Electronic Colloquium on Computational Complexity, Report No. 84 (2005)

Approximating the Entropy of Large Alphabets

Mickey Brautbar * Alex Samorodnitsky *

Abstract

We consider the problem of approximating the entropy of a discrete distribution P on a domain of size q, given access to n independent samples from the distribution. It is known that $n \geq q$ is necessary, in general for a good additive estimate of the entropy. A problem of multiplicative entropy estimate was recently addressed by Batu, Dasgupta, Kumar, and Rubinfeld. They show that $n = q^{\alpha}$ suffices for a factor-α approximation, $\alpha < 1$.

We introduce a new parameter of a distribution - its *effective alphabet size* $q_{ef}(P)$. This is a more intrinsic property of the distribution depending only on its entropy moments. We show $q_{ef} \leq \tilde{O}(q)$. When the distribution P is essentially concentrated on a small part of the domain $q_{ef} \ll q$. We strengthen the result of Batu et al. by showing it holds with q_{ef} replacing q.

This has several implications. In particular the rate of convergence of the maximum-likelihood entropy estimator (the empirical entropy) for both finite and infinite alphabets is shown to be dictated by the effective alphabet size of the distribution. Several new, and some known, facts about this estimator follow easily.

Our main result is algorithmic. Though the effective alphabet size is, in general, an unknown parameter of the distribution, we give an efficient procedure (with access to the alphabet size only) that achieves a factor-α approximation of the entropy with $n = \widetilde{O}\left(\exp\left\{\alpha^{1/4} \cdot \log^{3/4} q \cdot \log^{1/4} q_{ef}\right\}\right)$. Assuming (for instance) $\log q_{ef} \ll \log q$ this is smaller than any power of q. Taking $\alpha \to 1$ leads in this case to efficient *additive* estimates for the entropy as well.

Several extensions of the results above are discussed.

*School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel.

ISSN 1433-8092

1 Introduction

1.1 Background

Stochastic sources and entropy

Let X be a random variable with values in a finite or countable alphabet $\mathcal{A}$. For $a \in \mathcal{A}$, let p_a be the probability of the event $X = a$. The number $H(X) = \sum_{a \in \mathcal{A}} p_a \log \frac{1}{p_a}$ is called the Shannon entropy of X. [1]

A stochastic, or random, source $\mathcal{X}$ is a sequence of random variables $X_1, ...X_n, ...$ defined on a common alphabet $\mathcal{A}$. The entropy of a random source $\mathcal{X}$ is defined as $\lim_{n\to\infty} \frac{1}{n} H(X_1...X_n)$, if the limit exists. Here $(X_1...X_n)$ is a random variable with values in $\mathcal{A}^n$, distributed according to the joint distribution of $X_1, X_2, ..., X_n$. The entropy of a random source is well-defined for a large family of random sources - stationary and ergodic sources. In this paper we deal only with the simplest representative of this family, for which all the variables X_i are independent and identically distributed. In this case, clearly, $H(\mathcal{X}) = H(X_1)$. Such a source is often called a discrete memoryless source.

The setting

We are considering the black-box scenario in which a random source is given as a sample $(a_1, ..., a_n) \in \mathcal{A}^n$ produced according to the (unknown) joint distribution of $X_1...X_n$. The goal is to estimate the entropy of the source. Efficiency in this model means viewing the smallest possible sample. Appropriate notions of quality of approximation will be defined in the next subsection.

Motivation

Random sources model a large number of phenomena in natural and life sciences. In many cases the exact mechanism behind a specific source is not well-understood and therefore a black-box scenario is appropriate. The entropy of a source is frequently an important parameter, however computing it exactly might be infeasible [2, 10, 13]. The realistic goal then is to try and approximate it efficiently as well as possible.

Example 1.1: The following is the layman's view of [9]. A visual stimulus is applied to a blowfly. This generates a neural spike train, i.e. a binary string of length $k \approx 30$. After a short time interval (a second) a new stimulus is applied, and a new string is generated, and so on. This produces a sequence of random variables on a common alphabet of neural responses (binary strings of length k). These random variables are presumed to be identically distributed and independent if the time intervals between stimuli are long enough. The entropy of this random source is an important measure for the complexity of the fly's response to environment. ∎

1.2 Approximation of entropy

There are two natural notions of approximating an unknown quantity $H(\mathcal{X})$. In both an efficiently computable functional $\hat{H}$ of the sample $\bar{a} = a_1...a_n$ is constructed. $\hat{H}$ provides an additive approximation of H within a constant c if $\hat{H} \leq H \leq \hat{H} + c$ for most samples $\bar{a}$. $\hat{H}$ gives a multiplicative approximation of H within factor c if $\hat{H} \leq H \leq c \cdot \hat{H}$ for most samples $\bar{a}$.

Example 1.2: A sample $\bar{a} = a_1, ..., a_n$ defines an empirical distribution $\hat{p}_a = \frac{\left|\{i : a_i = a\}\right|}{n}$ on the alphabet $\mathcal{A}$. The entropy of this distribution is a natural estimator for H. We denote it by H_{MLE} and refer to it as the maximum-likelihood entropy estimator. ∎

[1] All the logarithms in this paper are to base 2, unless stated otherwise.

Most of the research in the area concentrated on additive approximation. The two important parameters are the alphabet size q and the sample size n. It is convenient to present some of the results for memoryless sources, classifying them according to relative sizes of q and n. Most of the results here are for the maximum-likelihood estimator. We mention that there are many other entropy estimators. Some of them work better for more general sources. Others arise naturally in the context of data compression. Our discussion is by necessity brief (see [6],[10], [13] for more).

1. q is fixed and $n \to \infty$. The maximum-likelihood estimator H_{MLE} converges to the entropy H with asymptotic rate of $O\left(\frac{1}{\sqrt{n}}\right)$. In other words $|H_{MLE} - H| \leq O\left(\frac{1}{\sqrt{n}}\right)$ with probability tending to 1 as $n \to \infty$.

2. q grows with n, but $n \gg q$. The maximum-likelihood estimator H_{MLE} gives an $o(1)$ additive approximation of the entropy. Namely for any $\epsilon > 0$ the probability of $|H_{MLE} - H| \leq \epsilon$ goes to 1 as n goes to infinity.

3. q grows at the same rate as n. In this case, the maximum-likelihood estimator H_{MLE} fails to achieve a vanishing additive error. In fact it is known to have a constant negative bias. However estimators with vanishing additive error are known to exist, though so far the proofs of this fact are existential [11].

4. $n = q^\alpha$ for $0 < \alpha < 1$. In this case there are no consistent additive estimators for the entropy [10]. Batu, Dasgupta, Kumar, and Rubinfeld [2] construct a *multiplicative* estimator achieving (essentially) an α-approximation. Let us quote [2] the relevant result of [2] here, since we will need it later on.

   Theorem 1.3: *[2] Let $0 < \alpha < 1$. Let $\bar{a} = a_1 ... a_n$ be a sample from a discrete memoryless source with distribution P. Let $\hat{p}$ be the empirical distribution defined by the sample. Let q be the alphabet size, and $t = q^{-\alpha}$. Define an entropy estimator $\hat{H} = \sum_{a:\ \hat{p}_a \geq t} \hat{p}_a \log \frac{1}{\hat{p}_a} + \log \frac{1}{t} \cdot \sum_{a:\ \hat{p}_a < t} \hat{p}_a$. Then, assuming $n \geq \widetilde{\Omega}\left(q^\alpha\right)$*

   $$\hat{H} \leq H(P) \leq \frac{1}{\alpha}\hat{H},$$

   with high probability over the samples $\bar{a}$. (The $\widetilde{\Omega}$ notation hides poly-logarithmic in q factors.)

5. $q = \infty$, and $n \to \infty$. This case is treated in [1], [13]. Assuming the first two entropy moments (see below) of the distribution are bounded, the maximum-likelihood estimator H_{MLE} converges to the entropy H with asymptotic rate of $O\left(\frac{1}{\log n}\right)$. This is best possible for any sequence of universal entropy estimators.

In all these cases, the sample size needed to obtain a required entropy approximation is taken to be a function of the alphabet size q. It seems reasonable to ask which additional parameters of the distribution bear on the required sample size. As an extremal example consider a distribution on a large (even infinite) alphabet, which is very concentrated, so that its entropy is small.

Example 1.4: The geometric distribution. Let $0 < p < 1$. Consider an integer-valued random variable X with $Pr(X = k) = p(1-p)^{k-1}$, for $k \geq 1$. The entropy of X is $\frac{H(p)}{p}$. (Here $H(x) = -x \log x - (1-x)\log(1-x)$ is the binary entropy function.) Setting $p = \frac{1}{4}$ yields $H(X) \approx 3.25$. This entropy is attained within an additive error of $\frac{1}{1000}$ by considering just the 28 most frequent symbols of the distribution. ∎

[2] This is a streamlined version of theorem 1 in [2].

1.3 Our results

Our point of view is that of sub-linear algorithms. We are working in the same regime as [2] assuming the alphabet size to be too large to allow dealing with linear-size samples. It turns out however that the alphabet size is not necessarily the most appropriate measure for the difficulty of the problem, and it is useful to consider other parameters of the distribution.

We introduce the notion of *effective alphabet size* of a distribution P. Its relevance will be demonstrated by the results below (theorems 1.9 and 2.3). Here we offer a brief informal discussion of this notion, relating it to the *asymptotic equipartition property* [3] for memoryless sources. Let $\mathcal{X} = (X_i)$ be a memoryless source. Let $P = \{p_a\}_{a\in\mathcal{A}}$ be the distribution of X_i. Consider a random variable $Y = \log\frac{1}{Pr(X_1)}$ taking value $\log\frac{1}{p_a}$ with probability p_a. We need to estimate the expectation of Y, which is the entropy of P. For this purpose we would like to define a random variable with the same expectation and a small variance. Let $H_2(P) = \sum_{a\in\mathcal{A}} p_a \log^2\frac{1}{p_a}$ be the second moment of Y. For a length-t sequence $X_i...X_{i+t-1}$, the distribution of the logarithm of inverse probability $\log\frac{1}{Pr(X_i...X_{i+t-1})}$ is that of a sum of t independent copies of Y. The Chebyshev inequality for $\frac{1}{t}\log\frac{1}{Pr(X_i...X_{i+t-1})}$ implies that for t of order $H_2(P)/H^2(P)$, a typical value of this random variable is close to $H(P)$. In other words, a typical length-t sample of the source has probability close to $2^{-tH(P)} \approx 2^{-\frac{H_2(P)}{H(P)}}$. We note that this is precisely the asymptotic equipartition property for the source. In order to sample this random variable with a reasonable accuracy (i.e. to estimate the probability of a a typical length-t sample) we need to look at as many as $2^{\frac{H_2(P)}{H(P)}}$ typical sequences. Consider typical length-t sequences as symbols in a new alphabet of size $2^{\frac{H_2(P)}{H(P)}}$. This is the *typical subset* of $\mathcal{A}^t$. The typical subset provides a comprehensive description of the source (e.g. for the sake of data compression [3]). Here we show its relevance to entropy approximation.

Definition 1.5: The effective alphabet size of a distribution P is $q_{ef}(P) = 2^{\frac{H_2(P)}{H(P)}}$. ∎

Example 1.6: Let P be a uniform distribution on alphabet of size q. Then $q_{ef}(P) = q$. ∎

The effective size of an alphabet is (essentially) majorized by its real size.

Proposition 1.7: *Let P be a distribution on alphabet of size q, and assume* [3] *the entropy of P is at least 1. Then $q_{ef}(P) \leq q\log^2(q)$.* [4]

On the other hand, the effective size of an alphabet could be much smaller than its real size.

Example 1.8: Let P be a geometric distribution with a parameter p. Then

$$q_{ef}(P) = 2^{\frac{H^2(p)+(1-p)\log^2(1-p)}{pH(p)}}$$

Taking $p = \frac{1}{4}$ gives $q_{ef}(P) \approx 16$, while $q = \infty$. ∎

The following claim is a special case of our main technical result theorem 2.3. It illustrates the relevance of the notion of effective size of the alphabet.

[3] The assumption that the entropy of the sampled distribution is bounded away from zero in necessary, even if we want only to distinguish between this distribution and a 1-point distribution. We will assume the entropy to be at least 1. This choice affects only constants in the following discussion.

[4] It is not hard to construct distributions on q points with $q_{ef}(P) \geq \Omega\left(q\log^c(q)\right)$, for some $c > 0$.

Theorem 1.9: *Let $0 < \alpha \leq \frac{1}{2}$. Let $\bar{a} = a_1 ... a_n$ be a sample from a discrete memoryless source with distribution P. Let $\hat{p}$ be the empirical distribution defined by the sample. Let $q_e = q_{ef}(P)$, and $t = q_e^{-\alpha}$. Define an entropy estimator $\hat{H} = \sum_{a:\ \hat{p}_a \geq t} \hat{p}_a \log \frac{1}{\hat{p}_a} + \log \frac{1}{t} \cdot \sum_{a:\ \hat{p}_a < t} \hat{p}_a$. Then, assuming $n \geq \widetilde{\Omega}\left(q_e^{\alpha}\right)$*

$$\hat{H} \leq H(P) \leq \frac{1}{\alpha} \hat{H},$$

with high probability over the samples $\bar{a}$. (The $\widetilde{\Omega}$ notations hides polylogarithmic in q_e factors.)

This result should be compared to theorem 1.3. In fact our investigation began as an attempt to understand theorem 1.3.

We replace the alphabet size with the effective alphabet size. By proposition 1.7 this makes theorem 1.9 at least as strong as theorem 1.3 for $\alpha < \frac{1}{2}$. If $q_{ef} \ll q$ theorem 1.9 is a strengthening of theorem 1.3.

For larger $\alpha < 1$ the sample size has to grow or more stringent assumptions on the distribution are required. In [2] for $\alpha \to 1$ the assumption $H(P) \geq \Omega\left(\frac{1}{1-\alpha}\right)$ is made. Here we deal with this problem by increasing the sample size. We will define an increasing sequence of "effective alphabet sizes" $q_{ef}^{(k)}(P)$ so that $q_{ef}(P) = q_{ef}^{(2)}(P)$, and such that if we replace $q_{ef}(P)$ with $q_{ef}^{(k)}(P)$ the theorem holds for a larger range of approximation factors $0 < \alpha \leq \frac{k-1}{k}$. On the other hand, for k as high as $O(\log q)$, the k-th effective alphabet size continues to be majorized by the alphabet size. Thus the claim of theorem 1.3 is essentially recovered, without requiring the entropy to grow.

This allows us to connect two relative scales of sample size versus alphabet size. So far we have considered sample sizes sublinear in the alphabet size. Taking the sample size to be linear in the alphabet size, we can take the approximation factor α close to 1, as close as $\alpha = 1 - \widetilde{O}\left(\frac{1}{\log q}\right)$. Therefore, given a linear number of samples, the functional $\hat{H}$ in theorem 1.9 becomes a very good multiplicative estimator for the entropy.

We now turn to the maximum-likelihood entropy estimator H_{MLE} defined in example 1.2. This estimator calculates the entropy of the empirical distribution defined by the sample. H_{MLE} is a natural and a thoroughly investigated functional. Our methods allow us to prove new properties of H_{MLE}. The key step is the simple observation that the two estimators $\hat{H}$ (in theorems 1.9 and 2.3) and H_{MLE} are comparable. In fact it is always true that $\hat{H} \leq H_{MLE}$. On the other hand, it is known (see section 3) that H_{MLE} is smaller than the entropy (with high probability). This means that H_{MLE} lies between $\hat{H}$ and $H(P)$, and everything we prove for $\hat{H}$ holds for H_{MLE}. [5]

Corollary 1.10: *The convergence rate of the maximum-likelihood estimator is dictated by the effective alphabet size of the distribution.*

The discussion so far provides a new glimpse into behaviour of the maximum-likelihood estimator for sample sizes linear in the alphabet size. We believe this is important enough to warrant stating it as a proposition.

Proposition 1.11: *Let P be a distribution on alphabet of size q, and let $\bar{a} = a_1 ... a_n$ be a sample from a discrete memoryless source with distribution P. Then, assuming $n \geq \Omega(q)$,*

$$H_{MLE} \leq H(P) \leq \left(1 + \widetilde{O}\left(\frac{1}{\log q}\right)\right) \cdot H_{MLE},$$

with high probability over the samples $\bar{a}$.

[5] The only advantage of $\hat{H}$ is that it seems to be easier to analyse.

It is a well-established fact that H_{MLE} is not a good *additive* estimator when the sample size is linear in the alphabet size. In this case it has a constant negative bias [10]. In fact several variations of this estimator [4, 5, 8, 12] were constructed to resolve this problem. However none of them seems to work for all distributions. Here we suggest another point of view: when the sample size decreases from super-linear to linear in the alphabet size, the maximum-likelihood estimator transforms from a good additive entropy estimator to a good *multiplicative* estimator. So good, in fact, that if the entropy is not very large ($H(P) \ll \widetilde{O}(\log q)$), it will be approximated within $o(1)$ additive error.

Replacing the alphabet size with an effective alphabet size allows us to consider sample sizes which are linear or even superlinear in the effective alphabet size even if we wish to keep the costs sublinear in the alphabet size. An extremal example is a distribution on an infinite alphabet with bounded first and the second entropy moments. Taking the sample size $n = q_{ef}^{\alpha}(P)$, and allowing α to grow to infinity, leads to the following result.

Proposition 1.12:

$$|H(P) - H_{MLE}| \leq \frac{H_2(P)}{2\log n - O(\log\log n)} + O\left(\frac{\log n}{\sqrt{n}}\right)$$

with probability tending to 1 *as* $n \to \infty$.

This recovers a known result of [13] with a better constant. Replacing $q_{ef}(P)$ with k-th effective alphabet size implies that for distributions with a bounded k-th entropy moment the maximum likelihood estimator has a convergence rate of $O\left(\frac{1}{k\log^{k-1} n}\right)$. This, in turn, implies polynomial convergence rate for distributions all of whose moments are bounded and do not grow too fast.

Returning to theorem 1.9 we point out a difficulty in using this theorem algorithmically. The definition of the effective alphabet size depends on the unknown parameters we actually want to estimate. This however suggests a sequential approach.

Let the alphabet size q be given. The first step will use an appropriate generalization of theorem 1.3 with a small approximation factor to obtain crude estimates of the first two entropy moments of the distribution, and thus a crude upper bound on its effective alphabet size. If the distribution is properly concentrated even this bound will be significantly smaller than q. The second step uses theorem 1.9 with this bound substituted for $q_{ef}(P)$ and with a *larger* approximation factor. (Of course the number of steps could be larger than two.) Theorem 4.1 gives sustenance to this plan by upgrading theorem 2.3 to estimate higher entropy moments.

This discussion leads to the main results of the paper. We present them in a special case matching theorem 1.9. They are easily generalized using theorem 2.3 and theorem 4.1 in full generality.

Theorem 1.13: *Let* $0 < \beta < \alpha \leq \frac{1}{2}$. *Let* $\bar{a} = a_1 ... a_n$ *be a sample from a discrete memoryless source with distribution* P. *Let* q *be the alphabet size, and* q_{ef} *the effective alphabet size.*

There is an efficient procedure that, given access to α, β, *and* q *only, computes a functional* $\hat{H}$ *of the sample such that assuming* $n \geq \widetilde{\Omega}\left(q^{\beta} + q_{ef}^{\frac{\alpha}{\beta^3}}\right)$

$$\hat{H} \leq H(P) \leq \frac{1}{\alpha}\hat{H},$$

with high probability over the samples $\bar{a}$. *(The* $\widetilde{\Omega}$ *notations hides polylogarithmic in* q *factors.)*

Let us fix the value of α and view β as a variable. The theorem holds for any constant $\beta < \alpha$. In fact, it is not hard to check that we can let β be a function of q. A natural idea then is to optimize over β, using binary search, say. [6] This leads to the following corollary.

[6] We are glossing over technical details.

Corollary 1.14: *Let $0 < \alpha \leq \frac{1}{2}$. Let $\bar{a} = a_1...a_n$ be a sample from a discrete memoryless source with distribution P. Let q be the alphabet size, and q_{ef} the effective alphabet size. There is an efficiently computable functional $\hat{H}$ of the sample, given access to α and q only, such that assuming $n \geq \widetilde{\Omega}\left(\exp\left\{\alpha^{1/4} \cdot \log^{3/4} q \cdot \log^{1/4} q_{ef}\right\}\right)$*

$$\hat{H} \leq H(P) \leq \frac{1}{\alpha}\hat{H},$$

with high probability over the samples $\bar{a}$.

Example 1.15: Let P be a geometric distribution with a parameter p. Turn P into a distribution P_q on q points by retaining only the first q most frequent symbols, and assigning the remaining probability mass to the last symbol. Let q be chosen so that $\frac{1}{\log q} \leq p \leq 1 - \frac{1}{\log q}$. Then $H(P_q) \geq 1$, and the effective alphabet size q_{ef} is bounded from above by $\log^3 q$. Therefore a $\frac{1}{2}$-approximation of $H(P_q)$ can be achieved using a sample of size at most $\exp\left\{\widetilde{O}\left((\log q)^{3/4}\right)\right\}$. ∎

Taking $\alpha \to 1$ and using general versions of theorem 1.13 and corollary 1.14 we obtain tighter multiplicative (and in some cases even additive) estimates for the entropy given sample of size $n = \exp\left\{\widetilde{\Omega}\left(\log^{3/4} q \cdot \log^{1/4} q_{ef}\right)\right\}$.

A word on our methods. Our proofs are similar to those of [2], with an occasional complication arising from the fact that we do not restrict the alphabet size. This means we are essentially dealing with distributions on infinite alphabets. Estimating higher entropy moments requires handling additional technicalities.

1.4 Discussion

Consider a a random source modelling a certain complex (natural world) system. The alphabet of this source describes a short-term behaviour of the system. It might be very large, since there are many possible patterns of short-term behaviour. However, one would expect a certain *typical behaviour* to emerge, in the sense that a very small subset of the alphabet would support most of the probability mass and in fact most of the entropy of the source distribution. These would be the patterns of short-term behaviour following the internal logic of the system. The huge majority of the remaining symbols would be, in this simplistic view, a noise stemming from external factors. If such is the case, the typical subset of the alphabet becomes the 'effective alphabet' of the distribution.

In particular the size of the typical set rather than the size of the entire alphabet should be the parameter of interest for the sake of entropy approximation. The notion of the effective alphabet size of a distribution attempts to capture this point of view. We have argued for its conceptual and algorithmic relevance.

2 Effective alphabet size and entropy approximation

In this section we define a sequence of "effective alphabet sizes" $q_{ef}^{(k)}(P)$ of a distribution P and prove our main approximation result: it is possible to approximate the entropy within a multiplicative factor $0 < \alpha < \frac{k-1}{k}$, given a sample of size $n \geq \widetilde{O}\left(q_e^{\alpha}\right)$, where $q_e = q_{ef}^{(k)}$.

First we define higher entropy moments.

Definition 2.1: For an integer $k \geq 1$ the k-th entropy moment of a distribution P is $H_k(P) = \sum_{a \in \mathcal{A}} p_a \log^k \frac{1}{p_a}$. ∎

Definition 2.2: For an integer $k \geq 2$, the k-th effective alphabet size of P is $q_{ef}^{(k)}(P) = 2^{\sqrt[k-1]{\frac{H_k(P)}{H(P)}}}$. ∎

The sequence $q_{ef}^{(k)}$ is increasing. (This is a simple consequence of Hölder's inequality.)

For the rest of this section, whenever the value of k is clear from the context, we set $q_e = q_{ef}^{(k)}(P)$.

Theorem 2.3: *Let $\alpha > 0$, and let $k \geq 1$ be an integer. Let $\bar{a} = a_1 ... a_n$ be a sample from a discrete memoryless source with distribution P. Let $\hat{p}$ be the empirical distribution defined by the sample. Set $t = q_e^{-\alpha}$. Define an entropy estimator $\hat{H} = \sum_{x: \hat{p}_a \geq t} \hat{p}_a \log \frac{1}{\hat{p}_a} + \log \frac{1}{t} \cdot \sum_{a:\ \hat{p}_a < t} \hat{p}_a$. Then, assuming $n \geq \Omega\left(q_e^{\alpha} \log^3 q_e^{\alpha}\right)$*

1. *If $\alpha < \frac{k-1}{k}$*
$$\hat{H} \leq H(P) \leq \frac{1}{\alpha} \cdot \hat{H}.$$
2. *If $\frac{k-1}{k} \leq \alpha \leq 1$*
$$\hat{H} \leq H(P) \leq \frac{k\left(-k\alpha + 3k - 2\right)}{2k^2 - 3k + 1} \cdot \hat{H}.$$ [7]
3. *If $\alpha > 1$,*
$$\hat{H} \leq H(P) \leq \left(1 + \frac{1}{(2k-1)\alpha^{k-1}}\right) \cdot \hat{H}.$$

with high probability over the samples $\bar{a}$.

Remark 2.4: The theorem is presented in a slightly imprecise form to gain clarity. In fact the stated approximation factors should be increased by a factor of $1 + \frac{\delta}{\log q_e}$, where δ may be taken as small as desired with an additional multiplicative cost factor of $O\left(\frac{1}{\delta^2}\right)$. [8] Furthermore, the number of samples has to satisfy an additional constraint $n \geq \widetilde{\Omega}\left(\frac{H_2}{H_1^2}\right)$. This constraint is relevant (larger than q_e^{α}) only for uninteresting settings of parameters, and we have chosen not to show it here. A complete statement of the theorem can be easily recovered from propositions 2.6 and 2.7 below. ∎

Proof: The theorem is proved in two steps. In the first step we define a functional $\widetilde{H}$ of the distribution which is conveniently thought of as a deterministic version of $\hat{H}$. We show this functional to provide a good estimate of the entropy. Then we show that with high probability either $\hat{H}$ is very close to $\widetilde{H}$ or $\hat{H}$ is very close to $H(P)$ itself.

Definition 2.5: For a parameter $0 < t < 1$ let $\widetilde{H} = \sum_{a:\ p_a \geq t} p_a \log \frac{1}{p_a} + \log \frac{1}{t} \cdot \sum_{a:\ p_a < t} p_a$. ∎

Proposition 2.6: *If $t \leq q_e^{-\alpha}$ then the three claims of theorem 2.3 hold with $\widetilde{H}$ replacing $\hat{H}$.*

[7] The factor on the right hand side is the equation of a straight line connecting the points $\left(\frac{k-1}{k}, \frac{k}{k-1}\right)$ and $\left(1, \frac{2k}{2k-1}\right)$.

[8] This overhead stems from the stipulation to approximate a deterministic quantity by a probabilistic one, and thus we need to pay the cost of ensuring typical behavior of the approximator.

The proposition is proved in the Appendix.

This is the first (deterministic) step in the proof of theorem 2.3. Now for the probabilistic part. Let ϵ be a small positive number, whose precise value will be chosen later. Set $t = q_e^{-\alpha}$. Let $\hat{H} = \sum_{a:\ \hat{p}_a \geq t} \hat{p}_a \log \frac{1}{\hat{p}_a} + \log \frac{1}{t} \cdot \sum_{a:\ \hat{p}_a < t} \hat{p}_a$, and $\widetilde{H} = \sum_{a:\ p_a \geq (1-\epsilon)t} p_a \log \frac{1}{p_a} + \log \frac{1}{t} \cdot \sum_{a:\ p_a < (1-\epsilon)t} p_a$. Let $r = \sum_{a:\ p_a < (1-\epsilon)t} p_a$.

Proposition 2.7: *Let the number of samples n be at least $\Omega\left(\max\left\{\frac{1}{\epsilon^2} q_e^\alpha \log q_e^\alpha,\ \frac{H_2}{\epsilon^2 H_1^2} \log q_e^\alpha\right\}\right)$. Then, for a sufficiently large constant implicit in the asymptotic notation* [9]

- *If $r \geq \frac{\epsilon^2 H_1^2}{H_2}$ then with high probability*
$$(1-\epsilon)^2 \cdot \hat{H} \leq \widetilde{H} \leq (1+\epsilon)^2 \cdot \hat{H}.$$
- *If $r < \frac{\epsilon^2 H_1^2}{H_2}$ then with high probability*
$$(1+2\sqrt{\epsilon})^{-1} \cdot \hat{H} \leq H(P) \leq (1-\epsilon)^{-2} \cdot \hat{H}.$$

The proposition is proved in the Appendix.

Theorem 2.3 follows by combining propositions 2.6 and 2.7, and taking $\epsilon = \frac{1}{\log q_e}$.

The efficiency of entropy approximation provided by the theorem depends on the rate of growth of the sequence of effective alphabet sizes. It turns out that this sequence does not grow too fast.

Proposition 2.8: *The k-th effective alphabet size of a distribution is essentially bounded by its alphabet size. Specifically, let q be the alphabet size. Then for $k \leq \ln q - \ln \ln q$*
$$\log q_{ef}^{(k)} = \sqrt[k-1]{\frac{H_k(P)}{H(P)}} \leq \log q + 2 \log \log q.$$

The proposition if proved in the Appendix.

Theorem 1.3 is a corollary of theorem 2.3 and the proposition. In fact we obtain a stronger claim. (Recall that in theorem 1.3 the approximation factor α has to be bounded away from 1.)

Corollary 2.9: *Theorem 1.3 is true for all $0 < \alpha < 1 - O\left(\frac{1}{\log q}\right)$. The only assumption is that the entropy of the source is bounded from below by a constant.*

3 The maximum-likelihood estimator

H_{MLE} is better than $\hat{H}$

Clearly the maximum-likelihood estimator H_{MLE} is always at least as large as the estimator $\hat{H}$ defined in theorems 1.9 and 2.3.

Next, we quote known properties of H_{MLE}. This random variable is strongly concentrated around its expectation, which is smaller than the entropy. The following is true for both finite and infinite alphabets [1]: let n be the size of the sample, then
$$\mathbf{E}(H_{MLE}) \leq H \qquad \text{and} \qquad Var(H_{MLE}) \leq \frac{\log^2(n)}{n}.$$
This means that with high probability $\hat{H} \leq H_{MLE} \leq H(P)$, [10] and H_{MLE} is the better estimator

[9] All the hidden constants in this paper are easy to compute explicitly. None of them exceeds 10.

[10] Up to a lower order error of $O\left(\frac{\log n}{\sqrt{n}}\right)$.

among the two.

Proof of proposition 1.11

It suffices to prove the proposition for $\hat{H}$. This new claim follows immediately from corollary 2.9 with $\alpha = 1 - O\left(\frac{\log\log q}{\log q}\right)$.

An infinite alphabet

The problem of entropy estimation for memoryless sources on a countable alphabet with bounded entropy moments was raised in [1]. In particular the question of convergence rate for the maximum-likelihood estimator is raised, and dealt with for specific examples. This question (among other things) is fully resolved in [13]. Assuming the first and the second entropy moments to be bounded, it is shown that

$$|H(P) - H_{MLE}| \leq \frac{4H_2(P)}{\log n} + O\left(\frac{1}{\sqrt{n}}\right)$$

with probability tending to 1 as $n \to \infty$. The lower order second term of $O\left(\frac{1}{\sqrt{n}}\right)$ appears as a result of an application of the Central Limit Theorem.

We now prove proposition 1.12 and show this result with better constants to be a simple consequence of theorem 2.3 (and known properties of H_{MLE}). Our result has an additional advantage of not relying on CLT. Consequently it is somewhat less asymptotic.

By the preceding discussion, it is enough to prove the proposition with $\hat{H}$ instead of H_{MLE}. This, in turn, is a direct consequence of the third claim of theorem 2.3, in the special case $k = 2$.

For a general k theorem 2.3 implies that for distributions with bounded k-th entropy moments the maximum likelihood estimator has a convergence rate of $\frac{H_k}{k\left(\log^{k-1} n \,-\, O(\log\log n)\right)} + O\left(\frac{\log n}{\sqrt{n}}\right)$. From this, if all the moments of the distribution are finite and do not grow too fast, a faster convergence rate can be attained. For instance, if $H_k \leq k^{O(k)}$, then H_{MLE} converges to $H(P)$ polynomially fast in n.

We remark that the main result of [13] is a matching lower bound of $\Omega\left(\frac{1}{\log n}\right)$ on the convergence rate of any sequence of universal entropy estimators over the class of distributions with bounded first and second moments. In particular, the rate of convergence of H_{MLE} is essentially optimal. This argument can be extended to show that this is true for distributions with bounded k-th entropy moments as well.

4 Sequential entropy approximation

The goal of this section is to prove theorem 1.13. We start with a generalization of theorem 2.3 for higher entropy moments.

Recall that for an integer $m \geq 1$ the m-th entropy moment of a distribution P is $H_m(P) = \sum_{a\in\mathcal{A}} p_a \log^m \frac{1}{p_a}$. For $k, m \geq 1$ define the (k,m)-th effective alphabet size $q_{ef}^{(k,m)}(P) = 2^{\sqrt[(k-1)m]{\frac{H_{km}(P)}{H_m(P)}}}$. Let $q_e = q_{ef}^{(k,m)}(P)$.

Theorem 4.1: *Let $\alpha > 0$, and let $k, m \geq 1$ be integers. Let $\bar{a} = a_1 ... a_n$ be a sample from a discrete memoryless source with distribution P. Let $\hat{p}$ be the empirical distribution defined by the sample. Set $t = 2^{-\alpha^{1/m}\log q_e}$. Define an entropy estimator $\hat{H}_m = \sum_{x:\hat{p}_a \geq t} \hat{p}_a \log^m \frac{1}{\hat{p}_a} + \log^m \frac{1}{t} \cdot \sum_{a:\ \hat{p}_a < t} \hat{p}_a$. Assume $n \geq \Omega\left(\frac{1}{t}\log^{2m+1}\left(\frac{1}{t}\right)\right)$. Then the three claims of theorem 2.3 hold with $H_m(P)$ replacing $H(P)$ and $\hat{H}_m$ replacing $\hat{H}$.*

A caveat: this is not a completely precise formulation of the theorem. In several places things have been glossed over to gain clarity (compare remark 2.4). In fact dealing with higher moments introduces additional difficulties. We do not go into details.

With all this said, the theorem is essentially true for any fixed moment m (independent of the alphabet size q), and we can even let m be a slowly (sub-logarithmically) growing function of q. The proof is very similar to the proof of theorem 2.3 and will be given in the full version of the paper.

We use the theorem with $m = 2$.

We will also need an upper bound on the new effective alphabet size. The following proposition generalizes proposition 2.8.

Proposition 4.2: *Let P be a distribution on alphabet of size q. Then for $k \leq \ln q - m \ln \ln q$*

$$\log q_{ef}^{(k,m)}(P) = \sqrt[(k-1)m]{\frac{H_{km}(P)}{H_m(P)}} \leq \log q + 2m \log \log q.$$

The proof of the proposition is similar to the proof of proposition 2.8. It will be given in the full version of the paper.

Proof of theorem 1.13

Let α and β be given.

We describe a quasi-algorithm to approximate the entropy of a distribution P within factor α given a sufficiently large sample from this distribution.

Use theorem 1.9 with approximation factor set to β and theorem 4.1 with β and $k = m = 2$ to retrieve a value for $t_1 = \Omega\left(q^{-\beta}\right)$ and for an integer $n_1 = \widetilde{\Omega}\left(\frac{1}{t}\right)$. Define functionals $\hat{H}$ and $\hat{H}_2$ with the parameter t set to t_1. Theorem 1.9 and theorem 4.1 guarantee that for a typical sample of size n_1 holds

$$\hat{H} \leq H(P) \leq \frac{1}{\beta} \cdot \hat{H} \qquad \text{and} \qquad \hat{H}_2 \leq H_2(P) \leq \frac{1}{\beta^2} \cdot \hat{H}_2.$$

Split the sample into two samples, such that one of them is of size n_1. Compute $\hat{H}$ and $\hat{H}_2$ on this sub-sample.

Set $\hat{q}_{ef} = \frac{\hat{H}_2}{\beta^2 \hat{H}}$. Then, assuming the above inequalities,

$$q_{ef} \leq \hat{q}_{ef} \leq \frac{q_{ef}}{\beta^3}.$$

Set $t_2 = \hat{q}_{ef}^{-\alpha}$. Define a new functional $\hat{H}' = \hat{H}$ with a parameter t set to t_2. Theorem 1.9 implies that for a typical sample of size $n_2 \geq \widetilde{\Omega}\left(\frac{1}{t_2}\right)$ holds

$$\hat{H}' \leq H(P) \leq \frac{1}{\alpha}\hat{H}'.$$

Compute $\hat{H}'$ on the remaining sub-sample. This is an efficiently computable functional of the original sample. By the above discussion if the size of the sample is at least $\widetilde{\Omega}\left(q^{\beta} + q_{ef}^{\frac{\alpha}{\beta^3}}\right)$ this functional gives an α-approximation of the entropy with high probability. The theorem is proved.

References

[1] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *RSA: Random Structures and Algorithms*, 19, 2001.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC-02)*, pages 678–687, New York, May 19–21 2002. ACM Press.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, USA, 2000.

[4] B. Efron and C. Stein. The jackknife estimate of variance. *Ann. Stat.*, 9(3):586–596, 1981.

[5] P. Grassberger. Entropy estimates from insufficient samplings, E-print physics/0307138 , July 2003.

[6] C. Haixiao, S.R. Kulkarni, and S. Verdu. Universal entropy estimation via block sorting. *IEEE Transactions on Information Theory,50*, July,2004.

[7] C. McDiarmid. *Surveys in Combinatorics*, chapter On the method of bounded differences, pages 148–188. Cambridge University Press, 1989.

[8] G. Miller. *Information theory in Psychology*, chapter II-B : Note on the bias of information estimates, pages 95–100. Glencoe IL: Free Press, 1955.

[9] I. Nemenman, W. Bialek, V. Steveninck, and R. de Ruyter. Entropy and information in neural spike trains: Progress on the sampling problem, March 12 2003.

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

[11] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50, 2004.

[12] S. P. Strong, R. Koberle, R. de Ruyter, V. Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Let.*, 1998.

[13] A. Wyner and D. Foster. On the lower limits of entropy estimation. *Submitted to IEEE Transactions on Information Theory*, July,2003.

5 Appendix

5.1 Proof of proposition 2.6

First, the entropy contribution $s = \sum_{p_a \le t} p_a \log \frac{1}{p_a}$ of small elements is bounded from above by a function of their probability mass $r = \sum_{p_a \le t} p_a$ and the k-th entropy moment.

Lemma 5.1:

$$s \le \sqrt[k]{r^{k-1} \cdot s_k} \le \sqrt[k]{r^{k-1} \cdot H_k}.$$

Proof:

$$s_k = \sum_{p_i \le t} p_i \log^k \frac{1}{p_i} = r \cdot \sum_{p_i \le t} \frac{p_i}{r} \log^k \frac{1}{p_i} \ge r \cdot \left(\sum_{p_i \le t} \frac{p_i}{r} \log \frac{1}{p_i} \right)^k = \frac{s^k}{r^{k-1}}.$$

The inequality follows from convexity of the function x^k. ∎

Now, clearly $\tilde{H} \le H$, for any choice of t. For the second direction,

$$\tilde{H} \ge \min_{0 \le s \le H} \left\{ (H - s) + r \log \frac{1}{t} \right\},$$

where $\log \frac{1}{t} \ge \alpha \log q_e$, and $r \ge \left(\frac{s^k}{H_k} \right)^{1/(k-1)}$. These two inequalities imply

$$\tilde{H} \ge \min_{0 \le s \le H} \left\{ (H - s) + \frac{\alpha s^{k/k-1}}{H^{1/(k-1)}} \right\}.$$

Let $f(s) = (H-s) + \frac{\alpha s^{k/k-1}}{H^{1/(k-1)}}$. At the endpoints of the interval $f(0) = H$ and $f(H) = \alpha H$. Differentiating over s, $\frac{df}{ds} = 0$ for $s = s_0 = \left(\frac{k-1}{k} \right)^{k-1} \cdot \alpha^{-(k-1)} \cdot H$. In particular, if $\alpha < \frac{k-1}{k}$, the minimum of f in the interval is attained at $s = H$, and therefore $\tilde{H} \ge \alpha H$, proving the first part of the proposition.

If $\alpha \ge \frac{k-1}{k}$, the minimum is at $s = s_0$. Substituting,

$$\tilde{H} \ge f(s_0) = \left(1 - \frac{(k-1)^{k-1}}{k^k} \cdot \alpha^{-(k-1)} \right) \cdot H.$$

This, with some easy algebra, gives the third part of the proposition. For $\frac{k-1}{k} \le \alpha \le 1$ we have

$$H \le \frac{1}{\left(1 - \frac{(k-1)^{k-1}}{k^k} \cdot \alpha^{-(k-1)} \right)} \cdot \hat{H}.$$

The factor on the right hand side is decreasing and convex in α. It is $\frac{k}{k-1}$ for $\alpha = \frac{k-1}{k}$ and is at most $\frac{2k}{2k-1}$ for $\alpha = 1$. Therefore a straight line passing through the points $\left(\frac{k-1}{k}, \frac{k}{k-1} \right)$ and $\left(1, \frac{2k}{2k-1} \right)$ can be used an upper bound for this factor. This completes the proof of the proposition.

5.2 Proof of proposition 2.7

The proof will be based on several lemmas. First, we will state the lemmas, and prove the proposition assuming the lemmas. After that we will prove the lemmas.

It is convenient to introduce notation for sets of frequent symbols for the distribution and its empirical counterpart. Let $F_p = \{a : p_a \ge (1-\epsilon)t\}$ and $F_q = \{a : \hat{p}_a \ge t\}$. Recall r stands for the sum of the rare probabilities $r = \sum_{\mathcal{A} \setminus F_p} p_a$. Let also $\tilde{r} = \sum_{\mathcal{A} \setminus F_p} \hat{p}_a$, and $\hat{r} = \sum_{\mathcal{A} \setminus F_q} \hat{p}_a$. Recall $\hat{H} = \sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \hat{r} \log \frac{1}{t}$, and $\tilde{H} = \sum_{F_p} p_a \log \frac{1}{p_a} + r \log \frac{1}{t}$.

Lemma 5.2: *The following two events hold with high probability: $F_q \subseteq F_p$ and, for all $a \in F_p$, $|\hat{p}_a - p_a| \leq \epsilon p_a$.*

Lemma 5.3: *If $r \geq \frac{\epsilon^2 H^2}{H_2}$ then with high probability*

$$|r - \tilde{r}| \leq \epsilon \cdot \tilde{r}.$$

Lemma 5.4: *If $r < \frac{\epsilon^2 H^2}{H_2}$ then*

$$\sum_{a \notin F_p} p_a \log \frac{1}{p_a} \leq \epsilon H$$

Lemma 5.5: *Let $0 < x, y \leq \frac{1}{e}$, and assume $|x - y| \leq \epsilon x$, for some $0 < \epsilon < 1$. Then*

$$(1-\epsilon) y \log \frac{1}{y} \leq x \log \frac{1}{x} \leq (1+\epsilon) y \log \frac{1}{y}.$$

Now we prove the proposition. Let us assume the state of affairs described by lemma 5.2. We will also make a couple of technical assumptions to streamline the proof. First, we assume that no element of $\mathcal{A}$ has probability mass greater than $\frac{1}{e}$. (Briefly, if such a heavy element a exists, we can learn its weight with a desired degree of precision in $O\left(\frac{1}{1-\alpha}\right)$ number of steps, and then condition on the event $\neg a$.) Next, we assume t is small enough so that $\log \frac{1}{(1-2\epsilon)t} \leq (1+\epsilon) \log \frac{1}{t}$. [11]

The first claim. Let $r \geq \frac{\epsilon^2 H^2}{H_2}$. By lemma 5.3 we may and will assume $|r - \tilde{r}| \leq \epsilon \cdot \tilde{r}$.

Now,

$$\widetilde{H} = \sum_{F_p} p_a \log \frac{1}{p_a} + r \log \frac{1}{t} \leq (1+\epsilon) \cdot \left(\sum_{F_p} \hat{p}_a \log \frac{1}{\hat{p}_a} + \tilde{r} \log \frac{1}{t} \right) =$$

$$(1+\epsilon) \cdot \left(\sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \sum_{F_p \setminus F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \tilde{r} \log \frac{1}{t} \right) \leq$$

$$(1+\epsilon) \cdot \left(\sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + (1+\epsilon) \cdot \sum_{F_p \setminus F_q} \hat{p}_a \log \frac{1}{t} + \tilde{r} \log \frac{1}{t} \right) \leq$$

$$(1+\epsilon)^2 \cdot \left(\sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \hat{r} \log \frac{1}{t} \right) = (1+\epsilon)^2 \hat{H}.$$

The first inequality follows from lemmas 5.3, 5.5, and the assumption there are no heavy elements in $\mathcal{A}$. The second inequality follows from assuming t is small enough.

Replacing $(1+\epsilon)$ with $(1-\epsilon)$ and reversing the inequalities everywhere we obtain

$$\widetilde{H} \geq (1-\epsilon)^2 \hat{H}.$$

This completes the proof of the first claim.

[11] It is sufficient to assume $t \leq 1/30$. This is a reasonable assumption for the interesting choice of parameters, in which $t \approx q^{-\alpha}$, the alphabet size q is very large, and $0 < \alpha < 1$ is a non-negligible fraction.

The second claim. Let $r < \frac{\epsilon^2 H^2}{H_2}$. Then by lemmas 5.2 and 5.4

$$\hat{H} \geq (1-\epsilon) \cdot \sum_{F_p} p_a \log \frac{1}{p_a} \geq (1-\epsilon)^2 H.$$

On the other hand, $\tilde{r}$ is a random variable with expectation r. By Markov's inequality, with probability at least $1 - \sqrt{\epsilon}$ holds $\tilde{r} \leq \frac{1}{\sqrt{\epsilon}} r$. Therefore

$$\hat{H} = \sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \hat{r} \log \frac{1}{t} \leq$$

$$\sum_{F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \sum_{F_p \setminus F_q} \hat{p}_a \log \frac{1}{\hat{p}_a} + \tilde{r} \log \frac{1}{t} \leq$$

$$(1+\epsilon) \sum_{F_p} p_a \log \frac{1}{p_a} + \frac{1}{\sqrt{\epsilon}} r \log \frac{1}{t} \leq$$

$$(1+\epsilon) \sum_{F_p} p_a \log \frac{1}{p_a} + \frac{1}{\sqrt{\epsilon}} \sum_{\mathcal{A} \setminus F_p} p_a \log \frac{1}{p_a} \leq$$

$$\left(1 + \sqrt{\epsilon} + \epsilon\right) H < (1 + 2\sqrt{\epsilon}) H.$$

∎

Proof: (of Lemma 5.2) We start with the second claim of the lemma. For any $a \in \mathcal{A}$ the empirical probability $\hat{p}_a$ counts the average number of appearances of a in a sample of size n. Therefore $\hat{p}_a$ is an empirical mean of a binomial random variable with probability p_a of success, and we can apply Chernoff's bound [7] to measure its probability of deviating from its mean p_a. (Recall that for all $a \in F_p$ holds $p_a \geq (1-\epsilon)t$.)

There are at most $\frac{1}{(1-\epsilon)t}$ symbols in F_p. The probability that there is one for which $|\hat{p}_a - p_a| > \epsilon p_a$ is at most

$$\frac{2}{(1-\epsilon)t} \cdot \exp\left\{-\frac{1}{3}\epsilon^2 (1-\epsilon) t n\right\}$$

We have used the union bound to account for all the symbols in F_p and Chernoff's bound to bound individual deviations. Since $n \geq \Omega\left(\frac{\frac{1}{t}\log\frac{1}{t}}{\epsilon^2}\right)$, choosing the constant implicit in the asymptotic notation to be sufficiently large will make this probability as small as $O(t)$, say, which is sufficiently good for our purposes.

The first part of the lemma is harder. The complementary event is that there is a symbol in F_q but not in F_p. Let $\mathcal{I}$ be the (countable) set of all such symbols. Let $\mathcal{I}_1 = \{a \in \mathcal{I},\ p_a \geq t/2\}$, and let $\mathcal{I}_2 = \mathcal{I} \setminus \mathcal{I}_1$.

The cardinality of $\mathcal{I}_1$ is at most $\frac{2}{t}$, and we can use the union bound and the Chernoff bound as before, to bound its probability by

$$\frac{4}{t} \cdot \exp\left\{-\frac{1}{6}\epsilon^2 t n\right\}.$$

As to $\mathcal{I}_2$, we use the large deviation inequality for the binomial random variable [7] to obtain

$$Pr\left(a \in \mathcal{I}_2\right) \leq \left(\left(\frac{p_a}{\hat{p}_a}\right)^{\hat{p}_a} \left(\frac{1-p_a}{1-\hat{p}_a}\right)^{1-\hat{p}_a}\right)^n.$$

Therefore

$$Pr\left(\mathcal{I}_2\right) \leq \sum_{a:\ p_a \leq \frac{t}{2}} \left(\left(\frac{p_a}{\hat{p}_a}\right)^{\hat{p}_a} \left(\frac{1-p_a}{1-\hat{p}_a}\right)^{1-\hat{p}_a} \right)^n .$$

Consider the function $f(x) = \left(\frac{p_a}{x}\right)^x \left(\frac{1-p_a}{1-x}\right)^{1-x}$, for a fixed a. For x larger than p_a, this function is easily seen to be decreasing. Therefore $f(x) \leq f(t)$ for all $x \geq t$. Since $q_a \geq t$ for all $a \in \mathcal{I}_2$,

$$Pr\left(\mathcal{I}_2\right) \leq \sum_{a:\ p_a \leq \frac{t}{2}} \left(\left(\frac{p_a}{t}\right)^{t} \left(\frac{1-p_a}{1-t}\right)^{1-t} \right)^n .$$

Now consider the function $g(x) = \left(x^t(1-x)^{1-t}\right)^n$. It is not hard to see that if $n \geq \frac{5}{t}$ (which we may assume) the second derivative of g is nonnegative for $0 \leq x \leq \frac{t}{2}$, and therefore g is convex in this interval.

It follows that the bound is maximized when all the p_a are as large as possible. Recall that $p_a \leq \frac{t}{2}$ and $\sum_a p_a \leq 1$. Assuming that there are $\frac{2}{t}$ symbols a in $\mathcal{I}_2$ with probability $\frac{t}{2}$ each leads to the bound

$$Pr\left(\mathcal{I}_2\right) \leq \frac{2}{t} \cdot \left(\left(\frac{1}{2}\right)^{t} \left(\frac{2-t}{2-2t}\right)^{1-t} \right)^n \leq \frac{2}{t} \cdot \left(\frac{1}{2}\right)^{tn} .$$

Summing up,

$$Pr\left(\mathcal{I}\right) \leq \frac{4}{t} \cdot e^{-\frac{1}{6}\epsilon^2 tn} + \frac{2}{t} \cdot \left(\frac{1}{2}\right)^{tn} .$$

This probability can be made as small as $O\left(t\right)$, by an appropriate choice of constant in the definition of n. ∎

Proof: (of lemma 5.3) The random variable $\tilde{r}$ counts the average number of appearances of symbols not in F_p in the sample. Therefore $\tilde{r}$ is an empirical mean of a binomial random variable with probability r of success, and by Chernoff's bound

$$Pr\left(|r - \tilde{r}| > \epsilon \cdot r\right) \leq 2 \cdot \exp\left\{-\frac{1}{3}\epsilon^2 rn\right\}.$$

This probability can be made as small as as $O\left(t\right)$ by an appropriate choice of constant in the definition of n. ∎

Proof: (of lemma 5.4)

$$H_2 = \sum_{\mathcal{A}} p_a \log^2 \frac{1}{p_a} \geq \sum_{a \notin F_p} p_a \log^2 \frac{1}{p_a} = r \cdot \sum_{a \notin F_p} \frac{p_a}{r} \log^2 \frac{1}{p_a} \geq$$

$$r \cdot \left(\sum_{a \notin F_p} \frac{p_a}{r} \log \frac{1}{p_a} \right)^2 = \frac{1}{r} \cdot \left(\sum_{a \notin F_p} p_a \log \frac{1}{p_a} \right)^2 .$$

The second inequality uses convexity of the function x^2.

Therefore

$$\sum_{a \notin F_p} p_a \log \frac{1}{p_a} \leq \sqrt{r H_2}.$$

∎

Proof: (of lemma 5.5) We may assume log stands for the natural logarithm.

If $x \geq y$ then $x \log \frac{1}{x} \leq (1+\epsilon) y \log \frac{1}{y}$.

Suppose $x < y$. By assumption $y \leq \frac{1}{e}$, and therefore $\log \frac{1}{y} \geq 1$. Consequently $(1+\epsilon)\log \frac{1}{y} \geq \log \frac{1}{y} + \log(1+\epsilon) = \log \frac{1+\epsilon}{y} \geq \log \frac{1}{x}$. Therefore $x \log \frac{1}{x} \leq (1+\epsilon) y \log \frac{1}{y}$.

The second direction is proved similarly. ∎

5.3 Proof of proposition 2.8

The set of all q-point distributions with entropy at least 1 is a compact set in the q-dimensional Eucliden space. $\frac{H_k}{H}$ is a continuous function on this set. As such it attains its maximum in the domain. Let P be a point on which a maximum is attained, and let $H(P) = C$. Thus the point P solves an additional optimization problem:

$$\max \left\{ H_k \;\middle|\; H = C \right\}.$$

Let $P = (p_1 ... p_q)$. The Lagrange conditions for P imply that there are constants λ, μ such that [12]

$$\log^k p_i + k \log^{k-1} p_i + \lambda \log p_i + \mu = 0.$$

Namely for any i a negative number $x_i = \log p_i$ is a root of a polynomial $P(x) = x^k + kx^{k-1} + \lambda x + \mu$.

Now we need a simple lemma about roots of sparse polynomials.

Lemma 5.6: *Let $P(x)$ be a real polynomial of the form $P(x) = x^d + ax^{d-1} + bx + c$. Then P has at most three negative roots.*

Proof: If $d \leq 3$ the claim is immediate, so assume $d > 3$. Take the second derivative to obtain $P''(x) = d(d-1)x^{d-2} + a(d-1)(d-2)x^{d-3}$. Therefore P'' has at most one negative root. Since between any two roots of a continuously differentiable function there is a root of its derivative, the lemma follows. ∎

This implies there are at most three distinct values of p_i, call them $r \leq s \leq t$. Set R be the total of all the probabilities that equal r, and similarly for S, T. Then

$$\frac{H_k(p)}{H(p)} = \frac{R \log^k \frac{1}{r} + S \log^k \frac{1}{s} + T \log^k \frac{1}{t}}{R \log \frac{1}{r} + S \log \frac{1}{s} + T \log \frac{1}{t}}.$$

Let $\rho = \frac{1}{q \log q}$. Now there are three cases to consider, depending on how large ρ is compared to r, s, t. Observe that $\rho < t$ since $t \geq \frac{1}{q}$.

1. $$s < \rho.$$

   The maximum of the function $f(x) = x \log^k \frac{1}{x}$ in the interval $[0,1]$ is attained at e^{-k}. Since $k \leq \log q - \log\log q = \log \frac{1}{\rho}$ the maximum of f in the interval $[0, \rho]$ is attained at the right endpoint. Therefore $R \log^k \frac{1}{r} + S \log^k \frac{1}{s} \leq q\rho \log^k \frac{1}{\rho} = \frac{1}{\log q} \cdot \log^k \frac{1}{\rho}$. Since $t \geq \frac{1}{q}$, and $H \geq 1$

   $$\frac{H_k(p)}{H(p)} \leq \log^{k-1} q + \frac{1}{\log q} \cdot \log^k \frac{1}{\rho}.$$

[12] Without loss of generality in this proof all the logarithms are natural.

We need to show this is at most $\log^{k-1}\left(q\log^2 q\right)$. Since, for $y > x$ and $t \geq 1$ holds $y^t - x^t \geq (y-x)t\left(\frac{x+y}{2}\right)^{t-1}$, we have

$$\log^{k-1}\left(q\log^2 q\right) - \log^{k-1} q \geq 2(k-1)\ \log\log q \cdot \log^{k-1}\frac{1}{\rho} \geq \frac{1}{\log q}\cdot\log^k\frac{1}{\rho}.$$

2.

$$r < \rho < s.$$

This is analysed as the previous case, but now we cluster s and t in the 'large probability' group, and r in the 'lightweight' group. We have

$$\frac{H_k(p)}{H(p)} \leq \log^{k-1}\frac{1}{\rho} + \frac{1}{\log q}\cdot\log^k\frac{1}{\rho} \leq \log^{k-1}\left(q\log^2 q\right),$$

as before.

3.

$$\rho < r.$$

In this case

$$\frac{H_k(p)}{H(p)} \leq \log^{k-1}\frac{1}{\rho} \leq \log^{k-1}\left(q\log^2 q\right).$$

∎

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'