The complexity of approximating the entropy*

Tuğkan Batu[†] Sanjoy Dasgupta[‡] Ravi Kumar[§] Ronitt Rubinfeld[¶]

March 7, 2002

Abstract

We consider the problem of approximating the entropy of a discrete distribution under several models. If the distribution is given explicitly as an array where the i-th location is the probability of the i-th element, then linear time is both necessary and sufficient for approximating the entropy.

We consider a black-box model in which the algorithm is given access only to independent samples from the distribution. In this model, we show that a γ-multiplicative approximation to the entropy can be obtained in $O\left(n^{(1+\eta)/\gamma^2}\mathrm{poly}(\log n)\right)$ time for distributions with entropy $\Omega(\gamma/\eta)$, where n is the size of the domain of the distribution and η is an arbitrarily small positive constant. We show that one cannot get a multiplicative approximation to the entropy in general in this model. Even for the class of distributions to which our upper bound applies, we obtain a lower bound of $\Omega\left(n^{\max(1/(2\gamma^2),2/(5\gamma^2-2))}\right)$.

We next consider a hybrid model in which the distribution is available both as an explicit array and as a black-box. In this model, significantly more efficient algorithms can be achieved: a γ-multiplicative approximation to the entropy can be obtained in $O\left(\frac{\gamma^2\log^2 n}{h^2(\gamma-1)^2}\right)$ time for distributions with entropy $\Omega(h)$; for these class of distributions, we show a lower bound of $\Omega\left(\frac{\log n}{h(\gamma^2-1)}\right)$.

Finally, we consider two special families of distributions: those for which the probability of an element decreases monotonically in the label of the element, and those that are uniform over a subset of the domain. In each case, we give more efficient algorithms for approximating the entropy.

---

*A preliminary version of this paper appeared in the 34th ACM Symposium on Theory of Computing, 2002, Montreal, Canada.

†Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104. This work was supported by ONR N00014-97-1-0505, MURI. Phone: (215) 898 9040, Fax: (215) 898 0587, Email: batu@saul.cis.upenn.edu

‡AT & T Labs – Research, Florham Park, NJ 07932. Phone: (973) 360 8328, Fax: (973) 360 8970, Email: dasgupta@research.att.com

§IBM Almaden Research Center, San Jose, CA 95120. Phone: (408) 927 1885, Fax: (408) 927 3215, Email: ravi@almaden.ibm.com

¶NEC Research Institute, Princeton, NJ 08540. Phone: (609) 951 2719, Fax: (609) 951 2488, Email: ronitt@research.nj.nec.com

1 Introduction

The Shannon entropy is a measure of the randomness of a distribution, and plays a central role in statistics, information theory, and data compression. Knowing the entropy of a random source can shed light on the compressibility of data produced by such a source. In this paper we consider the complexity of approximating the entropy under various different assumptions on the way the input is presented.

Suppose the distribution is given *explicitly* as an array where the i-th location contains the probability assigned to the i-th element of the domain. It is clear that an algorithm that reads the whole representation can calculate the exact value of the entropy. However, it is also easy to see that in this model, linear time in the size of the domain is required even to approximate the entropy: Consider two distributions, one with a singleton support set (zero entropy) and the other with a two-element support set (positive entropy). Any algorithm which approximates the entropy to within any multiplicative factor must distinguish these two distributions, however, distinguishing between two such distributions requires linear (even randomized) time in general.

Next suppose the distribution is given as a *black-box* which generates samples according to the distribution. This model has been considered in both the statistics and physics communities (cf., [6, 9, 7, 8]), though none of the previous works provides a rigorous analysis of the computational efficiency and sample complexity in terms of the approximation quality. Furthermore, to the best of our knowledge, the only algorithms which do not require superlinear (in the domain size) sample complexity are those presented in [7, 8]. These algorithms use estimates of the collision probability to give a reasonable lower bound estimate of the entropy, however, the quality of the estimate is not analyzed.

1.1 Our results

(1) THE BLACK-BOX MODEL: When the distribution is given as a black-box, we show that the entropy can be approximated well in sublinear time for a large class of distributions. Informally, a γ-multiplicative approximation to the entropy can be obtained in time $O(n^{(1+\eta)/\gamma^2}\text{poly}(\log n))$, where n is the size of the domain of the distribution and η is an arbitrarily small positive constant, provided that the distribution has $\Omega(\gamma/\eta)$ entropy. We show that one cannot get a multiplicative approximation to the entropy in general. But, even for the class of distributions to which our upper bound applies, we obtain an almost matching lower bound of $\Omega(n^{\max(1/(2\gamma^2), 2/(5\gamma^2-2))})$. Our algorithm is simple—we partition the elements in the domain as big or small based on their probability masses and approximate the entropy of the big and small elements separately.

It is interesting to consider what these bounds imply for the complexity of achieving a 2-approximation for distributions with sufficiently high entropy: Our upper bound yields an algorithm which runs in $O\left(n^{\frac{1+o(1)}{4}}\right)$ time. Our lower bound demonstrates that any algorithm that 2-approximates the entropy requires $\Omega(n^{1/8})$ time.

(2) THE HYBRID MODEL: We then consider a *hybrid* model, in which the distribution is given to the algorithm both explicitly and as a black-box. We assume that the two representations are consistent. (The work of [1] shows how to ascertain that the two representations are of distributions which are at least close to each other.) In the hybrid model, we give γ-approximation algorithms which run in time $O\left(\frac{\gamma^2 \log^2 n}{h^2(\gamma-1)^2}\right)$ for distributions with entropy $\Omega(h)$; we also show a lower bound of $\Omega\left(\frac{\log n}{h(\gamma^2-1)}\right)$ for this class of distributions.

(3) SPECIAL FAMILIES OF DISTRIBUTIONS: Finally we consider two families of distributions for

which we show more efficient upper bounds. The first family is that of *monotone distributions*, in which the the probability of an element decreases monotonically in the label of the element. We give an $O(\log^2 n/\log\gamma)$-time (resp. $O((\log n)^{6+(3/(\gamma^{1/2}-1))}\text{poly}(\gamma))$-time) algorithm for γ-approximating the entropy in the explicit model (resp. black-box model). The second family is that of *subset-uniform distributions*, in which the distribution is uniform over some subset of the domain. In this case we give $O(\sqrt{k})$-time algorithms for approximating the entropy, where k is the size of the subset.

1.2 Related work

The work of Goldreich and Vadhan [5] considers the complexity of approximating the entropy in a different model in which a distribution Y is encoded as a circuit C such that $Y = C(X)$, where the input X to the circuit is uniformly distributed; in this model, they show that a version of the problem is complete for statistical zero-knowledge.

The work of [2] and [1] consider algorithms for testing other properties of distributions in the black-box model. The properties considered are whether two input distributions are close or far, and whether a joint distribution is independent, respectively. Both works give algorithms whose sample complexity is sublinear in the domain size as well as lower bounds showing the algorithms to be nearly optimal.

1.3 Organization

In Section 2, we introduce the basic definitions used in this paper. In Section 3, we give algorithms and lower bounds for the case when the input is presented in the black-box model. In Section 4, we give algorithms and lower bounds for the case when the input is presented in the hybrid model. We give more efficient algorithms for two families of distributions in Section 5 and Section 6. Finally, we close with a remark comparing our methods to those of [7, 8] in Section 7.

2 Preliminaries

We consider discrete distributions over a domain of size n, which we denote by $[n] \stackrel{\text{def}}{=} \{1,\ldots,n\}$. Let $\mathbf{p} = \langle p_1,\ldots,p_n\rangle$ be such a distribution where $p_i \geq 0, \sum_{i=1}^n p_i = 1$. The representation of $\mathbf{p}$ is called *explicit* if we are given an array where the i-th entry is p_i. The representation of $\mathbf{p}$ is called *black-box* if we are given a source which generates samples independently according to $\mathbf{p}$. The representation of $\mathbf{p}$ is called *hybrid* if we are given both the explicit and black-box representation of $\mathbf{p}$.

The entropy of the distribution $\mathbf{p}$ is defined as

$$H(\mathbf{p}) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \log p_i$$

(all the logarithms are to the base 2). For a set $S \subseteq [n]$, we define $w_{\mathbf{p}}(S) \stackrel{\text{def}}{=} \sum_{i\in S} p_i$ and we define the entropy restricted to this subset as

$$H_S(\mathbf{p}) \stackrel{\text{def}}{=} -\sum_{i\in S} p_i \log p_i.$$

Notice that $H_S(\mathbf{p}) + H_{[n]\setminus S}(\mathbf{p}) = H(\mathbf{p})$. For a distribution $\mathbf{p}$, the set of indices of high probabilities is defined as:

$$B_\alpha(\mathbf{p}) \stackrel{\text{def}}{=} \{i \in [n] \mid p_i \geq n^{-\alpha}\}.$$

Given $\gamma > 1$, we say that $\mathcal{A}$ is a *γ-approximation algorithm* in model $\mathcal{R}$ for the entropy, if for every input $\mathbf{p}$ given in representation $\mathcal{R}$, $\mathcal{A}$ outputs $\mathcal{A}(\mathbf{p})$ such that $H(\mathbf{p})/\gamma \leq \mathcal{A}(\mathbf{p}) \leq \gamma H(\mathbf{p})$ with probability at least $3/4$.

Let $\mathcal{D}_h$ be the family of distributions with entropy at least h. For a distribution $\mathbf{p}$, we denote its L_2-norm by $\|\mathbf{p}\| \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^n p_i^2}$ and its L_∞-norm by $\|\mathbf{p}\|_\infty \stackrel{\text{def}}{=} \max_{i=1}^n p_i$. For distributions $\mathbf{p}, \mathbf{q}$, we denote the L_1-distance between them by $|\mathbf{p} - \mathbf{q}| \stackrel{\text{def}}{=} \sum_{i=1}^n |p_i - q_i|$.

3 The black-box model

3.1 Upper bounds

In this section we obtain an algorithm for estimating the entropy of a large class of distributions in the black-box model. We prove the following theorem:

Theorem 1 *For any $\gamma > 1$ and $0 < \epsilon_o < 1/2$, there exists an algorithm in the black-box model that can approximate the entropy of a distribution on $[n]$ to within a multiplicative factor of $(1 + 2\epsilon_o)\gamma$ with probability at least 3/4 in $O((n^{\frac{1}{\gamma^2}}/\epsilon_o^2) \cdot \mathrm{poly}(\log n))$ time, provided that the entropy of the distribution is at least $\frac{3\gamma}{2\epsilon_o(1-2\epsilon_o)}$.*

Given $\eta > 0$ and $\gamma' > 1$, one can set $\gamma = \gamma'/(1+2\epsilon_o)$ in the above algorithm and choose ϵ_o small enough to yield a γ'-approximation algorithm whose running time is $O(n^{(1+\eta)/\gamma'^2}\mathrm{poly}(\log n))$. Note that choosing η to be small affects both the running time and the family of distributions to which the algorithm can be applied.

The main idea behind the algorithm is the following. We classify elements in $[n]$ as big or small based on their probability mass. We then approximate the contribution of the entropy of the big and small elements separately. Section 3.1.1 shows how to approximate the entropy of the big elements, Section 3.1.2 shows how to approximate the entropy of the small elements, and Section 3.1.3 combines these approximations to yield Theorem 1.

3.1.1 Approximating the entropy of the big elements

To estimate the entropy of big elements, we approximate the probability of each of the big elements by sampling the black-box sufficiently many times.

Lemma 2 *For every $0 < \alpha, \epsilon_o \leq 1$ and sufficiently large n, there is an algorithm that uses $O((n^\alpha/\epsilon_o^2) \cdot \log n)$ samples from $\mathbf{p}$ and outputs $\mathbf{q}$ such that with probability at least $1 - n^{-1}$, the following hold for all i:*

1. *if $i \in B_\alpha(\mathbf{p})$, then $|p_i - q_i| \leq \epsilon_o p_i$, and*
2. *if $p_i \leq \frac{1-\epsilon_o}{1+\epsilon_o} n^{-\alpha}$, then $q_i \leq (1 - \epsilon_o)n^{-\alpha}$*

Proof. Let $m = O((n^\alpha/\epsilon_o^2) \cdot \log n)$. Fix i and let X_j be the indicator variable that indicates j-th sample is i. Let $q_i = \sum X_j/m$. By Chernoff bounds, if $p_i \geq n^{-\alpha}$, then

$$\Pr\left[q_i > (1 + \epsilon_o)p_i\right] \leq \exp\left(-\frac{\epsilon_o^2 p_i m}{3}\right) \leq \exp\left(-\frac{\epsilon_o^2 n^{-\alpha} m}{3}\right) \leq \frac{1}{n^2}.$$

Using a similar argument for the other direction, we can bound the probability that any element i such that $p_i \geq n^{-\alpha}$ is not estimated within $1+\epsilon_o$. Using Chernoff bounds again, we can show that for i such that $p_i < \frac{1-\epsilon_o}{1+\epsilon_o} n^{-\alpha}$,

$$\Pr\left[q_i > (1-\epsilon_o)n^{-\alpha}\right] \leq n^{-2}.$$

Hence, if $i \in B_\alpha(\mathbf{p})$ then $|p_i - q_i| \leq \epsilon_o p_i$. Now, (1) and (2) of the lemma follow from a union bound over all i. ∎

The following lemma shows that the entropy of elements in $B_\alpha(\mathbf{p})$ can be approximated well using $\mathbf{q}$ instead of $\mathbf{p}$.

Lemma 3 *For any set $B \subseteq [n]$ such that for each $i \in B$, $|p_i - q_i| \leq \epsilon_o p_i$,*

$$|H_B(\mathbf{q}) - H_B(\mathbf{p})| \leq \epsilon_o H_B(\mathbf{p}) + 2\epsilon_o w_{\mathbf{p}}(B).$$

Proof. For $i \in B$, let $q_i = (1+\varepsilon_i)p_i$ such that $|\varepsilon_i| \leq \epsilon_o$.

$$\begin{aligned}
H_B(\mathbf{q}) - H_B(\mathbf{p}) &= -\sum (1+\varepsilon_i)p_i \log((1+\varepsilon_i)p_i) + \sum p_i \log p_i \\
&= -\sum (1+\varepsilon_i)p_i \log p_i - \sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i) + \sum p_i \log p_i \\
&= -\sum \varepsilon_i p_i \log p_i - \sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i).
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
|H_B(\mathbf{q}) - H_B(\mathbf{p})| &\leq \left|-\sum \varepsilon_i p_i \log p_i\right| + \left|\sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i)\right| \\
&\leq \sum -|\varepsilon_i| p_i \log p_i + \sum p_i \left|(1+\varepsilon_i)\log(1+\varepsilon_i)\right| \\
&\leq \epsilon_o H_B(p) + 2\epsilon_o w_{\mathbf{p}}(B).
\end{aligned}$$

The last step above uses the fact that for $|\varepsilon| \leq \epsilon_o \leq 1$, $|(1+\varepsilon)\log(1+\varepsilon)| \leq 2|\varepsilon| \leq 2\epsilon_o$. ∎

3.1.2 Approximating the entropy of the small elements

Now, we obtain estimates on the entropy of the small elements. Suppose set S is such that $S \subseteq [n] \setminus B_\alpha(\mathbf{p})$.

First of all, if $w_{\mathbf{p}}(S) \leq n^{-\alpha}$, the contribution of entropy from S is below any constant and can be ignored. So, we can assume without loss of generality that $w_{\mathbf{p}}(S) \geq n^{-\alpha}$. In this case, by considering the set S as a single element and using a similar argument to that in the proof of Lemma 2, the following holds with high probability: $(1-\epsilon_o)w_{\mathbf{p}}(S) \leq w_{\mathbf{q}}(S) \leq (1+\epsilon_o)w_{\mathbf{p}}(S)$. (Note that this is stronger than a $(1+\epsilon_0)$-approximation.)

The following lemma shows how to approximate the entropy of small elements.

Lemma 4 $\alpha w_{\mathbf{p}}(S)\log n \leq H_S(\mathbf{p}) \leq w_{\mathbf{p}}(S)\log n + 1/e$.

Proof. Observe that $H_S(\mathbf{p})$ is a symmetric concave function of $p_1, \ldots, p_n$. To find the maximum value of $H_S(\mathbf{p})$ subject to the constraint that $\sum_{i\in S} p_i = w_{\mathbf{p}}(S)$, we use Lagrange multipliers. Let $u(\mathbf{p},\lambda) = H_S(\mathbf{p}) + \lambda((\sum_{i\in S} p_i) - w_{\mathbf{p}}(S))$. The maximum is attained when $\partial u/\partial p_i = -\log p_i - (\ln 2)^{-1} + \lambda = 0$ for $i = 1,\ldots,n$ and $\partial u/\partial\lambda = \sum_{i\in S} p_i - w_{\mathbf{p}}(S) = 0$, which yields $p_i = w_{\mathbf{p}}(S)/|S|, \forall i$. This concludes the proof of the upper bound, since in this case,

$$H_S(\mathbf{p}) = w_{\mathbf{p}}(S)\log(|S|/w_{\mathbf{p}}(S)) = w_{\mathbf{p}}(S)\log|S| - w_{\mathbf{p}}(S)\log w_{\mathbf{p}}(S) \leq w_{\mathbf{p}}(S)\log n + 1/e$$

for these values of p_i's. Here, we used $-w_{\mathbf{p}}(S)\log w_{\mathbf{p}}(S) \le 1/e$.

Since $H_S(\mathbf{p})$ is a symmetric concave function it will take its minimum value when as many as possible of its variables are at their extreme points, namely, 0 and 1. This follows from the following: for $f(x) \stackrel{\text{def}}{=} -x\log x$, $f(a)+f(b) \le f(a+\xi)+f(b-\xi)$ when $a < a+\xi < b-\xi < b$ and consequently, the entropy value of small elements could be reduced further when they are not on one of their extreme points. So, $H_S(\mathbf{p})$ will take its minimum value when $n^{\alpha}w_{\mathbf{p}}(S)$ of p_i's have the value $n^{-\alpha}$, and the rest is 0. In this case, $H_S(\mathbf{p}) = \alpha w_{\mathbf{p}}(S)\log n$. ∎

3.1.3 Putting it together

In this section we describe our approximation algorithm to $H(\mathbf{p})$ and prove Theorem 1. The following is our algorithm for obtaining a γ-approximation to the entropy:

Algorithm ApproximateEntropy(γ, ϵ_o)

1. $\alpha = 1/\gamma^2$.
2. Get $O((n^{\alpha}/\epsilon_o^2)\cdot \text{poly}(\log n))$ samples from $\mathbf{p}$.
3. Let $\mathbf{q}$ be the normalized frequencies of $[n]$ in the sample and $B = \{i \mid q_i > (1-\epsilon_o)n^{-\alpha}\}$ (Notice $B_\alpha(\mathbf{p}) \subseteq B$).
4. Output $H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}([n]\backslash B)\log n}{\gamma}$.

Proof. (of Theorem 1) Let $S = [n]\backslash B$. Using Lemma 3 and Lemma 4,

$$\begin{aligned}
H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}(S)\log n}{\gamma} &\le (1+\epsilon_o)H_B(\mathbf{p}) + 2\epsilon_o + \frac{1+\epsilon_o}{\gamma}w_{\mathbf{p}}(S)\log n \\
&\le (1+\epsilon_o)(H_B(\mathbf{p}) + \gamma H_S(\mathbf{p})) + 2\epsilon_o \\
&\le (1+\epsilon_o)\gamma H(\mathbf{p}) + 2\epsilon_o \\
&\le (1+2\epsilon_o)\gamma H(\mathbf{p}),
\end{aligned}$$

if $H(\mathbf{p}) \ge 2/\gamma$. Similarly,

$$\begin{aligned}
H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}(S)\log n}{\gamma} &\ge (1-\epsilon_o)H_B(\mathbf{p}) - 2\epsilon_o + \frac{1-\epsilon_o}{\gamma}w_{\mathbf{p}}(S)\log n \\
&\ge (1-\epsilon_o)(H_B(\mathbf{p}) + \frac{(H_S(\mathbf{p}) - e^{-1})}{\gamma}) - 2\epsilon_o \\
&= (1-\epsilon_o)(H_B(\mathbf{p}) + H_S(\mathbf{p})/\gamma) - \frac{1-\epsilon_o}{\gamma}e^{-1} - 2\epsilon_o \\
&\ge H(\mathbf{p})/((1+2\epsilon_o)\gamma),
\end{aligned}$$

if $H(\mathbf{p}) \ge \frac{3\gamma}{2\epsilon_o(1-2\epsilon_o)} \ge 2/\gamma$. ∎

3.2 Lower bounds

In this section we prove lower bounds on the number of samples needed to approximate the entropy of a distribution to within a multiplicative factor of $\gamma > 1$. All of our lower bounds are shown by giving pairs of distributions that are hard to distinguish and have entropy ratio at least γ^2.

The lower bounds follow since an algorithm which γ-approximates the entropy would allow one to distinguish the distributions.

First, we show that because distributions could have zero entropy, there is no algorithm which can γ-approximate the entropy of *every* distribution.

Theorem 5 *For any $\gamma > 1$, there is no algorithm which γ-approximates the entropy of every distribution in the black-box model.*

Proof. Assume that $\mathcal{A}$ is an algorithm which approximates the entropy of any distribution. For some small constant $0 < c < 1$, let cn^α be an upper bound on the runtime of $\mathcal{A}$ on distributions over $[n]$. Consider the two distributions $\mathbf{p}$ and $\mathbf{q}$ where $\mathbf{p} = \langle 1, 0, \ldots, 0\rangle$ and $\mathbf{q} = \langle 1 - n^{-\alpha}, n^{-\alpha-1}, \ldots, n^{-\alpha-1}\rangle$. Any algorithm which uses only cn^α samples cannot distinguish between $\mathbf{p}$ and $\mathbf{q}$ with high probability. Since the entropy of $\mathbf{p}$ is 0, any algorithm which gives a multiplicative approximation should output 0. On the other hand, any algorithm which approximates the entropy of $\mathbf{q}$ to within a multiplicative factor of γ should output a value which is at least $\frac{1}{\gamma}\alpha n^{-\alpha}\log n > 0$. Thus, any algorithm which γ-approximates the entropy would be able to distinguish between $\mathbf{p}$ and $\mathbf{q}$. ■

We now show (Theorem 9) a lower bound of $\Omega(n^{\frac{2}{5\gamma^2-2}})$ samples to γ-approximate the entropy for any distribution in $\mathcal{D}_{(5\log n)/(10\gamma^2-4)}$. Before we show this lower bound, we show a simpler lower bound of $\Omega(n^{1/(2\gamma^2)})$ samples.

Theorem 6 *For any $\gamma > 1$ and sufficiently large n, any algorithm in the black-box model that γ-approximates the entropy of a distribution over $[n]$ in $\mathcal{D}_{(\log n)/\gamma^2}$ is required to make $\Omega(n^{1/(2\gamma^2)})$ samples from it.*

Proof. Consider two distributions $\mathbf{p}$ and $\mathbf{q}$ on n elements where $\mathbf{p}$ is uniform on the set $[n]$ and $\mathbf{q}$ is uniform on a set S that is a randomly chosen subset of $[n]$ of size n^{1/γ^2}. It is easy to see that $H(\mathbf{p})/H(\mathbf{q}) = \gamma^2$. By the Birthday Paradox, with probability 1/3, we do not see any repetitions in the sample before we take $\delta n^{1/(2\gamma^2)}$ samples from either distribution for some constant $\delta < 1$. Hence, $\Omega(n^{1/(2\gamma^2)})$ samples are needed to distinguish these distributions. ■

We now prove a better lower bound when $\gamma < \sqrt{2}$. Our lower bound proof borrows the outline of the arguments in [2, 1]. We start by giving a canonical form for entropy approximation algorithms. In particular, we describe an aggregate representation of the samples such an algorithm takes. We then prove that we can assume, without loss of generality, that the algorithm is given this representation of the samples as input instead of the samples themselves.

Let $S = \{x_1, \ldots, x_s\}$ be a set of samples from a distribution over $[n]$.

Definition 7 (Fingerprint) *The* fingerprint d_S *is the function* $d_S : [s] \to [n] \cup \{0\}$ *such that* $d_S(i)$ *is the number of elements x such that x appears i times in S.*

We will just use $d(i)$ when S is clear from the context. The next lemma shows that the fingerprint of a set of samples is just as useful as the samples themselves to approximate the entropy.

Lemma 8 *Given algorithm $\mathcal{A}$ that γ-approximates the entropy of a distribution from samples, there exists an algorithm $\mathcal{A}'$ which gets as input only the fingerprint of the generated sample and has the error probability upper bounded by that of $\mathcal{A}$.*

Proof. We will construct algorithm $\mathcal{A}'$ using calls to algorithm $\mathcal{A}$. Let the input distribution be $\mathbf{p}$. Upon getting the fingerprint of a sample S, $\mathcal{A}'$ chooses $d_S(i)$ elements at random from $[n]$ without replacement for each i. Then, $\mathcal{A}'$ passes the multiset S' such that an element chosen for i appears exactly i times in S'. Finally, $\mathcal{A}'$ outputs the value that $\mathcal{A}$ outputs on S'.

Let π be a permutation on $[n]$. Define $\pi(\mathbf{p})$ to be distribution such that $\pi(\mathbf{p})_i = p_{\pi(i)}$ (i.e., the relabeling of the domain elements according to π). Define $\pi(S)$ for a sample set S to be the set S' that is the relabeling of the samples in S according to π.

Note that $\mathcal{A}'$ actually simulates $\pi(\mathbf{p})$ for $\mathcal{A}$ for some randomly chosen permutation π. Denote the output of $\mathcal{A}$ on samples S by $\mathcal{A}(S)$. Hence,

$$\begin{aligned}
&\Pr\left[\mathcal{A}' \ \gamma\text{-approximates } H(\mathbf{p})\right] \\
&= \sum_S \Pr\left[\mathbf{p} \text{ generates } S\right] \mathrm{E}\left[\Pr\left[\mathcal{A}(\pi(S)) \ \gamma\text{-approximates } H(\mathbf{p})\right]\right] \\
&= \mathrm{E}\left[\sum_S \Pr\left[\mathbf{p} \text{ generates } S\right] \cdot \Pr\left[\mathcal{A}(\pi(S)) \ \gamma\text{-approximates } H(\pi(\mathbf{p}))\right]\right] \\
&= \mathrm{E}\left[\sum_S \Pr\left[\pi(\mathbf{p}) \text{ generates } \pi(S)\right] \cdot \Pr\left[\mathcal{A}(\pi(S)) \ \gamma\text{-approximates } H(\pi(\mathbf{p}))\right]\right] \\
&= \mathrm{E}\left[\Pr\left[\mathcal{A} \ \gamma\text{-approximates } H(\pi(\mathbf{p}))\right]\right] \\
&\geq \min_\pi \Pr\left[\mathcal{A} \ \gamma\text{-approximates } H(\pi(\mathbf{p}))\right],
\end{aligned}$$

which is the correctness probability of $\mathcal{A}$. ∎

Theorem 9 *For any $\gamma > 1$, sufficiently large n, and any algorithm $\mathcal{A}$ using $o(n^{2/(5\gamma^2-2)})$ samples, there exist two distributions $\mathbf{p}$ and $\mathbf{q}$ over $[n]$ with $H(\mathbf{p})/H(\mathbf{q}) \geq \gamma^2$, such that $\mathcal{A}$ cannot distinguish between $\mathbf{p}$ and $\mathbf{q}$ with probability greater than* $2/3$.

Proof. Let $\alpha = 2/(5\gamma^2 - 2)$ and $\beta = 3\alpha/2$. Fix an algorithm $\mathcal{A}$ that uses $o(n^\alpha)$ samples. Next, we define the distributions $\mathbf{p}$ and $\mathbf{q}$ from the theorem statement. Using Lemma 8, we will assume that $\mathcal{A}$ uses the fingerprints of the samples instead of the samples themselves.

$$p_i \stackrel{\text{def}}{=} \begin{cases} n^{-\alpha} & 1 \leq i \leq n^\alpha/2 \\ n^{-1} & n/2 < i \leq n \\ 0 & \text{otherwise} \end{cases}$$

$$q_i \stackrel{\text{def}}{=} \begin{cases} n^{-\alpha} & 1 \leq i \leq n^\alpha/2 \\ n^{-\beta} & n/2 < i \leq (n + n^\beta)/2 \\ 0 & \text{otherwise} \end{cases}$$

Call i *big* if $1 \leq i \leq n^\alpha/2$ and *small* if otherwise and $p_i > 0$.

Note that $H(\mathbf{p}) = (\alpha+1)/2 \cdot \log n$ and $H(\mathbf{q}) = (\alpha+\beta)/2 \cdot \log n$. Therefore,

$$\frac{H(\mathbf{p})}{H(\mathbf{q})} = \frac{\alpha+1}{\alpha+\beta} = \left(\frac{5\gamma^2}{5\gamma^2-2}\right)\left(\frac{5\gamma^2-2}{5}\right) = \gamma^2.$$

When restricted to the big elements, both distributions are identical. The only difference between $\mathbf{p}$ and $\mathbf{q}$ comes from the small elements, and the crux of the proof will be to show that this difference will not change the relevant statistics significantly. Then, we conclude the proof by showing that distributions on the fingerprints when the samples are taken from $\mathbf{p}$ or $\mathbf{q}$ are indistinguishable.

The following lemma shows that it is only the big elements, which have identical distributions in both $\mathbf{p}$ and $\mathbf{q}$, that contribute to most of the entries in the fingerprint.

Lemma 10 *The expected number of small elements that occur at least three times in the sample is $o(1)$ for both* $\mathbf{p}$ *and* $\mathbf{q}$.

Proof. For a fixed small element, the probability that at least three samples from $\mathbf{q}$ will be this element is $o(n^{3(\alpha-\beta)})$. Since there are $n^{-\beta}/2$ small elements, by the linearity of expectation, the expected number of such small elements in the sample is $o(1)$. ■

It would be useful if we could assume that the frequency of each element is independent of the frequencies of the other elements. To allow this, we assume that algorithm $\mathcal{A}$ first chooses an integer s_1 from the Poisson distribution with the parameter $\lambda = s = o(n^\alpha)$. The Poisson distribution with the positive parameter λ has the probability mass function $p(k) = \exp(-\lambda)\lambda^k/k!$. Then, after taking s_1 samples from the input distribution, $\mathcal{A}$ decides whether to accept or reject the distribution. We will show that when $\mathcal{A}$ chooses s_1 this way, the frequencies of the elements will be independent.

Next, we show that $\mathcal{A}$ cannot distinguish $\mathbf{p}$ from $\mathbf{q}$ with success probability at least $2/3$. Since s_1 will have a value larger than $s/2$ with probability at least $1 - o(1)$ and we will show an upper bound on the statistical distance of the distributions of two random variables (i.e., the distributions on the fingerprints), it will follow that no symmetric algorithm with sample complexity $s/2$ can distinguish $\mathbf{p}$ from $\mathbf{q}$.

Let F_i be the random variable corresponding to the number of times the element i appears in the sample. The following fact is well-known (cf. [3], pp. 216).

Fact 11 *For each i, F_i is distributed identically to the Poisson distribution with parameter $\lambda = sr$, where r is the probability of element i. Furthermore, all F_i's are mutually independent.*

Using this fact, the total number of samples from the big elements and the total number of samples from the small elements are independent.

Let $D_\mathbf{p}$ and $D_\mathbf{q}$ be the distributions on all possible fingerprints when samples are taken from $\mathbf{p}$ and $\mathbf{q}$, respectively. The rest of the proof proceeds as follows. We first construct two processes $D'_\mathbf{p}$ and $D'_\mathbf{q}$ that generate distributions on fingerprints such that $D'_\mathbf{q}$ is statistically close to $D_\mathbf{p}$ and $D'_\mathbf{q}$ is statistically close to $D_\mathbf{q}$. Then, we prove that the distributions $D'_\mathbf{p}$ and $D'_\mathbf{q}$ are statistically close. Hence, the theorem follows by the indistinguishability of $D_\mathbf{p}$ and $D_\mathbf{q}$.

Each process has two phases. The first phase is the same in both processes. They choose integers s' and s'' independently from the Poisson distribution with parameter $\lambda = s/2$. The integers s' and s'' fix the the number of samples from the big and the small elements, respectively. By the properties of Poisson distribution mentioned above, $s' + s''$ is distributed identically to s_1. Then, they randomly generate the frequency counts for each big element i using the random variables F_i defined above. The processes know which elements are big and which elements are small, although any distinguishing algorithm does not. At the end of the phase, the processes check whether the number of samples from the big elements is equal to s'. If not, they start generating the frequency counts from scratch. This concludes the first phase of the processes.

In the second phase, processes $D'_\mathbf{p}$ and $D'_\mathbf{q}$ determine the frequency counts of each small element according to $\mathbf{p}$ and $\mathbf{q}$, respectively (using F_i's). If any small element is given a total frequency count of at least three or the total number of samples generated in this phase is not exactly s'', the second phase of the process is restarted from scratch.

Since the frequency counts for all elements are determined at this point, both processes output the fingerprint of the sample they have generated.

Now, we treat the outputs of $D'_\mathbf{p}$ and $D'_\mathbf{q}$ to be distributions on $[n]$. The following lemmas conclude the proof.

Lemma 12 $\left|D'_\mathbf{p} - D_\mathbf{p}\right| = o(1)$ *and* $\left|D'_\mathbf{q} - D_\mathbf{q}\right| = o(1)$.

Proof. The distribution that $D'_{\mathbf{p}}$ generates is the distribution $D_{\mathbf{p}}$ conditioned on the event that all small elements appear at most twice in the combined sample. Since this conditioning holds true with probability at least $1-o(1)$ by Lemma 10, $|D'_{\mathbf{p}} - D_{\mathbf{p}}| \leq o(1)$. A similar argument applies to $D'_{\mathbf{q}}$ and $D_{\mathbf{q}}$. ∎

Lemma 13 $\left|D'_{\mathbf{p}} - D'_{\mathbf{q}}\right| \leq 1/6$.

Proof. By the generation process, the L_1 distance between $D'_{\mathbf{p}}$ and $D'_{\mathbf{q}}$ can only arise from the second phase. We show that the second phases of the processes do not generate an L_1 distance larger than $1/6$.

Let G (respectively, H) be the random variable that corresponds to the values $d(2)$ when the input distribution is $\mathbf{p}$ (respectively, $\mathbf{q}$). Let $d' \stackrel{\text{def}}{=} \langle d(3), d(4), \ldots, d(s) \rangle$. We will use the fact that for any d', $\Pr\left[D'_{\mathbf{p}} \text{ generates } d', s', s''\right] = \Pr\left[D'_{\mathbf{q}} \text{ generates } d', s', s''\right]$ in the following calculation.

$$
\begin{aligned}
|D'_{\mathbf{p}} - D'_{\mathbf{q}}| &= \sum_{d} \left|\Pr\left[D'_{\mathbf{p}} \text{ generates } d\right] - \Pr\left[D'_{\mathbf{q}} \text{ generates } d\right]\right| \\
&= \sum_{d',s',s''} \Pr\left[D'_{\mathbf{p}} \text{ generates } d', s', s''\right] \cdot \\
&\qquad \sum_{k,l\geq 0} |\Pr\left[D'_{\mathbf{p}} \text{ generates } (d(1), d(2)) = (k,l) \mid d', s', s''\right] \\
&\qquad\qquad - \Pr\left[D'_{\mathbf{q}} \text{ generates } (d(1), d(2)) = (k,l) \mid d', s', s''\right]| \\
&= \sum_{d',s',s''} \Pr\left[D'_{\mathbf{p}} \text{ generates } d', s', s''\right] \cdot \\
&\qquad \sum_{k\geq 0} \left|\Pr\left[D'_{\mathbf{p}} \text{ generates } d(2) = k \mid d', s', s''\right] - \Pr\left[D'_{\mathbf{q}} \text{ generates } d(2) = k \mid d', s', s''\right]\right| \\
&= \sum_{k\geq 0} \left|\Pr\left[D'_{\mathbf{p}} \text{ generates } d(2) = k\right] - \Pr\left[D'_{\mathbf{q}} \text{ generates } d(2) = k\right]\right| \\
&= |G - H|,
\end{aligned}
$$

where the third equality follows since $s', s'', d', d(2)$ determine $d(1)$.

Consider the composition of G and H in terms of big and small elements. In the case of $\mathbf{p}$, let G_h be the number of big elements that contribute to $d(2)$ and G_l be the number of such small elements. Hence, $G = G_h + G_l$. Define H_h, H_l analogously. Then, G_h and H_h are distributed identically. In the rest of the proof, we show that the fluctuations in G_h, H_h dominate the magnitude of G_l, H_l.

Let ξ_i be the indicator random variable that takes value 1 when element i has been sampled twice. Then, $G_h = \sum_{\text{big } i} \xi_i$. By the assumption about the way samples are generated, the ξ_i's are independent. Therefore, G_h is distributed identically to the binomial distribution on the sum of n^α Bernoulli trials with success probability $\Pr[\xi_i = 1] = \exp(-s/n^\alpha)(s^2/2n^{2\alpha})$. An analogous argument shows that G_l is distributed identically to the binomial distribution with parameters $n/2$ and $\exp(-s/n)(s^2/2n^2)$. Similarly, H_l is distributed identically to the binomial distribution with parameters $n^\beta/2$ and $\exp(-s/n^\beta)(s^2/2n^{2\beta})$.

As n and m grow large enough, both G_h and G_l can be approximated well by normal distributions. Therefore, by the independence of G_h and G_l, G is also approximated well by a normal distribution. That is,

$$
\Pr[G = t] \to \frac{1}{\sqrt{2\pi}\sigma_G} \exp\left(-\frac{(t - \mathrm{E}[G])^2}{2\mathrm{Var}[G]}\right)
$$

as $n \to \infty$. Similarly, H is approximated well by a normal distribution.

Thus, $\Pr[G = t] = \Omega(1/\sigma_G)$ over an interval I_1 of length $\Omega(\sigma_G)$ centered at $\mathrm{E}[G]$. Similarly, $\Pr[H = t] = \Omega(1/\sigma_H)$ over an interval I_2 of length $\Omega(\sigma_H)$ centered at $\mathrm{E}[H]$. Since

$$\begin{aligned} \mathrm{E}[H] - \mathrm{E}[G] &= \mathrm{E}[H_l] - \mathrm{E}[G_l] \\ &\leq \mathrm{E}[H_l] \leq \exp\left(\frac{-s}{n^\beta}\right)\left(\frac{s^2}{4n^\beta}\right) = o(\sigma_H), \end{aligned}$$

$I_1 \cap I_2$ is an interval of length $\Omega(\sigma_{H_h})$. Therefore,

$$\sum_{t \in I_1 \cap I_2} |\Pr[G = t] - \Pr[H = t]| \leq o(1)$$

because for $t \in I_1 \cap I_2$, $|\Pr[G = t] - \Pr[H = t]| = o(1/\sigma_G)$. We can conclude that $\sum_t |\Pr[G = t] - \Pr[H = t]|$ is less than $1/6$ after accounting for the probability mass of G and H outside $I_1 \cap I_2$. ■

It can be seen that Theorem 9 follows from Lemma 12 and Lemma 13. ■

Thus, we obtain:

Theorem 14 *For any $\gamma > 1$ and sufficiently large n, any algorithm in the black-box model that γ-approximates the entropy of a distribution over $[n]$ in $\mathcal{D}_{(\log n)/\gamma^2}$ needs $\Omega\left(n^{\max\{1/2\gamma^2, 2/(5\gamma^2-2)\}}\right)$ samples from it.*

4 The hybrid model

In this section we consider the hybrid model where an algorithm is given access to both explicit and black-box versions of the same distribution. Our algorithm for this case hinges on an alternate interpretation of the entropy: the entropy of the distribution $\mathbf{p}$ is the expected value of $-\log p_i$ where i is distributed according to $\mathbf{p}$. The algorithm, given that the entropy value is not too small, needs only a polylogarithmic number of samples and probes.

Theorem 15 *For any $\gamma > 1$, there exists an algorithm in the hybrid model that can approximate the entropy of a distribution on $[n]$ to within a multiplicative factor of γ with probability at least $3/4$ in $O\left(\frac{\gamma^2 \log^2 n}{h^2(\gamma-1)^2}\right)$ time, provided that the entropy of the distribution is at least h.*

Proof. Let $m \geq O\left(\frac{\gamma^2 \log^2 n}{h^2(\gamma-1)^2}\right)$. The algorithm takes m samples, say $i_1, \ldots, i_m$ from $\mathbf{p}$. The output of the algorithm is $X = (1/m)\sum_{j \in m} -\log p_{i_j}$.

Define the random variable $X_j \stackrel{\text{def}}{=} -\log p_{i_j}$ for $j = 1, \ldots, m$. Clearly, $\mathrm{E}[X_j] = H(\mathbf{p})$. All that remains to show is that the variance of the summation is not too large, so that we get a γ-approximation to the entropy value. Since the X_j's are independent, it will suffice to bound the variance of an individual X_j. We bound this variance in the following lemma.

Lemma 16 $\mathrm{Var}[X_j] \leq 3 + \log^2 n$.

Proof. By the definition of variance,

$$\mathrm{Var}[X_j] = \mathrm{E}\left[X_j^2\right] - \mathrm{E}[X_j]^2 = \left(-\sum_i p_i \log^2 p_i\right) - H(\mathbf{p})^2.$$

Let $A \stackrel{\text{def}}{=} \{i \mid p_i \leq 1/e\}$, and Y be a random variable that takes value $(1/p_i)$ with probability p_i for $i \in A$, and 0 with probability $1 - w_{\mathbf{p}}(A)$. Finally, let $f(x) \stackrel{\text{def}}{=} \log^2 x$. Note that $f(x)$ is concave for $x \geq e$ and therefore by Jensen's inequality, $\mathrm{E}[f(Y)] \leq f(\mathrm{E}[Y])$ for the random variable Y. Now, we can write

$$\begin{aligned}
-\sum_i p_i \log^2 p_i &= -\sum_{i \in A} p_i \log^2 p_i - \sum_{i \notin A} p_i \log^2 p_i \\
&\leq \mathrm{E}[f(Y)] + w_{\mathbf{p}}([n] \setminus A) \log^2 e \\
&\leq f(\mathrm{E}[Y]) + w_{\mathbf{p}}([n] \setminus A) \log^2 e \\
&= \log^2 |A| + w_{\mathbf{p}}([n] \setminus A) \log^2 e \\
&\leq \log^2 n + 3
\end{aligned}$$

■

By the independence of the X_j's, $\mathrm{Var}[X] = \mathrm{Var}[X_j]/m \leq (3 + \log^2 n)/m$.

At this point, we use Chebyshev's inequality, which states that for $\rho > 0$, $\Pr[|X - \mathrm{E}[X]| \geq \rho] \leq \mathrm{Var}[X]/\rho^2$, to bound the error probability of the algorithm.

$$\begin{aligned}
\Pr[\mathcal{A} \text{ does not } \gamma\text{-approximate } \mathrm{H}(\mathbf{p})] &= \Pr[X \leq H(\mathbf{p})/\gamma \text{ or } X \geq \gamma H(\mathbf{p})] \\
&\leq \Pr[|X - H(\mathbf{p})| \geq (\gamma - 1) H(\mathbf{p})/\gamma] \\
&\leq \frac{\gamma^2 (3 + \log^2 n)}{m (H(\mathbf{p}))^2 (\gamma - 1)^2} \leq 1/3,
\end{aligned}$$

where the last inequality follows from the choice of m. ■

Corollary 17 *There exists an algorithm $\mathcal{A}$ in the hybrid model which γ-approximates $H(\mathbf{p})$ in $O((\frac{\gamma}{\gamma-1})^2)$ time when $H(\mathbf{p}) = \Omega(\log n)$,*

The next theorem gives a lower bound for the hybrid model when the entropy of the distribution is not lower bounded. The distribution used to show this lower bound has very small entropy and thus does not fall into the family of distributions for which the above upper bound applies.

Theorem 18 *For $\gamma > 1$ and sufficiently large n, any algorithm in the hybrid model that γ-approximates the entropy of a distribution over $[n]$ (with non-zero entropy) is required to take $\Omega(n^{\frac{1-o(1)}{\gamma^2}})$ samples from it.*

Proof. Let $\alpha = \frac{1-o(1)}{\gamma^2}$. Consider the following distributions $\mathbf{p}$ and $\mathbf{q}$ defined as follows:

$$p_i \stackrel{\text{def}}{=} \begin{cases} 1 - n^{-\alpha} & i = 1 \\ n^{-\alpha} & i = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$q_i \stackrel{\text{def}}{=} \begin{cases} 1 - n^{-\alpha} & i = 1 \\ n^{-1} & 2 \leq i \leq n^{1-\alpha} + 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that $H(\mathbf{p}) = (1+o(1))\alpha n^{-\alpha} \log n$ and $H(\mathbf{q}) > n^{-\alpha} \log n$. By the choice of α, $H(\mathbf{q})/H(\mathbf{p}) \geq \gamma^2$.

Let $\mathcal{P}$ be the family of distributions obtained from $\mathbf{p}$ by permuting the labels of the elements. Define $\mathcal{Q}$ similarly for $\mathbf{q}$. It is simple to verify that any algorithm taking $o(n^\alpha)$ samples and making

$o(n^{\alpha})$ probes will fail to distinguish between a randomly chosen member of $\mathcal{P}$ and a randomly chosen member of $\mathcal{Q}$ with high probability. ■

The next theorem gives a lower bound on the complexity of approximating the entropy in the hybrid model when the distribution has a lower bound on the entropy value. The proof generalizes the counterexample in Theorem 18.

Theorem 19 *For $\gamma > 1$ and sufficiently large n, any algorithm in the hybrid model that γ-approximates the entropy of a distribution over $[n]$ in $\mathcal{D}_h$ is required to take $\Omega(\log n/(h(\gamma^2 - 1))$ samples from it.*

Proof. Let $\alpha < 1$ and $k \stackrel{\text{def}}{=} \lceil 2^{h/(1-n^{-\alpha})} \rceil$. Consider the following distributions $\mathbf{p}$ and $\mathbf{q}$ defined as follows:

$$p_i \stackrel{\text{def}}{=} \begin{cases} (1-n^{-\alpha})/k & 1 \le i \le k \\ n^{-\alpha} & i = k+1 \\ 0 & \text{otherwise} \end{cases}$$
$$q_i \stackrel{\text{def}}{=} \begin{cases} (1-n^{-\alpha})/k & 1 \le i \le k \\ n^{-1} & k+1 \le i \le k+n^{1-\alpha} \\ 0 & \text{otherwise} \end{cases}$$

Note that $H(\mathbf{p}) = h + (1+o(1))\alpha n^{-\alpha} \log n$ and $H(\mathbf{q}) > h + n^{-\alpha} \log n$.

Let $\mathcal{P}$ be the family of distributions obtained from $\mathbf{p}$ by permuting the labels of the elements. Define $\mathcal{Q}$ similarly for $\mathbf{q}$. It is simple to verify that any algorithm taking $o(n^{\alpha})$ samples and making $o(n^{\alpha})$ probes will fail to distinguish between a randomly chosen member of $\mathcal{P}$ and a randomly chosen member of $\mathcal{Q}$ with high probability.

If we choose α such that $n^{\alpha}/(1-\alpha) \le (\log n)/(h(\gamma^2 - 1))$, then the entropy ratio

$$\frac{H(\mathbf{q})}{H(\mathbf{p})} > \frac{h + n^{-\alpha} \log n}{h + (1+o(1))\alpha n^{-\alpha} \log n} \ge \gamma^2.$$

This concludes the proof. ■

5 Monotone distributions

A *monotone distribution* $\mathbf{p} = \langle p_1, \ldots, p_n \rangle$ is one for which $p_i \ge p_{i+1}$ for all i. The structure of a monotone distribution makes it much easier to approximate the entropy.

5.1 The explicit model

We show that given an explicit monotone distribution, we can approximate the entropy in polylogarithmic time.

Theorem 20 *For $\gamma > 1$, there exists an algorithm in the explicit model that can approximate the entropy of a monotone distribution on $[n]$ to within a multiplicative factor of γ with probability at least $3/4$ in $O((\log^2 n)/\log \gamma)$ time, provided that the entropy of the monotone distribution is $\Omega(\gamma^2/(\sqrt{\gamma} - 1))$.*

Proof. The algorithm $\mathcal{A}$ partitions the domain elements into sets such that all the elements in a set have similar probability values, and then uses the sizes of each of the sets to estimate the entropy. The algorithm constructs this partition by using binary search on the explicit representation of the distribution for predetermined values.

Consider the following partition of $[n]$. Let $k \stackrel{\text{def}}{=} 2\log n/\log\gamma$. Let $B_0 = \{i \mid p_i \le n^{-2}\}$, and $B_j = \{i \mid \gamma^{j-1}n^{-2} < p_i \le \gamma^j n^{-2}\}$ for $j = 1, \ldots, k$. Algorithm $\mathcal{A}$ determines the boundaries of the B_j's and hence the number of elements in each B_j by binary search. At this point, with no further queries, $\mathcal{A}$ has a $\sqrt{\gamma}$-approximation to every p_i for $i \notin B_0$ (for each B_j, $\mathcal{A}$ uses the geometric mean of the upper and lower threshold to approximate the p_i values).

Using these approximate p_i values, the algorithm computes entropy values that are a γ-approximation to $H_{B_j}(\mathbf{p})$'s for $j > 0$ as follows.

Let $B = \cup_{j>1} B_j$. For each $i \in B$, let $q_i \stackrel{\text{def}}{=} c_i p_i$ be the representative probability value that the algorithm uses for B_j to which i belongs. So, $\frac{1}{\sqrt{\gamma}} \le c_i \le \sqrt{\gamma}$ for all $i \in B$. For all the other i's, let $q_i = 0$.

$$\begin{aligned}
H(\mathbf{q}) &= -\sum_i q_i \log q_i \\
&= -\sum_{i\in B} c_i p_i \log(c_i p_i)) \\
&= -\sum_{i\in B} c_i p_i \log p_i - \sum_{i\in B} c_i p_i \log c_i \\
&\le \sqrt{\gamma} H_B(\mathbf{p}) + (w_{\mathbf{p}}(B)\log e)/e \le \gamma H(\mathbf{p})
\end{aligned}$$

when $H(p) \ge \log e/(e(\gamma - \sqrt{\gamma}))$. Similarly,

$$\begin{aligned}
H(\mathbf{q}) &= -\sum_i q_i \log q_i \\
&= -\sum_{i\in B} c_i p_i \log(c_i p_i) \\
&= -\sum_{i\in B} c_i p_i \log p_i - \sum_{i\in B} c_i p_i \log c_i \\
&\ge \frac{1}{\sqrt{\gamma}} H_B(\mathbf{p}) - w_{\mathbf{p}}(B)\sqrt{\gamma}\log\sqrt{\gamma} \\
&\ge \frac{1}{\sqrt{\gamma}}\left(H(\mathbf{p}) - \frac{2\log n}{n}\right) - w_{\mathbf{p}}(B)\sqrt{\gamma}\log\sqrt{\gamma} \\
&\ge H(\mathbf{p})/\gamma
\end{aligned}$$

when $H(p) \ge (\gamma^2 + (2\sqrt{\gamma}n^{-1}\log n))/(\sqrt{\gamma} - 1)$.

Finally, the algorithm outputs $-\sum_{i\in B} q_i \log q_i$, which is a γ-approximation to $H(\mathbf{p})$.

The number of probes that the algorithm makes is $O(\log n)$ per search; therefore, the complexity of the algorithm is $O((\log^2 n)/\log\gamma)$ ■

5.2 The black-box model

We show that the entropy of a monotone distribution can be approximated in polylogarithmic time even in the black-box model. Our algorithm rests on the following observation that is formally stated in Lemma 21: if a monotone distribution $\mathbf{p}$ over $[n]$ is such that $w_{\mathbf{p}}([n/2])$ and $w_{\mathbf{p}}([n]\backslash[n/2])$ are very close, then the distribution must be close to uniform. In such a case, we can approximate the entropy of the distribution by the entropy of the uniform distribution. The main idea behind our algorithm is to recursively partition the domain into half, stopping the recursion if the total sum of the probabilities of each half are close or if they are both too small to contribute much to the total entropy. Our algorithm can be viewed as forming a tree based on the set of samples S, where the root is labeled by the range $[1, n]$, and if the node labeled by the range $[i, j]$ is partitioned, its children are labeled by $[i, (i + j)/2]$ and $[(i + j)/2 + 1, j]$, respectively. Once the partition tree is determined, the algorithm estimates the entropy by assuming that the distribution restricted to the ranges labeling each leaf is uniform, and combines the estimates in proportion to their total weight. For an interval I, let S_I denote the set of samples that are in I and $|I|$ the length of the interval.

More specifically, the procedure $\mathbf{BuildTree}(S, \beta)$ takes as input a parameter $\beta > 1$ and a multiset S of m samples from $\mathbf{p}$ and outputs a rooted binary tree T_S as follows: Let v be a node in the tree that is currently a leaf corresponding to the interval $[i, j]$ for some i and j. We determine that v will remain a leaf if either of the following two conditions is satisfied:

- $|S_{[i,j]}| < m\beta / \log^3 n$ (call v *light*), or
- $|S_{[i,(i+j)/2]}| \le \beta |S_{[(i+j)/2+1,j]}|$ (call v *balanced*).

Otherwise, we split v's interval by attaching two children to v, corresponding to the intervals $[i, (i + j)/2]$ (the left child) and $[(i + j)/2 + 1, j]$ (the right child). Let $\mathcal{I}(T_S)$ denote the set of intervals corresponding to the balanced leaves of T_S.

For each balanced interval $I \in \mathcal{I}(T_S)$, we estimate the contribution of the interval to the total entropy of the distribution. We define a function $\alpha(I)$ which is supposed to approximate the entropy in the balanced interval I. If the interval is big, we use the samples themselves to approximate the entropy and if the interval is small, we use the algorithm for heavy elements in the black-box model (Section 3) to approximate the entropy. Formally, if

$$|I| > 2\left(\frac{\log^3 n}{2\beta}\right)^{1/(\beta-1)},$$

(call such intervals *long*) then let

$$\alpha(I) \stackrel{\text{def}}{=} \frac{|S_I|}{m} \log \frac{|I|}{2}.$$

Otherwise, for *short* intervals, the procedure $\alpha(I)$ takes $O(|I| \log^6 n)$ new samples from $\mathbf{p}$ and computes the normalized frequency vector $\mathbf{q}$ of these samples. Then, $\alpha(I)$ returns $H_I(\mathbf{q})$ for any short interval I. Since the length of any short interval is bounded, the total sample complexity of $\alpha(\cdot)$ is $O(\log^{6+(3/(\beta-1))} n)$. We now give the top level description of our algorithm:

Algorithm MonotoneApproximateEntropy(γ)

1. $\beta = \sqrt{\gamma\left(1 - \frac{\log\log n}{\log n}\right)}$.
2. Get a multiset S of $O((\beta^5 \log^4 n)/(\beta - 1)^2)$ samples from $\mathbf{p}$.
3. $T_S = \mathbf{BuildTree}(S, \beta)$.
4. Output $\sum_{I \in \mathcal{I}(T_S)} \alpha(I)$.

First we show that the maximum and the minimum entropy values for an interval corresponding to a balanced leaf are fairly close.

Lemma 21 *Let I be an interval of length $2k$ in $[n]$, I_1 and I_2 be a bisection of I, and $\mathbf{p}$ be a monotone distribution over $[n]$. Then,*

$$H_I(\mathbf{p}) \leq w_{\mathbf{p}}(I) \log k - w_{\mathbf{p}}(I_1) \log w_{\mathbf{p}}(I_1) - w_{\mathbf{p}}(I_2) \log w_{\mathbf{p}}(I_2)$$

and

$$H_I(\mathbf{p}) \geq 2w_{\mathbf{p}}(I_2) \log k - w_{\mathbf{p}}(I_2)(\log w_{\mathbf{p}}(I_1) + \log w_{\mathbf{p}}(I_2)).$$

In particular, the ratio of the upper bound to the lower bound on $H_I(\mathbf{p})$ is at most $w_{\mathbf{p}}(I)/2w_{\mathbf{p}}(I_2)$.

Proof. The proof of the upper bound follows from the concavity of $H(\mathbf{p})$. The maximum value is attained when the available weight is distributed uniformly over the elements.

Let $w_1 \stackrel{\text{def}}{=} w_{\mathbf{p}}(I_1)$ and $w_2 \stackrel{\text{def}}{=} w_{\mathbf{p}}(I_2)$. We will prove the lower bound by relaxing the monotonicity condition to a weaker condition: namely, the condition that for $i \leq k$, $p_i \geq w_2/k$, and for $i > k$, $p_i \leq w_1/k$. It is easy to verify that any monotone distribution will satisfy this new constraint. Again, by the concavity of the entropy function, we can plug w_2/k for the elements in I_1 and w_1/k for as many elements as possible in I_2 to give a lower bound on $H_I(\mathbf{p})$. Hence, we get

$$H_I(\mathbf{p}) \geq w_2 \log \frac{k}{w_2} + w_2 \log \frac{k}{w_1}.$$

■

For a balanced leaf corresponding to an interval I with the bisection I_1, I_2, the ratio $w_{\mathbf{p}}(I)/2w_{\mathbf{p}}(I_2)$ can be made small by choosing the parameter β appropriately.

Lemma 22 *Let I be an interval in $[1,n]$ such that $w_{\mathbf{p}}(I) \geq \log^{-3} n$. Then, with probability at least $1 - n^{-1}$, $|S_I|$ will be a β-approximation to $mw_{\mathbf{p}}(I)$.*

Proof. The lemma follows from a straight-forward application of the Chernoff bounds. The random variable X that corresponds to $|S_I|$ is a summation of m independent Bernoulli trials, each with success probability $w_{\mathbf{p}}(I)$. Therefore by the choice of m in the algorithm, the probability that this summation is less than $\mathrm{E}[X]/\beta$ or more than $\beta \mathrm{E}[X]$ is at most $1/n$. Since $\mathrm{E}[X] = mw_{\mathbf{p}}(I)$, the lemma follows. ■

Lemma 23 *Let I be an interval in $[1,n]$ such that $w_{\mathbf{p}}(I) \geq \log^{-3} n$ and I_1, I_2 be a bisection of I. For $\beta > 1$,*

1. *if $w_{\mathbf{p}}(I_1)/w_{\mathbf{p}}(I_2) \geq 2\beta - 1$, then with probability at least $1 - 2n^{-1}$, $|S_{I_1}| \geq \beta \cdot |S_{I_2}|$;*
2. *if $w_{\mathbf{p}}(I_1)/w_{\mathbf{p}}(I_2) \leq (1+\beta)/2$, then with probability at least $1 - 2n^{-1}$, $|S_{I_1}| \leq \beta \cdot |S_{I_2}|$.*

Proof. By Lemma 22, we know that with probability at least $1 - n^{-1}$, $|S_I| \geq mw_{\mathbf{p}}(I)/\beta$. Fix any $t > mw_{\mathbf{p}}(I)/\beta$. We will consider the ratio of the number of samples from I_1 and I_2 conditioned on the event that there are exactly t samples from I. Let Y_i, for $i = 1, \ldots, t$, be an indicator random variable that takes the value 1 if the i-th of these t samples is in I_2, and $Y = \sum_i Y_i$. Therefore, we want to show that the probability that $(t - Y)/Y < \beta$ is at most $1/n$.

The rest of the proof is an application of the Chernoff bounds. Note that $(t-Y)/Y < \beta$ implies $Y > t/(\beta+1)$. Since $\mathrm{E}[Y] \leq t/(2\beta)$, we get

$$\begin{aligned} \Pr\left[Y > \frac{t}{\beta+1}\right] &\leq \Pr\left[Y > \mathrm{E}[Y] + \frac{t(\beta-1)}{2\beta(\beta+1)}\right] \\ &\leq \exp\left(\frac{-t(\beta-1)^2}{\beta^2(\beta+1)^2}\right). \end{aligned}$$

Conditioned on the event that $t \geq mw_{\mathbf{p}}(I)/\beta$, this probability is less than $1/n$. Combining this with Lemma 22, we can conclude that with probability at least $1 - 2n^{-1}$, we have $|S_{I_1}| \geq \beta \cdot |S_{I_2}|$.

Similarly, the second part of the lemma can be proved. ■

By Lemma 23, we can assume that for a "balanced interval" I, the ratio of the weights for the halves is at most $2\beta - 1$ and that two intervals associated with two sibling nodes have weight ratio at least $(1+\beta)/2$.

Let assumption $(*)$ be that no bad event happens, i.e., that (1) for each interval we decide to split, we have that $w_p(I_1)/w_p(I_2) \geq (1+\beta)/2$, (2) for each interval we decide not to split, we have that $w_p(I_1)/w_p(I_2) \leq 2\beta - 1$, and (3) each light leaf has weight at most $\beta^2/\log^3 n$.

Corollary 24 *Assume that (*) holds. For $I \in \mathcal{I}(T_S)$, if $w_{\mathbf{p}}(I) \geq \log^{-3} n$, then $\alpha(I)$ is a β^2-approximation to $H_I(\mathbf{p})$.*

Proof. Let I_1, I_2 be the bisection of I. By the fact that (*) holds, $|S_I|/(m\beta) \leq w_{\mathbf{p}}(I) \leq |S_I|\beta/m$ and $w_{\mathbf{p}}(I_1)/w_{\mathbf{p}}(I_2) \leq 2\beta - 1$. Combining these with Lemma 21, we can conclude that the maximum and minimum possible entropy value for $H_I(\mathbf{p})$ have ratio β. Therefore, $\alpha(I) \leq \beta w_{\mathbf{p}}(I)\log(|I|/2) \leq \beta^2 H_I(\mathbf{p})$.

Suppose I is a long interval. Then since $w_{\mathbf{p}}(I) > 1/\log^3 n$, we have $-\log w_{\mathbf{p}}(I_2)/\log(|I|/2) \leq \beta - 1$. From this, we get $\beta w_{\mathbf{p}}(I)\log(|I|/2) \geq w_{\mathbf{p}}(I)\log(|I|/2) - w_{\mathbf{p}}(I_1)\log w_{\mathbf{p}}(I_1) - w_{\mathbf{p}}(I_2)\log w_{\mathbf{p}}(I_2)$. Thus, we can see that $\alpha(I) \geq (w_{\mathbf{p}}(I)/\beta)\log(|I|/2) \geq H_I(\mathbf{p})/\beta^2$.

Suppose I is a short interval. In this case, we use the lower bound on the entropy in the interval given by Lemma 23, i.e., $H_I(\mathbf{p}) \geq (w_{\mathbf{p}}(I)/2)\log(|I|/2)$. The total entropy contribution of the elements in this interval with probability at most $(\log^{-5} n)/|I|$ is $o((\log^{-4} n)/(\beta - 1))$. By Lemma 3, the entropy of the elements with probability at least $(\log^{-5} n)/|I|$ are estimated well since if we let $\epsilon_0 \leq (\beta - 1)/(2\beta(\beta+2))$, then $(1 - \epsilon_0 - 2\epsilon_0\beta + O(\log^{-1} n))^{-1} \leq \beta^2$ and so the output of Lemma 3 will be a β^2-approximation to $H_I(\mathbf{p})$. ■

Next, we show a bound on the number of nodes in the tree.

Lemma 25 *Assume that (*) holds. Given $\beta > 1$, the number of nodes in T_S is at most*

$$\frac{6\log n \log\log n}{\log(\beta+1) - 1}.$$

Proof. Each node v in the tree corresponds to an interval $I_v = [i,j]$ of probabilities $p_i, \ldots, p_{j-1}$; let $w(v) = p_i + \cdots + p_{j-1}$. For each level h in the tree, let N_h (resp. L_h) denote the number of internal nodes (resp. leaves). The total number of nodes in the tree (whose depth is at most $\log n$) is at most $\sum_{h=1}^{\log n} N_h + L_h$. But, $\sum_h L_h = 1 + \sum_h N_h$. Now, we show an upper bound on N_h.

Fix a level h and let $v_1, \ldots, v_{N_h}$ be the internal nodes in level h ordered by the intervals they define. Note that the intervals are strictly non-overlapping since the nodes are from the same level. If v_i and v_{i+1} are siblings, then by assumption $(*)$, $w(v_i) \geq (1+\beta)w(v_{i+1})/2$. Otherwise, by monotonicity, $w(v_i) \geq w(v_{i+1})$. Since T_S is a full binary tree, every node in level h has a sibling in level h, so for at least half the i's in $[I_h]$, $w(v_i) \geq (1+\beta)w(v_{i+1})/2$. Furthermore, since $1/\log^3 n \leq w(v) \leq 1$,

$$N_h \leq 3\log_{(1+\beta)/2}\log n = \frac{3\log\log n}{(\log(1+\beta)) - 1}.$$

■

Now, we are ready to complete our proof.

Theorem 26 *For $\gamma > 1$, there is an algorithm that approximates the entropy of a monotone distribution on $[n]$ to within a multiplicative factor of γ with probability at least $3/4$ in*

$$O\left(\frac{\gamma^{5/2}\log^{(6\sqrt{\gamma}-3)/(\sqrt{\gamma}-1)} n}{(\sqrt{\gamma}-1)^2(\log(\sqrt{\gamma}+1)-1)}\right)$$

time provided that the entropy of the distribution is $\Omega(\gamma^2/(\log(\sqrt{\gamma}+1)-1))$.

Proof. Assume that (*) holds. By Corollary 24, we can assume that we have a $\gamma\left(1-\frac{\log\log n}{\log n}\right)$-approximation to the entropy of intervals that are associated with balanced leaves and each light leaf has weight at most $\beta^2/\log^3 n$. Since the total weight of the intervals associated with light leaves is at most

$$\frac{3\beta^2\log\log n}{(\log(\beta+1)-1)\log^2 n},$$

their combined entropy contribution is

$$O\left(\frac{\gamma^2\log\log n}{(\log(\sqrt{\gamma}+1)-1)\log n}\right).$$

Let $B = \cup_{I\in\mathcal{I}(T_S)} I$. Since the algorithm's output is a γ-approximation to $H_B(\mathbf{p})$, it is clearly at most $\gamma H(\mathbf{p})$. We can show the other direction as follows.

$$\frac{H_B(\mathbf{p})}{\gamma\left(1-\frac{\log\log n}{\log n}\right)} \geq \frac{H(\mathbf{p})-O\left(\frac{\gamma^2\log\log n}{(\log(\sqrt{\gamma}+1)-1)\log n}\right)}{\gamma\left(1-\frac{\log\log n}{\log n}\right)} \geq \frac{H(\mathbf{p})\left(1-\frac{\log\log n}{\log n}\right)}{\gamma\left(1-\frac{\log\log n}{\log n}\right)} \geq \frac{H(\mathbf{p})}{\gamma}.$$

Since (*) fails to hold true with probability $o(1)$, the error probability of the algorithm is $o(1)$. The running time of the algorithm is sample size times the size of T_S plus the total sample complexity of $\alpha(I)$. ■

Note that the lower bound shown in Theorem 5 applies to monotone distributions. Therefore, a restriction on the entropy such as the one in the statement of Theorem 26 is necessary.

6 Subset-uniform distributions

Consider subset-uniform distributions $\mathcal{E}_k$ which are uniform over some subset $K \subset [n]$ with $|K| = k$. The entropy of this class of distributions is $\log k$. If we approximate k to within a multiplicative factor of γ, then we get a very strong additive approximation to $\log k$. Now, given a black-box distribution which is promised to be from $\mathcal{E}_k$ for some k, the question is to approximate k.

We first approximate k to within a constant factor. Then, we use this approximate value of k together with an algorithm for approximating the L_2-norm of a distribution to improve this constant approximation to any arbitrary factor. First, we show how to approximate k to within a factor of 5.

Lemma 27 *There exists an algorithm in the black-box model that, for a distribution in $\mathcal{E}_k$, outputs ℓ such that $k/5 \leq \ell \leq 5k$ with probability at least $4/5$ in $O(\sqrt{k})$ time.*

Proof. The algorithm is as follows: Sample until you see some element for the second time, say at the t-th sample. Output t^2.

In order to prove the lemma, we need to show that the probability of getting a collision before $\sqrt{k/5}$ samples or not seeing a collision until after $\sqrt{5k}$ samples is less than $1/5$.

$$\Pr\left[\text{No collisions after } t \text{ samples}\right] = \prod_{i=1}^{t}\left(1 - \frac{i-1}{k}\right).$$

For $t < \sqrt{k/5}$,

$$\prod_{i=1}^{t}\left(1 - \frac{i-1}{k}\right) \geq 1 - \frac{1}{k}\sum_{i=0}^{t-1} i \geq 1 - \frac{t^2}{2k} \geq 1 - \frac{1}{10} = 0.9.$$

For $t \geq \sqrt{5k}$,

$$\prod_{i=1}^{t}\left(1 - \frac{i-1}{k}\right) \leq \left(1 - \frac{t}{2k}\right)^t \leq e^{-t^2/2k} \leq e^{-5/2} < 0.1.$$

■

Next, we recall the result from [4] (see also [1]).

Theorem 28 (based on [4]) *Given a black-box distribution P over a domain R, there is an algorithm that uses $O(\sqrt{|R|}/\epsilon^2)$ samples and estimates $\|X\|$ to within a factor of $(1 \pm \epsilon)$ with probability at least $4/5$.*

We put these together to obtain an algorithm that obtains a γ-approximation to k for any $\gamma > 1$.

Theorem 29 *For any $\gamma > 1$, there exists an algorithm in the black-box model that, for a distribution $\mathbf{p} \in \mathcal{E}_k$, outputs ℓ such that $k/\gamma \leq \ell \leq \gamma k$ with probability at least $3/5$ in $O(\sqrt{k}/\gamma^2)$ time.*

Proof. The algorithm has two phases. In the first phase, we use the algorithm in Lemma 27 on $\mathbf{p}$ to obtain k', which is a 5-approximation to k. Now, observe that $\|\mathbf{p}\| = 1/k$. In the second phase, we use the algorithm from Theorem 28 to approximate the two norm of $\mathbf{p}$, viewing $\mathbf{p}$ as a distribution over a domain K' of size k' and with $\epsilon = \gamma - 1$; the important fact about this algorithm is since it is based purely on collision probabilities, it is oblivious to the identity of the domain K' itself. It can be seen that the output of the second phase is actually a γ-approximation to $1/k$, which in turn can be used to obtain a γ-approximation to k. The running time of the second phase is $O(\sqrt{k'}/\gamma^2)$ which is $O(\sqrt{k}/\epsilon^2)$. ■

7 Entropy via collisions

Several earlier work in statistical physics community [7, 8], suggest the use of the collision probability ($\|\cdot\|_2^2$) to estimate the entropy. Using Jensen's inequality, [8] showed that

$$\log \|\mathbf{p}\|^2 = \log \sum_i p_i^2 \geq \sum_i p_i \log p_i = -H(\mathbf{p}).$$

A converse relationship can be shown assuming a bound on the maximum probability of $\mathbf{p}$. If $\|\mathbf{p}\|_\infty \leq n^{-\alpha}$, using the relationship between norms, we get

$$\log \|\mathbf{p}\|^2 \leq \log(|\mathbf{p}|\, \|\mathbf{p}\|_\infty) \leq \log n^{-\alpha} = -\alpha \log n \leq -\alpha H(\mathbf{p}).$$

However, it is unclear how to use this to obtain an arbitrary multiplicative approximation with a better sample complexity than our results.

References

[1] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *Proc. 42nd Annual Symposium on Foundations of Computer Science*, 2001.

[2] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. *Proc. 41st Annual Symposium on Foundations of Computer Science*, pp. 259–269, 2000.

[3] W. Feller. *An Introduction to Probability Theory and Applications, I.* John Wiley & Sons Publishers, 1968.

[4] O. Goldreich and D. Ron. On testing expansion in bounded degree graphs. *ECCC*, TR00-020, 2000.

[5] O. Goldreich and S. Vadhan. Comparing entropies in statistical zero-knowledge with applications to the structure of SZK. *14th IEEE Conf. on Computational Complexity*, pp. 54–73, 1999.

[6] B. Harris. The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis János Bolyai, Topics in Information Theory*, 16:323–355, 1975.

[7] S.-K. Ma. Calculation of entropy from data of motion. *J. of Statistical Physics*, 26(2):221–240, 1981.

[8] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80:197–200, 1998.

[9] D. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. Part I. Bayes estimators and the Shannon entropy. *Physical Review E*, 52(6):6841–6854, 1995.