{M}\log n$. Since there are $M\overline{\boldsymbol{X}}$ heads blocks and $M(1-\overline{\boldsymbol{X}})$ tails blocks, the total entropy becomes $H[\boldsymbol{p}] = M\overline{\boldsymbol{X}} \cdot \frac{1}{M}\log n + M(1-\overline{\boldsymbol{X}}) \cdot \frac{.999999}{M}\log n = \overline{\boldsymbol{X}}\log n + .999999(1-\overline{\boldsymbol{X}})\log n = (.999999 + .000001\overline{\boldsymbol{X}})\log n$. Note that this function is monotone with respect to $\overline{\boldsymbol{X}}$.

Define two families of distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ constructed by the above process, taking d to be $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{2} + \lambda$, respectively. Let $\boldsymbol{p_1}$ (respectively $\boldsymbol{p_2}$) be a probability distribution randomly chosen from $\mathcal{P}_1$ (respectively $\mathcal{P}_2$).

Proposition 3. $\boldsymbol{p_1}$ *has entropy at most* $.9999995\log n + \varDelta$ *with high probability.*

Proof. We prove this by using a Chernoff bound on the number of heads blocks in the distribution.

$$\begin{aligned}\mathbf{Pr}\left[\boldsymbol{X} \geq \left(p_1 + \tfrac{10^6\varDelta}{\log n}\right)M\right] &\leq \exp\Bigg(-\frac{\frac{4\cdot 10^{12}\varDelta^2}{\log^2 n}}{2 + \frac{2\cdot 10^6\varDelta}{\log n}}\frac{M}{2}\Bigg)\\ &\leq \exp\Bigg(-\frac{10^{12}\cdot n^{.999999}/n^{.999998}}{\log^2 n(1+10^6)}\Bigg) = o(1).\end{aligned}$$

The last term indicates that the proportion of the heads blocks $\overline{\boldsymbol{X}} < \big(p_1 + \frac{10^6\varDelta}{\log n}\big)$ with high probability. Thus with high probability $H[\boldsymbol{p_1}] = (.999999 + .000001\overline{\boldsymbol{X}})\log n < .9999995\log n + \varDelta$. □

Proposition 4. $\boldsymbol{p_2}$ *has entropy at least* $.9999995\log n + 2\varDelta$ *with high probability.*

Proof. We find a similar bound by;

$$\mathbf{Pr}\left[\boldsymbol{X} \leq \left(p_2 - \tfrac{10^6\varDelta}{\log n}\right)M\right] \leq \exp\Bigg(-\frac{\frac{10^{12}\varDelta^2}{p_2^2\log^2 n}}{2}p_2 M\Bigg) \leq \exp\left(-\frac{n^{.000001}}{\log^2 n}\right) = o(1).$$

The last term indicates that the proportion of the heads blocks $\overline{\boldsymbol{X}} > \big(p_2 - \frac{10^6\varDelta}{\log n}\big)$ with high probability. Thus with high probability $H[\boldsymbol{p_2}] = (.999999 + .000001\overline{\boldsymbol{X}})\log n > .9999995\log n + .000001(\lambda - \frac{10^6\varDelta}{\log n})\log n \geq .9999995\log n + .000001(\frac{2\cdot 10^6\varDelta}{\log n})\log n = .9999995\log n + 2\varDelta$. □

Since the entropies of $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ are sufficiently far apart from each other, our hypothetical estimator $\mathcal{E}$ can be used to determine whether the underlying coin has probability p_1 or p_2 associated with it. To arrive at the contradiction we want, we must ensure that the coin is not thrown too many times during this process. This is achieved by constructing the distribution "on-the-fly" [CR14] during the execution of $\mathcal{E}$, throwing the coin only when it is required to determine the label of a previously undefined block:

When $\mathcal{E}$ makes a SAMP query, we choose a block I_m uniformly at random (since each block has probability mass $\frac{1}{M}$), and then flip the coin for I_m to decide its label if it is yet undetermined. We then draw a sample $i \sim \mathbf{d_m}$ from I_m, where $\mathbf{d_m}$ is the normalized distribution of the m^{th} block.

When $\mathcal{E}$ makes a PMF query on $i \in [n]$, we flip the coin to determine the label of the associated block I_m if it is yet undetermined. We then return the probability mass of i.

By this procedure, the queries of $\mathcal{E}$ about the probability distribution $\boldsymbol{p}$ (known to be either $\boldsymbol{p_1}$ or $\boldsymbol{p_2}$) can be answered by using at most one coin flip per query, i.e. $o\left(\frac{\log^2 n}{\Delta^2}\right)$ times in total.

Since we selected n so that $1/\lambda^2 = \Theta(\frac{\log^2 n}{\Delta^2})$, this would mean that it is possible to distinguish between the two coins using only $o(1/\lambda^2)$ throws, which is a contradiction, letting us conclude that no algorithm can estimate the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with high probability making $o\left(\frac{\log^2 n}{\Delta^2}\right)$ queries. □

We now give a similar lower bound for the SAMP+CDF model.

Corollary 5. *In the SAMP+CDF model, any algorithm estimating (with high probability) the Shannon entropy $H(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ must make $\Omega\left(\frac{\log^2 n}{\Delta^2}\right)$ queries.*

Proof. The construction is identical to the one in the proof of Theorem 2, except that we now have to describe how the CDF queries of the estimation algorithm must be answered using the coin:

When $\mathcal{E}$ makes a CDF query on $i \in [n]$, we flip the coin to determine the label of the associated block I_m if this is necessary. We then return the sum of the total probability mass of the blocks preceding I_m (which is $\frac{m-1}{M}$, since each block has a total probability mass of $\frac{1}{M}$ regardless of its label) and the probability masses of the elements from the beginning of I_m up to and including i itself. At most one coin flip per CDF query is therefore sufficient. □

3 Estimating Rényi Entropy

We start by demonstrating a lower bound.

Theorem 6. *For any $\alpha > 1$, $\Omega\left(\frac{n^{1-1/\alpha}}{2^{\Delta}}\right)$ SAMP+PMF queries are necessary to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Proof. We will first prove the theorem for rational α, and show that it remains valid for irrationals at the end.

The proof has the same structure as that of Theorem 2. One difference is that we reduce from the problem of distinguishing a maximally biased coin that never comes up tails from a less biased one (instead of the problem of distinguishing a fair coin from a biased one).

Suppose that we are given a coin whose probability of coming up heads is promised to be either $p_1 = 1$ or $p_2 = 1 - \lambda$ for a specified number λ, and we must determine which is the case. It is easy to show that this task requires at least $\varOmega(1/\lambda)$ coin throws. We will show that this fact is contradicted if one assumes that there exist natural numbers s and t, where $\alpha = \dfrac{s}{t} > 1$, such that it is possible to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\varDelta$ using an algorithm, say $\mathcal{R}$, that makes only $o(\dfrac{n^{1-1/\alpha}}{2^{\varDelta}})$ SAMP+PMF queries.

Let n be the smallest number of the form $\left(\lceil 2^{\varDelta}\rceil j\right)^s$ that satisfies $\frac{5\cdot\lceil 2^{\varDelta}\rceil}{n^{1-1/\alpha}} \le \lambda$, where j is some natural number. Partition $[n]$ into $M = \frac{n^{1-1/\alpha}}{\lceil 2^{\varDelta}\rceil}$ consecutive blocks $I_1, I_2, \ldots I_M$, each of size $K = \lceil 2^{\varDelta}\rceil \cdot n^{1/\alpha}$. As in the proof of Theorem 2, a probability distribution $\boldsymbol{p}$ can be realized by throwing a given coin M times to obtain the outcomes $\boldsymbol{X_1}, \ldots, \boldsymbol{X_M}$, and setting the label of block I_m to $\boldsymbol{X_m}$, for all $m \in [M]$, where each member of each heads block again has probability mass $1/n$. The first member of each tails block has probability mass $\frac{\lceil 2^{\varDelta}\rceil}{n^{1-1/\alpha}}$, and the remaining members have probability mass 0. We again have that each block has total probability mass $\frac{K}{n} = \frac{\lceil 2^{\varDelta}\rceil n^{1/\alpha}}{n} = \frac{1}{M}$ regardless of its label, so this process always results in a legal probability distribution.

If the coin is maximally biased, then $\boldsymbol{p}$ becomes the uniform distribution, and $H_\alpha(\boldsymbol{p}) = \log n$. We will examine the probability of the same distribution being obtained using the less biased coin. Let $\mathcal{P}_2$ be the family of distributions constructed by the process described above, using a coin with probability p_2 of coming up heads. Let $\boldsymbol{p_2}$ be a probability distribution randomly chosen from $\mathcal{P}_2$.

The probability of the undesired case where $\boldsymbol{p_2}$ is the uniform distribution is

$$\mathbf{Pr}\left[\boldsymbol{p_2} = \mathcal{U}([n])\right] = p_2^M \le \left(1 - \frac{5\cdot\lceil 2^{\varDelta}\rceil}{n^{1-1/\alpha}}\right)^M \le e^{-\frac{5\cdot\lceil 2^{\varDelta}\rceil}{n^{1-1/\alpha}}M} = e^{-5} \le \frac{1}{1000} \quad .$$

That is, with probability $\ge .999$, $\boldsymbol{p_2}$ has at least one element with probability mass $\frac{\lceil 2^{\varDelta}\rceil n^{\frac{1}{\alpha}}}{n}$. Let $\boldsymbol{X}$ be the number of heads outcomes and $\boldsymbol{B}$ and $\boldsymbol{W}$ denote the number of elements with probability mass $\frac{1}{n}$ and $\frac{\lceil 2^{\varDelta}\rceil n^{\frac{1}{\alpha}}}{n}$, respectively. It is not difficult to see that $\boldsymbol{B} = K \cdot \boldsymbol{X}$ and $\boldsymbol{W} = M - \boldsymbol{X}$. We just showed that $\boldsymbol{X} < M$ with high probability.

Then the Rényi entropy of the constructed distribution $\boldsymbol{p_2} \in \mathcal{P}_2$ is, with high probability:

$$\begin{aligned} H_\alpha\left(\boldsymbol{p_2}\right) &= \frac{1}{1-\alpha}\log\left(\frac{K\cdot\boldsymbol{X}/n + (M-\boldsymbol{X})\lceil 2^{\varDelta}\rceil^\alpha}{n^{\alpha-1}}\right) \\ &\le \log n - \frac{1}{\alpha-1}\log\left(\lceil 2^{\varDelta}\rceil^\alpha\right) \le \log n - \varDelta. \end{aligned}$$

Because $H_\alpha(\mathcal{U}([n])) - H_\alpha(\boldsymbol{p_2}) \geq \mathit{\Delta}$, $\mathcal{R}$ has to be able to distinguish $\mathcal{U}([n])$ and $\boldsymbol{p_2}$ with high probability.

We can then perform a simulation of $\mathcal{R}$ involving an "on-the-fly" construction of distribution $\boldsymbol{p}$ exactly as described in the proof of Theorem 2. As discussed in Section 2, this process requires no more coin throws than the number of SAMP+PMF queries made by $\mathcal{R}$, allowing us to determine the type of the coin using only $o(\frac{n^{1-1/\alpha}}{2^{\mathit{\Delta}}})$, that is, $o(1/\lambda)$ tosses with high probability, a contradiction.

Having thus proven the statement for rational α, it is straightforward to cover the case of irrational α: Note that $H_\alpha(\boldsymbol{p})$ is a continuous function of α for fixed $\boldsymbol{p}$. Given any $\boldsymbol{p}$ and ε, for any irrational number α_i greater than 1, there exists a rational α_r which is so close to α_i such that $H_{\alpha_i}(\boldsymbol{p}) - H_{\alpha_r}(\boldsymbol{p}) < \varepsilon$. An efficient entropy estimation method for some irrational value of α would therefore imply the existence of an equally efficient method for some rational value, contradicting the result obtained above. □

These results are generalized to the SAMP+CDF model in the same way as in Section 2:

Corollary 7. *For any $\alpha > 1$, $\mathit{\Omega}\left(\frac{n^{1-1/\alpha}}{2^{\mathit{\Delta}}}\right)$ SAMP+PMF or SAMP+CDF queries are necessary to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\mathit{\Delta}$.*

We now show that PMF queries are useful for the estimation of H_α for non-integer α.

Lemma 8. *For any rational number $\alpha > 1$, there exists an algorithm estimating (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\mathit{\Delta}$ with $O\left(\frac{n^{1-1/\alpha}}{2^{\mathit{\Delta}}}\right)$ SAMP+PMF queries.*

Proof. Defining the α^{th} moment of $\boldsymbol{p}$ as

$$\mathcal{M}_\alpha(\boldsymbol{p}) = \sum_{i=1}^{n} (\boldsymbol{p}[i])^\alpha,$$

the Rényi entropy can be written as

$$H_\alpha(\boldsymbol{p}) = \tfrac{1}{1-\alpha} \log \mathcal{M}_\alpha(\boldsymbol{p}).$$

Observe that estimating $H_\alpha(\boldsymbol{p})$ to an additive accuracy of $\pm\mathit{\Delta}$ is equivalent to estimating $\mathcal{M}_\alpha(\boldsymbol{p})$ to a multiplicative accuracy of $2^{\pm\mathit{\Delta}(1-\alpha)}$. Therefore we construct an estimator for $\mathcal{M}_\alpha(\boldsymbol{p})$.

Let $M = \left\lceil \frac{100 n^{1-1/\alpha}}{2^{\mathit{\Delta}}} \right\rceil$, and let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_M}$ be i.i.d. random variables drawn from $\boldsymbol{p}$. Define $\boldsymbol{Y_i} = (\boldsymbol{p}[\boldsymbol{X_i}])^{\alpha-1}$, where $\boldsymbol{p}[\boldsymbol{X_i}]$ can be calculated using a PMF query on $\boldsymbol{X_i}$ for $1 \leq i \leq M$. Note that $\mathbf{E}[\boldsymbol{Y_i}] = \sum_{j=1}^{n} \boldsymbol{p}[j] (\boldsymbol{p}[j])^{\alpha-1} =$

$\sum_{j=1}^{n} (\boldsymbol{p}[j])^{\alpha} = \mathcal{M}_\alpha(\boldsymbol{p})$. Then $\frac{1}{M}\sum_{i=1}^{M} \boldsymbol{Y_i}$ is an unbiased estimator of $\mathcal{M}_\alpha(\boldsymbol{p})$, because $\mathbf{E}\left[\frac{1}{M}\sum_{i=1}^{M} \boldsymbol{Y_i}\right] = \frac{1}{M}\sum_{i=1}^{M}\mathbf{E}[\boldsymbol{Y_i}] = \mathcal{M}_\alpha(\boldsymbol{p})$. Moreover, $\mathbf{Var}(\boldsymbol{Y_i}) = \mathbf{E}\left[\boldsymbol{Y_i}^2\right] - \mathbf{E}^2[\boldsymbol{Y_i}] = \sum_{j=1}^{n} \boldsymbol{p}[j](\boldsymbol{p}[j])^{2\alpha-2} - \mathbf{E}^2[\boldsymbol{Y_i}] = \mathcal{M}_{2\alpha-1}(\boldsymbol{p}) - \mathcal{M}_\alpha^2(\boldsymbol{p})$. Since the $\boldsymbol{Y_i}$'s are also i.i.d. random variables, $\mathbf{Var}\left(\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i}\right) = \frac{1}{M^2}\sum_{i=1}^{M}\mathbf{Var}(\boldsymbol{Y_i}) = \frac{M}{M^2}\mathbf{Var}(\boldsymbol{Y}) = \frac{1}{M}\left(\mathcal{M}_{2\alpha-1}(\boldsymbol{p}) - \mathcal{M}_\alpha^2(\boldsymbol{p})\right)$.

We use the following fact from [AOST15] to find an upper bound for the variance of our empirical estimator.

Fact 9 *([AOST15], Lemma 1) For $\alpha > 1$ and $0 \le \beta \le \alpha$*

$$\mathcal{M}_{\alpha+\beta}(\boldsymbol{p}) \le n^{(\alpha-1)(\alpha-\beta)/\alpha}\mathcal{M}_\alpha^2(\boldsymbol{p}) \quad .$$

By taking $\beta = \alpha - 1$, we get

$$\sigma^2 = \mathbf{Var}\left(\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i}\right) \le \frac{1}{M}\mathcal{M}_\alpha^2(\boldsymbol{p})\left(n^{1-1/\alpha} - 1\right) \le \mathcal{M}_\alpha^2(\boldsymbol{p})\frac{2^{\Delta}}{100}.$$

By Chebyshev's inequality we have

$$\mathbf{Pr}\left[\left|\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i} - M_\alpha(\boldsymbol{p})\right| \ge 10\sigma\right] \le \frac{1}{100} \Rightarrow$$
$$\mathbf{Pr}\left[\left|\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i} - M_\alpha(\boldsymbol{p})\right| \le \mathcal{M}_\alpha(\boldsymbol{p})\,2^{\Delta}\right] \ge .99$$

Thus we can estimate $\mathcal{M}_\alpha(\boldsymbol{p})$ to a desired multiplicative accuracy with $O\left(\frac{n^{1-1/\alpha}}{2^{\Delta}}\right)$ queries, which ends the proof. □

Applying the generalization to irrational α discussed in the proof of Theorem 6 to Lemma 8, the results proven in this section up to now can be summarized in the following theorem.

Theorem 10. *Let $\alpha > 1$ be a real number. In both the SAMP+PMF and the SAMP+CDF models, $\Theta(n^{1-1/\alpha}/2^{\Delta})$ queries are necessary and sufficient to estimate (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.*

Finally, we show a similar upper bound for $\alpha < 1$.

Lemma 11. *Let $\alpha < 1$ and $1 > \epsilon > 0$ be rational numbers. There exists an algorithm estimating (with high probability) the Rényi entropy $H_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$ with $O\left(\frac{n^{\epsilon}}{2^{\Delta}}\right)$ SAMP+PMF queries.*

Proof. See Appendix A. □

4 Lower Bound for Estimating Support Size

For any probability distribution $\boldsymbol{p}$, $H_0(\boldsymbol{p}) = \log(\operatorname{supp}(\boldsymbol{p}))$, where $\operatorname{supp}(\boldsymbol{p})$ denotes the support size of $\boldsymbol{p}$. Canonne and Rubinfeld [CR14] have shown that $\Omega(1/\epsilon^2)$ SAMP+PMF queries are necessary for estimating $\operatorname{supp}(\boldsymbol{p})$ to within $\pm\epsilon n$ where $[n]$ is the domain of $\boldsymbol{p}$. We modify their proof to establish the same lower bound for this task in the SAMP+CDF model.

Theorem 12. *$\Omega\left(\frac{1}{\epsilon^2}\right)$ SAMP+CDF queries are necessary to estimate (with high probability) the support size of an unknown distribution $\boldsymbol{p}$ on domain $[n]$ to within $\pm\epsilon n$.*

Proof. Assume that there exists a program $\mathcal{S}$ which can accomplish the task specified in the theorem statement with only $o\left(\frac{1}{\epsilon^2}\right)$ queries. Let us show how $\mathcal{S}$ can be used to determine whether a given a coin is fair, or comes up heads is with probability $p_2 = \frac{1}{2} + \lambda$.

Set $\epsilon = \frac{\lambda}{6}$, and let n be the smallest even number satisfying $n \geq 10/\epsilon^2$. Partition the domain $[n]$ into $M = \frac{n}{2}$ blocks $I_1, \ldots, I_M$ where $I_m = \{2m-1, 2m\}$ for all $m \in [M]$. The construction of a probability distribution $\boldsymbol{p}$ based on coin flips is as follows:
- Throw the coin M times, with outcomes $\boldsymbol{X_1}, \ldots, \boldsymbol{X_M}$,
- for $m \in [M]$, set $\boldsymbol{p}[2m-1] = \frac{2}{n}$ and $\boldsymbol{p}[2m] = 0$ if $\boldsymbol{X_m}$ is heads, and set $\boldsymbol{p}[2m-1] = \boldsymbol{p}[2m] = \frac{1}{n}$ if $\boldsymbol{X_m}$ is tails.

Note that by construction $\boldsymbol{p}[2m-1] + \boldsymbol{p}[2m] = \frac{2}{n}$ for all $m \in [M]$.

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the families of distributions constructed by the above process, using the fair and biased coin, respectively. Let $\boldsymbol{p_1}$ (respectively $\boldsymbol{p_2}$) be a probability distribution randomly chosen from $\mathcal{P}_1$ (respectively $\mathcal{P}_2$). Then

$$\mathbf{E}[\operatorname{supp}(\boldsymbol{p_1})] = n - M\frac{1}{2} = n\left(1 - \frac{1/2}{2}\right) = \frac{3}{4}n$$

$$\mathbf{E}[\operatorname{supp}(\boldsymbol{p_2})] = n - Mp_2 = n\left(1 - \frac{p_2}{2}\right) = n\left(\frac{3}{4} - \frac{\lambda}{2}\right) = \left(\frac{3}{4} - 3\epsilon\right)n$$

and via additive Chernoff bound,

$$\mathbf{Pr}\left[\operatorname{supp}(\boldsymbol{p_1}) \leq \frac{3}{4}n - \frac{\epsilon}{2}n\right] \leq e^{-\frac{\epsilon^2 n}{2}} \leq e^{-5} < \frac{1}{1000}$$

$$\mathbf{Pr}\left[\operatorname{supp}(\boldsymbol{p_2}) \geq \frac{3}{4}n - \frac{5\epsilon}{2}n\right] \leq e^{-\frac{\epsilon^2 n}{2}} \leq e^{-5} < \frac{1}{1000}$$

In other words, with high probability the resulting distributions will satisfy $\operatorname{supp}(\boldsymbol{p_1}) - \operatorname{supp}(\boldsymbol{p_2}) > 2\epsilon n$, distant enough for $\mathcal{S}$ to distinguish between two families.

As in our previous proofs, we could use $\mathcal{S}$ (if only it existed) to distinguish between the two possible coin types by using the coin for an on-the-fly construction of $\boldsymbol{p}$. As before, SAMP and CDF queries are answered by picking a block randomly, throwing the coin if the type of this block has not been fixed before, and returning the answer depending on the type of the block. Since $o\left(\frac{1}{\epsilon^2}\right)=o\left(\frac{1}{\lambda^2}\right)$ coin tosses would suffice for this task, we have reached a contradiction. □

5 Concluding Remarks

Tsallis entropy, defined as [Tsa87]

$$S_\alpha(\boldsymbol{p}) = \frac{\mathbf{k}}{\alpha-1}\left(1 - \sum_{i=1}^{n} (\boldsymbol{p}[i])^\alpha\right),$$

where $\alpha \in \mathbb{R}$, and $\mathbf{k}$ is the Boltzmann constant, is a generalization of Boltzmann-Gibbs entropy. Harvey *et al.* [HNO08] gave an algorithm to estimate Tsallis entropy, and used it to estimate Shannon entropy in the most general streaming model. Recalling the link shown in Lemma 8 between the tasks of estimating Rényi entropy $H_\alpha(\boldsymbol{p})$ and the α^{th} moment $\mathcal{M}_\alpha(\boldsymbol{p})$, the results we obtained for Rényi entropy can be extended easily to Tsallis entropy:

Remark 13. Let $\alpha > 1$ be a real number. In both the SAMP+PMF and the SAMP+CDF models, $\Theta(n^{1-1/\alpha}/2^{\Delta})$ queries are necessary and sufficient to estimate (with high probability) the Tsallis entropy $S_\alpha(\boldsymbol{p})$ of an unknown distribution $\boldsymbol{p}$ on $[n]$ to within $\pm\Delta$.

One problem left open by our work is that of optimal lower bounds for estimating the Rényi entropy $H_\alpha(\boldsymbol{p})$ in the SAMP+PMF model for $\alpha < 1$. The work [AOST15] showed that in the model where only SAMP are allowed, $\widetilde{\Omega}(n^{1/\alpha})$ queries are necessary when $0 < \alpha < 1$. It is interesting to ask whether the bound in Lemma 11 is optimal.

Acknowledgments We thank Clément Canonne for his assistance with our questions about the literature.

References

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[AOST15] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating Rényi entropy. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015.

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[BG06] Lakshminath Bhuvanagiri and Sumit Ganguly. Estimating entropy over data streams. In *Proceedings of the 14th Annual European Symposium on Algorithms*, pages 148–159, 2006.

[Can15] Clement Canonne. A survey on distribution testing: Your data is big. But is it blue? Technical Report TR15-063, ECCC, 2015.

[CCM07] Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for computing the entropy of a stream. pages 328–335, 2007.

[CDM06] Amit Chakrabarti, Khanh Do Ba, and S Muthukrishnan. Estimating entropy and entropy norm on data streams. *Internet Mathematics*, 3(1):63–78, 2006.

[CR14] Clément Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. Technical Report 1402.3835, arXiv, 2014.

[GMV06] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 733–742. ACM, 2006.

[HNO08] Nicholas Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 489–498, 2008.

[KMR$^+$94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt. Rubinfeld, Robert Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pages 273–282, 1994.

[LSO$^+$06] Ashwin Lall, Vyas Sekar, Mitsunori Ogihara, Jun Xu, and Hui Zhang. Data streaming algorithms for estimating entropy of network traffic. In *Proceedings of ACM SIGMETRICS*, pages 145–156, 2006.

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

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

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012.

[Tsa87] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. Technical Report CBPF-NF-062/87, CBPF, 1987.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VV11a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. Technical Report TR10-179, Electronic Colloquium on Computational Complexity, 2011.

[VV11b] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTsse. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 685–694, 2011.

Appendix A

Proof (of Lemma 11). By definition

$$0 \leq H_\alpha(\boldsymbol{p}) \leq \log n.$$

Then

$$1 \leq \mathcal{M}_\alpha(\boldsymbol{p}) \leq n^{1-\alpha}, \ \alpha < 1 \quad \text{and} \quad n^{1-\alpha} \leq \mathcal{M}_\alpha(\boldsymbol{p}) \leq 1, \ \alpha > 1 \ . \tag{2}$$

Define $\theta = \dfrac{1}{1-\epsilon}$. Because $\theta > 1$ we use the empirical estimator constructed in Lemma 8, which has the property that

$$\mathbf{Pr}\left[\left|\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i} - M_\theta\left(\boldsymbol{p}\right)\right| \leq \mathcal{M}_\theta\left(\boldsymbol{p}\right)2^{\varDelta}\right] \geq .99 \ .$$

By using (2), it is easy to see that with probability at least .99,

$$\left|\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{Y_i} - \mathcal{M}_\theta(\boldsymbol{p})\right| \leq \mathcal{M}_\alpha(\boldsymbol{p})2^{\varDelta}.$$

Then

$$\begin{aligned}\left|\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i} - \mathcal{M}_\alpha(\boldsymbol{p})\right| &\leq \left|\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{Y_i} - \mathcal{M}_\theta(\boldsymbol{p})\right| + \left|\mathcal{M}_\theta(\boldsymbol{p}) - \mathcal{M}_\alpha(\boldsymbol{p})\right| \\ &\leq \mathcal{M}_\alpha(\boldsymbol{p})(2^{\varDelta}+1) \quad \text{with high probability.}\end{aligned}$$

Thus, we can estimate $H_\alpha\left(\boldsymbol{p}\right)$ to a desired additive accuracy of $\varDelta$ with $O\left(\dfrac{n^\epsilon}{2^{\varDelta}}\right)$ queries. □