

Notes on Estimating the Entropy

Consider a distribution \mathbf{p} over the set $[n] = \{1, \dots, n\}$, where the probability of element i is denoted by p_i . Recall that the entropy of the distribution \mathbf{p} is defined as follows:

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

Given access to samples $i \in [n]$ distributed according to \mathbf{p} , we would like to estimate $H(\mathbf{p})$ to within a multiplicative factor γ . That is, we seek an algorithm that obtains an estimate \hat{H} such that $H(\mathbf{p})/\gamma \leq \hat{H} \leq \gamma \cdot H(\mathbf{p})$ with probability at least $2/3$ (here too we can increase the success probability to $1 - \delta$ by running the algorithm $O(\log(1/\delta))$ times and outputting the median).

The main result we shall see today is such an algorithm whose running time is $O\left(n^{\frac{1+\eta}{\gamma^2}} \log n\right)$ conditioned on $H(\mathbf{p}) = \Omega(\gamma/\eta)$. If there is no lower bound on the entropy, then it is impossible to obtain any multiplicative factor (will explain this later), and even if the entropy is quite high (i.e., at least $\log n/\gamma^2$), then $\Omega\left(n^{\frac{1}{2\gamma^2}}\right)$ samples are necessary. For γ that is close to 1 it is possible to obtain a stronger lower bound of: $\Omega\left(n^{\frac{2}{6\gamma^2 - 3 + o(1)}}\right)$ (i.e., roughly $\Omega(n^{2/3})$ for γ close to 1). To make this more concrete, consider the case $\gamma = 2$. Then the upper bound is $\tilde{O}\left(n^{\frac{1+\eta}{4}}\right)$ for $H(\mathbf{p}) = \Omega(1/\eta)$, and in particular, if the entropy is more than constant, we get $\tilde{O}\left(n^{\frac{1+o(1)}{4}}\right)$, while the lower bound is $\Omega\left(n^{\frac{1}{8}}\right)$.

The Lower Bounds

To justify the necessity for a lower bound on $H(\mathbf{p})$, assume there is an algorithm \mathcal{A} that gives a γ approximation and whose running time is $t(n)$. Next consider the following two distributions: in \mathbf{p} the support is all on element 1, i.e., $p_1 = 1$, and $p_i = 0$ for every $i \neq 1$, and in \mathbf{q} all the support is on elements 1 and 2, where $p_1 = 1 - 1/(10t(n))$, $p_2 = 1/(10t(n))$ and $p_i = 0$ for every $i > 2$. Now, $H(\mathbf{p}) = 0$ while $H(\mathbf{q}) > 0$, so for \mathbf{p} the estimate of \mathcal{A} must be 0 (w.h.p), while for \mathbf{q} it must be greater than 0 (w.h.p). But if the running time of the algorithm is $t(n)$, in particular it can view at most $t(n)$ samples. When the distribution is \mathbf{p} it will always see element 1, and when the distribution is \mathbf{q} it will see element 1 in all $t(n)$ sample with probability $(1 - 1/(10t(n)))^{t(n)} \geq 9/10$. This implies that it cannot give a correct answer with sufficiently high probability for at least one of the two distributions.

The lower bound $\Omega\left(n^{\frac{1}{2\gamma^2}}\right)$ is not much more complicated. We give it informally. Let \mathbf{p} be uniform over $[n]$ so that $H(\mathbf{p}) = \log n$, and let \mathbf{q} be uniform over a random subset S of size n^{1/γ^2} , so that $H(\mathbf{q}) = \log(n^{1/\gamma^2}) = \log n/\gamma^2$ and $H(\mathbf{p})/H(\mathbf{q}) = \gamma^2$. If the algorithm takes a sample of size less than $\sqrt{|S|}/c = n^{1/(2\gamma^2)}/c$ for some constant c , then, by the (“negative side” of the) Birthday paradox, for both distributions it won’t see any element repeated. In other words, since S is selected uniformly at random, for both distributions it will just see a sample of different, uniformly selected elements, and hence won’t be able to distinguish between the two distributions.

This lower bound approaches $\Omega\left(n^{\frac{1}{2}}\right)$ and γ gets closer to 1. The other lower bound of $\Omega\left(n^{\frac{2}{6\gamma^2-3+o(1)}}\right)$, which gets close to $\Omega\left(n^{\frac{2}{3}}\right)$ is somewhat more complex. Its first building block is that it can be proved that, without loss of generality, the algorithm makes its decisions only as a function of the number of k -way collisions in the sample for different k . Namely, the algorithm bases its decision on the number of distinct elements, the number of elements that appear twice, the number that appear three times and so on. To get a lower bound of roughly $n^{2/3}$, we define two distributions whose entropies differ by more than $1/\gamma^2$ but such that: (1) The probability of seeing a 3-way collision in $o(n^{2/3})$ samples is very small. (2) the expected number of distinct elements and two-way collisions in the sample is the same. It then needs to be argued that not only the expectations are the same, but that they don’t vary by much from their expectations. The structures of the distributions are simple: one is uniform over a subset of “medium weight” elements. The other has two types of elements: “heavy” and “light”.

The Algorithm

The exact result is stated below.

Theorem 1 *For any $\gamma > 1$ and $0 < \epsilon_0 < 1/2$, there exists an algorithm that can approximate the entropy of a distribution over $[n]$ whose entropy is at least $\frac{4\gamma}{\epsilon_0(1-2\epsilon_0)}$ to within a multiplicative factor of $(1 + 2\epsilon_0)\gamma$ with probability at least $2/3$ in time $O\left(n^{1/\gamma^2} \log n/\epsilon_0^2\right)$.*

The main idea behind the algorithm is the following. Elements in $[n]$ are classified as either *big* or *small* depending on their probability mass. Specifically, for any choice of $\alpha > 0$,

$$B_\alpha(\mathbf{p}) \stackrel{\text{def}}{=} \{i \in [n] : p_i \geq n^{-\alpha}\} \quad (2)$$

The algorithm separately approximates the contribution to the entropy of the big elements and of the small elements, and then combines the two.

In order to describe the algorithm and analyze it, we shall need the following notation. For a distribution \mathbf{p} and a set T ,

$$w_{\mathbf{p}}(T) = \sum_{i \in T} p_i \quad \text{and} \quad H_T(\mathbf{p}) = - \sum_{i \in T} p_i \log(p_i) \quad (3)$$

Note that if T_1, T_2 are disjoint sets such that $T_1 \cup T_2 = [n]$ then $H(\mathbf{p}) = H_{T_1}(\mathbf{p}) + H_{T_2}(\mathbf{p})$.

Algorithm Approximate-Entropy(γ, ϵ_0)

1. Set $\alpha = 1/\gamma^2$.
2. Get $m = \Theta(n^\alpha \log n / \epsilon_0^2)$ samples from \mathbf{p} .
3. Let \mathbf{q} be the empirical probability vector of the n elements. That is, q_i is the number of times i appears in the sample divided by m .
4. Let $\widehat{B}_\alpha = \{i : q_i > (1 - \epsilon_0)n^{-\alpha}\}$.
5. Take an additional sample of size $m = \Theta(n^\alpha \log n / \epsilon_0^2)$ from \mathbf{p} and let $\widehat{w}(S)$ be the total empirical weight of elements in $S = [n] \setminus \widehat{B}_\alpha$ in the sample.
6. Output $H_{\widehat{B}_\alpha}(\mathbf{q}) + \frac{\widehat{w}(S) \log n}{\gamma}$.

In the next two subsections we analyze separately the contribution of the big elements and the contribution of the small elements.

Approximating the contribution of big elements

Lemma 2 For $m = 20n^\alpha \log n / \epsilon_0^2$ and \mathbf{q} as defined in the algorithm, with probability at least $1 - \frac{1}{n}$ the following two conditions hold for every $i \in [n]$:

1. If $p_i \geq \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$ (in particular this is true of $i \in B_\alpha(\mathbf{p})$) then $|p_i - q_i| \leq \epsilon_0 p_i$;
2. If $p_i < \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$ then $q_i < (1 - \epsilon_0)n^{-\alpha}$.

Proof: Recall the multiplicative Chernoff bound (stated here for 0/1 random variables): Let χ_1, \dots, χ_m be m independent 0/1 (Bernoulli) random variables, where $\Pr[\chi_j = 1] = \mu$ (so that $\text{Exp}[\chi_j] = \mu$ as well). Then, for every $\epsilon \in (0, 1]$, the following bounds hold:

$$\Pr \left[\frac{1}{m} \cdot \sum_{j=1}^m \chi_j > (1 + \epsilon)\mu \right] < \exp(-\epsilon^2 \mu m / 3)$$

and

$$\Pr \left[\frac{1}{m} \cdot \sum_{i=1}^m \chi_j < (1 - \epsilon)\mu \right] < \exp(-\epsilon^2 \mu m / 2)$$

For each element i we can define m random variables χ_1, \dots, χ_m where $\chi_j = 1$ if and only if the j 'th element in the sample is i . Therefore, $\mu = \Pr[\chi_j = 1] = p_i$ for every j . When $p_i \geq \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$ we have that $\mu \geq \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$ which is at least $\frac{1}{3}n^{-\alpha}$ since $\epsilon_0 < 1/2$. If we set $\epsilon = \epsilon_0$,

$$\Pr[|q_i - p_i| > \epsilon p_i] < \exp(-\epsilon_0^2 n^{-\alpha} m / 9) + \exp(-\epsilon_0^2 n^{-\alpha} m / 6) < \frac{1}{n^2} \quad (4)$$

On the other hand, if $p_i < \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$, then the probability that $q_i \geq (1 - \epsilon_0)n^{-\alpha}$ is upper bounded by the probability of such an event when $p_i = \frac{1-\epsilon_0}{1+\epsilon_0}n^{-\alpha}$, which we have already shown is at most $\frac{1}{n^2}$. The lemma follows by taking a union bound over all i . ■

By Lemma 2, we get that $B_\alpha(\mathbf{p}) \subseteq \widehat{B}_\alpha$, and for every $i \in \widehat{B}_\alpha$ (even if $i \notin B_\alpha(\mathbf{p})$), $|q_i - p_i| \leq \epsilon_0 p_i$. The next lemma bounds the deviation of $H_T(\mathbf{q})$ from $H_T(\mathbf{p})$ condition on q_i being close to p_i for every $i \in T$.

Lemma 3 For every set T such that for every $i \in T$, $|q_i - p_i| \leq \epsilon_0 p_i$,

$$|H_T(\mathbf{q}) - H_T(\mathbf{p})| \leq \epsilon_0 H_T(\mathbf{p}) + 2\epsilon_0 w_{\mathbf{p}}(T) \quad (5)$$

Proof: For $i \in T$, let ϵ_i be defined by $q_i = (1 + \epsilon_i)p_i$ where by the premise of the lemma, $|\epsilon_i| \leq \epsilon_0$.

$$H_T(\mathbf{q}) - H_T(\mathbf{p}) = - \sum_{i \in T} (1 + \epsilon_i) p_i \log((1 + \epsilon_i) p_i) + \sum_{i \in T} p_i \log p_i \quad (6)$$

$$= - \sum_{i \in T} (1 + \epsilon_i) p_i \log p_i - \sum_{i \in T} (1 + \epsilon_i) p_i \log(1 + \epsilon_i) + \sum_{i \in T} p_i \log p_i \quad (7)$$

$$= \sum_{i \in T} \epsilon_i p_i \log(1/p_i) - \sum_{i \in T} (1 + \epsilon_i) p_i \log(1 + \epsilon_i) \quad (8)$$

If we now consider the absolute value of this difference:

$$|H_T(\mathbf{q}) - H_T(\mathbf{p})| \leq \left| \sum_{i \in T} \epsilon_i p_i \log(1/p_i) \right| + \left| \sum_{i \in T} (1 + \epsilon_i) p_i \log(1 + \epsilon_i) \right| \quad (9)$$

$$\leq \sum_{i \in T} |\epsilon_i| p_i \log(1/p_i) + \sum_{i \in T} (1 + \epsilon_i) p_i \log(1 + \epsilon_i) \quad (10)$$

$$\leq \epsilon_0 H_T(\mathbf{p}) + 2\epsilon_0 w_T(\mathbf{p}) \quad (11)$$

■

Approximating the contribution of small elements

Recall that $S = [n] \setminus \widehat{B}_\alpha$ so that, By Lemma 2, with high probability $S \subseteq [n] \setminus B_\alpha(\mathbf{p})$.

Claim 4 Let \hat{w} be the fraction of samples, among $m = \Theta((n^\alpha/\epsilon_0^2) \log n)$ that belong to S . If $w_{\mathbf{p}}(S) \geq n^{-\alpha}$ then Then

$$(1 - \epsilon_0) w_{\mathbf{p}}(S) \leq \hat{w} \leq (1 + \epsilon_0) w_{\mathbf{p}}(S) \quad (12)$$

with probability $1 - 1/n$.

The claim directly follows by a multiplicative Chernoff bound.

Lemma 5 If $p_i \leq n^{-\alpha}$ for every $i \in S$ then

$$\alpha \cdot \log n \cdot w_{\mathbf{p}}(S) \leq H_S(\mathbf{p}) \leq \log n \cdot w_{\mathbf{p}}(S) + 1 \quad (13)$$

Proof: Conditioned on a particular weight $w_{\mathbf{p}}(S)$, the entropy $H_S(\mathbf{p})$ is maximized when $p_i = w_{\mathbf{p}}(S)/|S|$ for all i . In this case

$$H_S(\mathbf{p}) = w_{\mathbf{p}}(S) \log(|S|/w_{\mathbf{p}}(S)) \quad (14)$$

$$= w_{\mathbf{p}}(S) \log |S| + w_{\mathbf{p}}(S) \log(1/w_{\mathbf{p}}(S)) \quad (15)$$

$$\leq w_{\mathbf{p}}(S) \log n + 1 \quad (16)$$

On the other hand, $H_S(\mathbf{p})$ is minimized when its support is minimized. Since $p_i \leq n^{-\alpha}$ for every $i \in S$, this means that $n^\alpha w_{\mathbf{p}}(S)$ of the elements have the maximum probability $p_i = n^{-\alpha}$, and all others have 0 probability. In this case $H_S(\mathbf{p}) = \alpha w_{\mathbf{p}}(S) \log n$. ■

Putting it together

We now prove Theorem 1 based on Algorithm Approximate-Entropy. By lemma 2 we have that with high probability:

1. If $i \in B_\alpha(\mathbf{p})$ then $i \in \widehat{B}_\alpha$. That is, For ever $i \in S = [n] \setminus \widehat{B}_\alpha$, $p_i \leq n^{-\alpha}$.
2. Every $i \in \widehat{B}_\alpha$ satisfies $|q_i - p_i| \leq \epsilon_0 p_i$.

Assume from this point on that the above two properties hold. Let B be a shorthand for \widehat{B}_α and let $S = [n] \setminus B$ as defined in the algorithm. Assume first that $w_{\mathbf{p}}(S) \geq n^{-\alpha}$. In this case, Lemma 5 tells us that (since $\alpha = 1/\gamma^2$)

$$\frac{1}{\gamma^2} \cdot w_{\mathbf{p}}(S) \leq H_S(\mathbf{p}) \leq w_{\mathbf{p}}(S) \log n + 1 \quad (17)$$

or equivalently:

$$\frac{1}{\log n} \cdot (H_S(\mathbf{p}) - 1) \leq w_{\mathbf{p}}(S) \leq \frac{1}{\log n} \cdot \gamma^2 \cdot H_S(\mathbf{p}) \quad (18)$$

By Claim 4,

$$(1 - \epsilon_0)w_{\mathbf{p}}(S) \leq \hat{w}(S) \leq (1 + \epsilon_0)w_{\mathbf{p}}(S) \quad (19)$$

If we now use the above two equations and apply Lemma 3 (using $|q_i - p_i| \leq \epsilon_0 p_i$ for every $i \in B$), we get:

$$H_B(\mathbf{q}) + \frac{\hat{w}(S) \log n}{\gamma} \leq (1 + \epsilon_0)H_B(\mathbf{p}) + 2\epsilon_0 + \frac{(1 + \epsilon_0)w_{\mathbf{p}}(S) \log n}{\gamma} \quad (20)$$

$$\leq (1 + \epsilon_0) \cdot (H_B(\mathbf{p}) + \gamma H_S(\mathbf{p})) + 2\epsilon_0 \quad (21)$$

$$\leq (1 + \epsilon_0)\gamma H(\mathbf{p}) + 2\epsilon_0 \quad (22)$$

$$\leq (1 + 2\epsilon_0)\gamma H(\mathbf{p}) \quad (23)$$

where in the last inequality we used our lower bound on $H(\mathbf{p})$: $H(\mathbf{p}) \geq \frac{4\gamma}{\epsilon_0(1-2\epsilon_0)} > 2/\gamma$ so that $2\epsilon_0 < \epsilon_0 \cdot \gamma \cdot H(\mathbf{p})$. Similarly,

$$H_B(\mathbf{q}) + \frac{w_q(S) \log n}{\gamma} \geq (1 - \epsilon_0)H_B(\mathbf{p}) - 2\epsilon_0 + \frac{(1 - \epsilon_0)w_{\mathbf{p}}(S) \log n}{\gamma} \quad (24)$$

$$\geq (1 - \epsilon_0) \cdot \left(H_B(\mathbf{p}) + \frac{H_S(\mathbf{p}) - 1}{\gamma} \right) - 2\epsilon_0 \quad (25)$$

$$\geq \frac{H(\mathbf{p})}{\gamma(1 + 2\epsilon_0)} \quad (26)$$

(the last inequality follows from the lower bound on $H(\mathbf{p})$ by tedious by elementary manipulations). Finally, if $w_{\mathbf{p}}(S) < n^{-\alpha}$ then by Claim 4 $w_{\mathbf{q}}(S) \leq (1 + \epsilon_0)n^{-\alpha}$ with high probability. Therefore, $w_q(S) \log n/\gamma$ is at most $(1 + \epsilon_0)n^{-\alpha} \log n/\gamma$ (and at least 0). It is not hard to verify that the contribution to the error is negligible, assuming γ is bounded away from 1.

Special Cases and Alternative Models

MONOTONE DISTRIBUTIONS We say that \mathbf{p} is monotone if $p_i \geq p_{i+1}$ for all i . For this case there is a more sophisticated algorithm that uses $\text{poly}(\log n, \log \gamma)$ samples.

THE EVALUATION ORACLE MODEL. In this model it is assumed that the algorithm can ask, for any i of its choice, what is p_i . In general there is an $\Omega(n)$ lower bound (think of one distribution that has all its support on one element, and the other has on two elements (e.g., $1/2$ and $1/2$)). The upper bound on estimating (actually, computing exactly) the entropy in this model is of course trivially $O(n)$.

If $H(\mathbf{p}) \geq h$ then there is a lower bound of $\Omega(n2^{-\gamma^2(h+1)})$ (that “spreads” the weight a bit more in both distributions), but no known upper bound (other than the trivial $O(n)$). Here too monotonicity makes life easier: there is an $O(\log n \log(1/\gamma))$ algorithm. Finally, if one combines the two models, then in general there is a lower bound of $\Omega\left(n^{(1-o(1))/\gamma^2}\right)$, and when $H(\mathbf{p}) \geq h$, can show $\Omega\left(\frac{\log n}{h(\gamma^2-1)+\gamma^2}\right)$ and $O\left(\frac{\gamma^2 \log n}{h^2(\gamma-1)^2}\right)$. The idea of the lower bound is simple: use the sampling to get select i 's, and then query p_i for each. The output is the average over $\log(1/p_i)$ for those i 's sampled.