On the lower limits of entropy estimation

Abraham J. Wyner and Dean Foster
Department of Statistics,
Wharton School,
University of Pennsylvania,
Philadelphia, PA
e-mail: ajw@wharton.upenn.edu foster@wharton.upenn.edu

June 15, 2003

Abstract

In recent paper, Antos and Kontoyiannis [1] considered the problem of estimating the entropy of a countably infinite discrete distribution from independent identically distributed observations. They left several open problems regarding the convergence rates of entropy estimates, which we consider here. Our first result, is that the plug-in estimate of entropy is as efficient as a match length estimator when applied to the class of memoryless sources on a countable alphabet with finite entropy and finite entropy "variance". Our second result provides lower bounds on the convergence rate of any sequence of universal estimators of entropy over the same class of distributions. Finally, we consider an estimator based on match lengths that achieves this lower bound to first order. The surprising conclusion that follows is that a match-length estimator is first order optimal over the simplest class of distributions for which entropy estimation is non-trivial. We describe how this in turn implies that the Lempel-Ziv algorithm has an optimal convergence rate among the class of universal data compression algorithms over the arguably simplest class of non-trivial sources.

Keywords: Entropy estimation, Coding, Lempel-Ziv algorithm, countably infinite alphabets, lower bounds, redundancy.

1 Introduction

We begin with some definitions. The entropy $H(P)$ is a smooth functional of the distribution P. The problem of entropy estimation given sequences of i.i.d. random variables is quite naturally solved by a plug-in estimate

$$\hat{H}(P_n) = -\sum_{i \in A} P_n(i) \log P_n(i),$$

where $P_n(\cdot)$ is the empirical estimate of the distribution P on discrete alphabet A of i.i.d. random variables $X_1, X_2, \ldots, X_n$. For a memoryless source, P_n is the maximum likelihood estimate of P as well as a sufficient statistic for X_1^n. The Rao-Blackwell theorem implies that the best estimate of the source entropy should be a function of the sufficient statistic for example the plug-in estimate $\hat{H}(P_n)$. The plug-in estimate is universal and optimal not only for finite alphabet i.i.d sources but also for finite alphabet, finite memory sources. On the other hand, practically as well as theoretically, these problems are of little interest.

We consider here the simplest extension of the class of finite alphabet finite, finite memory sources, namely, the class $\mathcal{P}$ of memoryless sources on a countable alphabet with $H^{(x)} < \infty$ for some $r \geq 2$. Where

$$H^{(r)}(P) = E(-\log P(X))^r.$$

The special case of interest is with $r = 2$. Universal entropy estimation over $\mathcal{P}$ is again naturally accomplished using $\hat{H}(P_n)$ (since P_n is still sufficient for P). But now even consistency is not obvious and universal convergence rates can be arbitrarily slow (see [1]). In the finite alphabet case the convergence rate of $\hat{H}(P_n)$ to $H(P)$ occurs at rate

$$\approx \frac{\sigma}{\sqrt{n}}$$

where

$$\sigma^2 = \mathbf{Var}\{-\log P(X)\}.$$

We call σ^2 the "entropy variance". In contrast, the convergence of $\hat{H}_n$ to H can be arbitrarily slow for any universal estimator. Antos and Kontoyiannis prove that for a much smaller class of discrete distributions which satisfy a sharp tail condition it is possible to demonstrate that the plug-in estimator converges at rates

$$\approx O(n^{-\beta}) \text{ for } 0 < \beta < 1$$

where β depends on the precise rate at which the tail vanishes. On the other hand for the much more natural class $\mathcal{P}$ of discrete distributions on a countable alphabet with finite entropy and finite entropy variance they show that a non-parametric estimate based on match lengths converges at the tortoise like rate of

$$\approx O(\frac{1}{\sqrt{\log n}}).$$

They were unable to prove that the plug-in estimate converges even at this rate. Furthermore, they conjecture that

"the rate $\{n^{-\frac{r-1}{r}}\}$ can be achieved if we restrict ourselves to the case where $H^{(r)} < \infty$ at least for $1 < r \leq 2$".

The main aim of this paper is to resolve each of these conjectures. We begin by establishing that for the class of distributions $\mathcal{P}$ discussed above that there can be no universal estimator $\hat{H}$ that converges to H at a rate faster than $O(\frac{1}{(\log n)^{(1+\epsilon)}})$ for all sources $P \in \mathcal{P}$ and any $\epsilon > 0$. This is a generalization of the result in [1] where it is shown that there exist a family of distributions for which the plug-in estimator of entropy does not converge faster than $O(\frac{1}{(\log n)^{(1+\epsilon)}})$.
This lower bound is particularly interesting since it is so *slow*. The implication is that no method of entropy estimation can converge rapidly (in a true sense of the word) even when one assumes a source to be memoryless and with bounded entropy variance. Consequently, universal rapid rates of convergence do not exist under natural assumptions outside of the simple class of completely finite (finite alphabet and finite memory) distributions.

Remarkably this lower bound is in fact achieved by a non-parametric estimator based on match lengths. We shall prove that this estimator $\hat{H}_{ML}(n)$ is eventually almost surely within $\frac{Const.}{\log n}$ of $H(P)$ for any $p \in \mathcal{P}$. Since this is to first order the best possible rate of convergence among distributions in this natural class we conclude that the match length estimator has the best rate of convergence among all estimates of the entropy that are consistent for all $P \in \mathcal{P}$.

1.1 Coding and Entropy Estimation

Scientists, mathematicians, information theorists, computer scientists, and others have all longed for a precise and measurable way to define the complexity of an individual sequence. This tantalizing theory sets up three levels of problems of various difficulty. On the first level, we consider the individual sequences themselves with no restrictions. This is an impossible problem at best, and an attractive nuisance at worst. The theory of Kolmogorov complexity while elegant, can serve only to bound complexity but never to know it. On the second level, we can impose a stationary probability model and define complexity to be entropy. This condition transforms complexity into a property of the source and the problem becomes estimation. If we assume further that the data generating source is ergodic then we can define asymptotically consistent procedures for estimating entropy [3]. Finally, on the third level we talk about memoryless sources: the arguably simplest class of data generating probability models. The problem of estimating the entropy of a memoryless source, which we shall take up here, is actually not quite so trivial.

The Kolomogorov Complexity of an individual sequence is the length of the shortest universal computer program that can output the sequence. It cannot be computed but it can be bounded. When the data generating source is a stationary ergodic process then the Kolomogrov complexity will equal, with high probability and on average, the process

entropy rate. This sets up a nice duality: a good code is therefore a good estimate of entropy. Conversely, lower bounds on the performance of entropy estimators are also lower bounds for coding. In this section we stress the implications of this duality.

It is not widely known, but the motivations behind the Lempel-Ziv [6] data compression algorithm were centered only indirectly on coding and more directly on the problem of finding a practical scheme to estimate entropy of an individual sequence. It is well known ([11], [10], [9]) that the Lempel-Ziv algorithm (in any number of its versions) can achieve a coding redundancy, when applied to finite alphabet, finite memory sources, that is $\emptyset_P(\frac{1}{\log n})$. With countable alphabets the coding problem becomes harder (see [7]). With restrictions (such as montonicity) it is possible to determine asymptotic redundancies for countable processes (see [4]). But our results nevertheless imply that an entropy estimate based on the average code redundancy that is at least $\Omega(\frac{1}{logn})$. This suggest that the coding redundancy of the Lempel-Ziv algorithm as well as other "grammar" based codes [5] are to first order optimal with respect to the class of countable memoryless sources with bounded entropy and entropy variance. Indeed, this class is arguably the simplest class of sources outside of finite alphabet finite memory sources, which suggests that the slow rate of convergence to the entropy is an intrinsic difficulty and not a shortcoming of the algorithm.

2 A Universal Lower Bound

The asymptotic equipartition theorem connects the entropy to the exponent of the cardinality of the typical set. As such, it is easiest to think of the entropy as an "effective" population size and an estimator of entropy as an attempt to count that effective population size. This is particularly difficult if there are an extraordinarily large number of extraordinarily rare symbols. For a memoryless source, the counting process is analogous to a fisherman who wants to estimate a population of fish. Knowing only how to fish, he heads out to sea to capture, label and release. Clearly, he can only upper bound the population when a fish is recaptured. Until the first recapture point, no upper bound on the population can posited (at least not with finite mean squared error). In this work, we will play the part of the fisherman. The challenge is to construct two populations satisfying certain constraints that are as different as possible, and for which recapture is not likely for either population . It follows, that in such an event, no estimator can distinguish between the two populations. We will need to construct an uncountable family of distributions on our countable alphabet. This is a difficult construction and we will do it twice: first on sets and then using random variables.

We will define a family of random variables indexed by a infinite binary sequence $u_1u_2u_3...$ which we denote u. To this end we consider sets A_k. For every k each A_k is partitioned into subsets A_{ik}. Each of these sets consists of atoms $\{a_{ijk}\}$in arbitrary countable alphabet $\mathcal{A}$. We define a distribution P_u and a random variable X_u by assigning probabilities to the

atoms $\{a_{ijk}\}$ and the sets A_k and A_{ik}. Therefore X_u is a family of random variables on the countable space $\{a_{ijk}\}$.

We begin by assuming that we have a sequence of integers $\{n_k\}$ that goes to ∞. We will choose $\{n_k\}$ later when our ends become clearer. Our first step is to assign probability to the collection of atoms in A_i so that for every u

$$\Pr\{X_u \in A_k\} = \frac{1}{(\log n_k)^{2+\epsilon}}, \tag{1}$$

where $\epsilon > 0$ is arbitrary. Next we assign equal probability to the sets A_{ik} contained in A_k so that

$$\Pr\{X_u \in A_{ik}\} = \frac{1}{n_k^2}.$$

Since $A = \bigcup_i \{a_i\}$ we conclude that $i = 1, 2, \ldots, I_{n_k}$, where

$$I_{n_k} = \frac{n_k^2}{(\log n_k)^{2+\epsilon}}.$$

Now assume that A_{ik} contains n_k atoms. We introduce the dependence on u by letting X_u equal the first atom if $u_k = 0$ and a randomly chosen atom if $u_k = 1$. Formally, if $u_k = 0$ we let $Pr\{a_{ijk}\} = 0$ for all $j > 1$, which means that if $X_u \in A_{ik}$ then $X_u = a_{i1k}$. On the other hand, if $u_k = 1$ then we choose equally among the the atoms $\{a_{ijk} \in A_{ik}\}$ so that

$$\Pr\{a_{ijk}\} = \frac{1}{n_k^3}.$$

Now that we have carefully constructed P_u in context we can back out a simpler definition. Given binary $u \in [0, 1]$ and sequence n_k for $k = 1, 2, \ldots$, choose integer K with

$$P(K = k) = \frac{1}{(\log n_k)^{2+\epsilon}},$$

integer I conditionally with

$$P(I = i | K = k) = \frac{(\log n_k)^{2+\epsilon}}{n_k^2}$$

and J with

$$P(J = j | K = k, u_k = 1) = \frac{1}{n_k}$$

otherwise $P(J = 1 | K = k, u_k = 0) = 1$.

Let's begin by computing the entropy of X_u for any u.

$$\begin{aligned} H(X_u) &= -\sum_{t=1}^{\infty} P(A_t)[u_t \log n_t^3 + (1-u_t)\log n_t^2] \quad (2) \\ &= \sum_{t=1}^{\infty}[u_t \frac{3}{(\log n_t)^{1+\epsilon}} + (1-u_t)\frac{2}{(\log n_t)^{1+\epsilon}}]. \quad (3) \end{aligned}$$

Now let's check that for every u, X_u has finite entropy and entropy variance. We may assume that $u_k = 1$ for all k and observe that for this choice

$$\begin{aligned} H(X_u) &= \sum_{ijk} -P(a_{ijk}) \log P(a_{ijk}). \\ &= \sum_k \frac{1}{\log n_k^{2+\epsilon}} \log n_k^3 \\ &= \sum_k \frac{3}{\log n_k^{1+\epsilon}}. \end{aligned}$$

In order for $H(X_u)$ to be bounded we need n_k go to go to infinity at least as fast as 2^k. The *entropy variance* of a random variable X is the second moment of the the log likelihood. It is easy to see that for all u, X_u has bounded entropy variance for an appropriate choice of n_k. Again, we may assume that $u_k = 1$ for all k. We compute:

$$\begin{aligned} E[\log P(X_u)]^2 &= \sum_{ijk} -P(a_{ijk})(\log P(a_{ijk}))^2. \\ &= \sum_k \frac{1}{\log n_k^{2+\epsilon}} (\log n_k^3)^2 \\ &= \sum_k \frac{9}{(\log n_k)^{\epsilon}} \end{aligned}$$

Thus X_u will have bounded entropy and entropy variance for appropriate choices of n_k that tend to infinity fast. Two possible choices are $n_k = 2^{2^k}$ and $n_1 = 2$ with

$$n_{k+1} = 2^{n_k}.$$

Now suppose that we define u and u' to differ only on coordinate k for which $u_k = 0$ and $u'_k = 1$. Then it follows from (3) that

$$\begin{aligned} H(X_{u'}) - H(X_u) &= [\frac{3}{(\log n_k)^{1+\epsilon}} - \frac{2}{(\log n_k)^{1+\epsilon}}] \quad (4) \\ &= \frac{1}{(\log n_k)^{1+\epsilon}} \quad (5) \end{aligned}$$

If we suppose further that $n^{k+1} = 2^{n_k}$ we can consider the more general situation when $u_k = 0$, $u'_k = 1$ but with coordinates satisfying only the restriction $u_i = u'_i$ for $i < k$. In this case we compute

$$
\begin{aligned}
|H(X_{u'}) - H(X_u)| &= [\frac{3}{(\log n_k)^{1+\epsilon}} - \frac{2}{(\log n_k)^{1+\epsilon}}] + && (6)\\
&\quad \sum_{t=k+1}^{\infty} [u_t \frac{3}{(\log n_t)^{1+\epsilon}} + (1-u_t)\frac{2}{(\log n_t)^{1+\epsilon}}] && (7)\\
&\geq \frac{1}{(\log n_k)^{1+\epsilon}} - \frac{1}{n_k} && (8)
\end{aligned}
$$

To recap, we have now created two probability distributions each with finite entropy and entropy variance, whose entropy is "far" apart in the sense of (5). Our goal now is two connect the pair $(X_u, X_{u'})$ so that with high probability a sequence of independent copies $(X_{u,i}, X_{u',i})$ for $i = 1, 2, \ldots$ will be indistinguishable.To do this, we introduce a coupling which builds X_u and $X_{u'}$ on the same probability space so that $X_u = X_{u'}$ unless $X \in A_k$. In that event, $X_{u'}$ is chosen independently from the atoms in $\{a_{ijk}\} \in A_k$ while X_u is chosen independently from the atoms $\{a_{i1k} \in A_k\}$. It follows that

$$\Pr\{X_u = X_{u'}\} > 1 - \frac{1}{(\log n_k)^{2+\epsilon}}$$

since $\Pr\{X_u \notin A_k\} = 1 - \frac{1}{(\log n_k)^{2+\epsilon}}$.

Suppose we repeat the coupling to produce n_k independent identically distributed copies of the coupled random variables $(X_u, X_{u'})$. To make things more fun, we assume that we observe the entire sequence $X_{u,i}$ or $X_{u',i}$ but it is unknown whether we are observing X_u or $X_{u'}$. Furthermore, we assume that a competitor observes the opposing sequence! Let $\hat{H}(n_k)$ be our estimate and let $\hat{H}'(n_k)$ be our competitor's estimate.

Now let $\hat{H}_{n_k}$ be an estimate of entropy derived from n_k observations. We say that any estimate $\hat{H}(X_1^n)$ is a *consistent* estimate of entropy if $\lim_{n\to\infty} H_n(X_1^n) = H(X)$. It is said to be *universally* consistent for the class $\mathcal{P}$ of probability distributions if for every $P \in \mathcal{P}$ the estimate $\hat{H}_n$ is consistent. Let $\mathcal{H}$ be the set of universally consistent estimators over $\mathcal{P}$. Now the entropy is a function of the likelihood function $P(\cdot)$. If there is no connection between the alphabet $\mathcal{A}$ and the probability function mapping symbols $a \in \mathcal{A}$ into $[0, 1]$ then it may be assumed, without loss of generality with respect to worst case performance, that every estimator $\hat{H} \in \mathcal{H}$ is invariant with respect to one-to-one mappings of the symbols of $\mathcal{A}$ into itself. With this property in mind, we say that two sequences of random variables $X_{u,i} \in \mathcal{A}$ and $X_{u',i} \in \mathcal{A}'$ are equivalent ($\cong$) if there exists a one-to-one relabeling such that the relabeled sequences are permutations of each other. This will be true if the empirical probability distributions are the same. We now state and prove a lemma which establishes that with high probability $\{X_{u,i}\}_{i=1}^{n_k} \cong \{X_{u',i}\}_{i=1}^{n_k}$.

Lemma A: Let $\{X_u\}_1^{n_k}$ and $\{X_{u'}\}_1^{n_k}$be i.i.d couplings as defined above. Then there exists a $\delta_{n_k} < 1/2$ such that for all n_k:

$$Pr\{\{X_u\}_1^{n_k} \cong \{X_{u'}\}_1^{n_k}\} > 1 - \delta_{n_k}.$$

Proof: The result is actually quite simple given the construction. For an invariant one-to one mapping not to be possible, there must exist a symbol $a \in A_k$ that occurs at least twice in either sequence. The probability of this event is easily bounded. We begin first with $\{X_u\}_1^{n_k}$. Let Z be the number of atoms in A_k that occur twice in $\{X_u\}_1^{n_k}$. Then

$$EZ = \sum_{i=1}^{n_k-1} \sum_{j=i}^{n_k} \Pr\{X_{u,i} = X_{u,j} \in A_k\}$$

Now each atom in A_k has probability $\frac{1}{n_k^2}$ of occurring it follows that

$$\begin{aligned} EZ &= \frac{n_k(n_k-1)}{2} \Pr\{X_{u,i} \in A_k\} \Pr\{X_{u,j} = X_{u,i} | X_{u,i} \in A_k\} \\ &= \frac{n_k(n_k-1)}{2} \frac{1}{(\log n_k)^{2+\epsilon}} \frac{1}{n_k^2} \\ &< \frac{1}{2(\log n_k)^{2+\epsilon}} \\ &< \frac{1}{2} \end{aligned}$$

Now it follows from the Markov inequality that

$$\Pr\{Z \geq 1\} \leq EN \leq \frac{1}{2(\log n_k)^{2+\epsilon}}$$

Which implies that

$$\Pr\{Z = 0\} \geq 1 - \frac{1}{2(\log n_k)^{2+\epsilon}} = 1 - \delta_{n_k}.$$

Now let Z' be the number of symbols in A_k that occur twice in sequence $\{X_{u'}\}_1^{n_k}$. Since each $a \in A_k$ has probability $\frac{1}{n_k^3}$ it is easy to follow the steps of the preceding bound to show that

$$\Pr\{Z' = 0\} \geq 1 - \frac{1}{2n_k(\log n_k)^{2+\epsilon}} = 1 - \frac{\delta_{n_k}}{n_k}.$$

Now, observe that

$$Pr\{\{X_u\}_1^{n_k} \cong \{X_{u'}\}_1^{n_k}\} \geq 1 - \delta_{n_k}(1 + \frac{1}{n_k})$$

This completes the proof of the lemma.

To review, we are now in possession of two key facts: that there exist two distributions with finite entropy and finite entropy variance that are quite different with respect to entropy, yet are likely to be mathematically invariant with respect to universally consistent estimators of entropy. We formalize this in the following:

Theorem A: Let $\mathcal{P}$ be the space of probability distributions on countable sets with finite entropy and finite entropy variance. Let $\Delta > 0$ be arbitrary. Then there does not exist an invariant universal estimator $\hat{H}_n$ that is uniformly $O_P(\frac{1}{(\log n)^{1+\Delta}})$ consistent over $\mathcal{P}$.

To prove this, construct coupled X_u and $X_{u'}$ for any $\epsilon < \Delta$. Lemma A implies that with probability greater than 1/2 and tending to one, the two sequences are equivalent with respect to one-to-one relabeling of the symbols. We now proceed by contradiction. Suppose that X_1^n and Y_1^n sequences of random variables whose distributions belong to $\mathcal{P}$. Now suppose that $\hat{H}$ is a a universally consistent estimator for which there always exists an N so large such that for all $n > N$ the following inequalities hold with probability arbitrarily close to one:

$$|\hat{H}(X_1^n) - H(X)| \leq \frac{C_1}{(\log n)^{1+\Delta}} \tag{9}$$

$$|\hat{H}(Y_1^n) - H(Y)| \leq \frac{C_2}{(\log n)^{1+\Delta}} \tag{10}$$

Now let X_1^n and Y_1^n be the coupled pair $(X_u, X_{u'})$ for k large enough that $n_k = 2^{n_{k-1}} = n$ for some $n > N$. Now Lemma A implies that the two estimates are identical with positive probability. On the other hand, equation (1) reveals that the entropy of Y is greater than entropy of X by an amount larger than permitted by equations (2) or (3). Thus either (2) or (3) must be false, proving the theorem.

In the proof of theorem of A we chose to let

$$n_{k+1} = 2^{n_k}.$$

With respect to Theorem A, we could have gotten by with less since all that was required of our choice is that for any pair (u, u') that differ in the k^{th} coordinate, the random variables $X_u, X_{u'}$ had to be in $\mathcal{P}$. Any choice that caused the entropy variance to be finite would have worked. But we have a stronger theorem in mind, for which a more rigorous choice is required. We have so far only proved that convergence at the $O_P(\frac{1}{\log n})$ rate is not possible *uniformly* over $\mathcal{P}$ (not by any universally consistent estimator). We shall now prove that this convergence rate is not possible *pointwise*. That is, for any estimator $\hat{H}$ there exists a $P \in \mathcal{P}$ (whose choice depends on $\hat{H}$) that does not converge pointwise at the desired rate. That is, no matter how large N, there will always exist an $n = n_k > N$ with

$$|\hat{H}(X_1^n) - H(X)| \geq \frac{1}{(\log n)^{1+\epsilon}},$$

with non-vanishing probability.

The result is already almost in our grasp. The trick is to construct a limiting "bad" distribution (i.e. a distribution for which (9) fails to hold) out of a sequence of distributions that are bad for $n = n_k$. The main idea is as follows. If the first k coordinates of u and u' differ only in coordinate k, then $\hat{H}$ cannot be too close to both $H(X_u)$ *and* $H(X_{u'})$ with high probability. Here is where (8) becomes important. Since the entropy of X_u and $X_{u'}$ are determined to within $\frac{1}{n_k}$ by the first k coordinates of u and u' it follows that we can check if (9) fails based only on the first k coordinates. Let u_k^* be the value of either u_k or u_k' for which the estimator is bad (if both are bad let $u_k^* = 0$):

$$|\hat{H}(X_{u^*}) - H(X_{u^*})| \geq \frac{1}{(\log n_k)^{1+\epsilon}} \tag{11}$$

with probability at least 1/2. Now, Let u^* be the infinite binary sequence whose k^{th} coordinate is u_k^*. Let X_{u^*} be a random varaible with distribution P_{u^*}. We shall show that $\hat{H}$ does not converge pointwise almost surely faster than $\frac{1}{(\log n)^{1+\epsilon}}$.

Since the construction of P_{u^*} is inductive we need to show that for every k that (11) holds unconditionally on the selection of $u_{k+1}^*, u_{k+2}^*, \ldots$. To see that this is an issue, recall that in the proof of Theorem A, we require that sequences u and u' differ only at coordinate k. Consequently, it becomes possible that the selection of u_k^* will depend on coordinates of u that are larger than k. Thus, as we proceed through the inductive process to select u_k^* for ever larger k, we may alter the performance of the estimator at smaller values. To prevent this problem in the sequel to Theorem A, we must only require that $u_i = u_i'$ for all $i < k$. This is where our choice of n_k becomes critical. The coupling of X_u to $X_{u'}$ forces the pair to agree unless $X_u \in A_k$. Using n_k as defined and noting (1) it follows that

$$\Pr\{X_u \in \bigcup_{l>k} A_l\} = \sum_{l \leq k} \frac{1}{n_l^{2+2\epsilon}}.$$

Now let $G_k = \{\omega : \forall l \leq n^k \; X_{u,i}(\omega) \notin \bigcup_{i>k} A_l\}$. So G_k implies that no atoms in A_l appear for $l > k$ in the sequence $\{X_u\}_{i=1}^{n^k}$. A crude bounding operation shows that

$$\Pr\{G_k\} \leq \frac{1}{n_k} = \gamma_k \to 0 \text{ as } k \to \infty.$$

Since G_k does not occur except for a finite number of times, it follows that $\hat{H}$, except for a finite number of k, is independent of the values of u_l for $l > k$. It follows that for k sufficiently large we can determine the value of u_k for which the estimator will be bad based only on the values of $u_1^*, u_2^*, \ldots, u_{k-1}^*$. We can now conclude the section with a theorem.

Theorem B: For any estimator $\hat{H}$ and any $\epsilon > 0$ there exists a distribution $P \in \mathcal{P}$ and random variables $X_1, X_2, \ldots$ drawn i.i.d according to P for which the event

$$\{|\hat{H}(X_1^n) - H(X)| \geq \frac{1}{(\log n)^{1+\epsilon}}\}$$

occurs for infinitely many n.

2.1 Local Limit: moving truth

So far we have considered the problem of entropy estimation in the context of sequence of random variables X_1^∞ generated from a single source P from a fixed class $\mathcal{P}$. Within this context all finite alphabet memoryless sources exhibit fundamentally equivalent asymptotic behavior. In essence, any finite problem, including finite memory Markov sources, are easy: eventually, enough data can be accumulated to crush the problem into submission.

Statisticians have pursued a different approach to make the asymptotics of finite memory processes more interesting. Instead of assuming that the source probability distribution is a fixed P, we suppose that the data is actually a triangluar array $X_{i,n}$ with $n = 1, 2 \ldots$ and $i = 1, 2, \ldots, n$. For each n the sequence $X_{1,n}^{n,n}$ is assumed to be drawn i.i.d. from P_n. In this case, the truth "moves" with increasing n. The asympototics (often called local limits) of such problems can be very different. For example, the definition of coding redundancy has many definitions. It is most commonly defined in the sense as the difference between the average code length and the entropy rate of the process. Alternatively, the redundancy can be defined to be the *ratio* of the average code length to the entropy rate. In the usual asmptotics, where P is fixed, the two£redundancies are equivalent (up to a constant). Thus, if a code is optimal with respect to the usual redundancy it is also optimal with respect to the relative redundancy. This is not true with local asynmptotics. There exist codes and a sequence of probability distributions P_n for which the sequence of relative redundancies tend to infinity but the usual redundancy tends to zero. In fact, the Lempel-Ziv code is such a code.
We will now formulate a corollary to Theorem B. The goal is to construct a limit theorem that describes local behavior:

- Fix n. Assume that $X_{i,n}$ is i.i.d. $P_n \in \mathcal{P}$.
- For any estimator $\hat{H}_n$, find $r(P_n, n) = |\hat{H}_n - H(P_n)|$.

Corollary B: Let $\mathcal{P}_n$ be a sequence of finite alphabet i.i.d. models satisfying a common bound on the entropy variance. Then for all n and all estimators, there exists a $P_n \in \mathcal{P}_n$ such that

$$|\hat{H}_n - H(P_n)| = \Omega_P(\frac{1}{\log n}).$$

The proof of the corollary follows immediately from the proof of Theorem B. The virtue of the local limit theorem allows us to restrict our model class to*finite* models not countable models.

3 LZ upper bound

In this section we consider upper bounds on the accuracy of entropy estimates based on string matching. We begin, as usual, with a sequence of observations X_1^n from an i.i.d.

probability distribution $P \in \mathcal{P}$ the set countable distributions with finite entropy and finite entropy variances. The longest matching prefix $L(n)$ is defined as follows:

$$L(n) = \max\{k : X_1^k = X_{j+1}^{j+k} \text{ for some } 1 \geq n - k\}$$

The usual match length estimate is

$$\hat{H}_{ML}(n) = \frac{\log n}{L(n)}.$$

Antos and Kontoyiannis demonstrated that

$$\hat{H}_{ML}(n) = H(P) + O_P(\frac{1}{\sqrt{\log n}}).$$

We will provide an alternative proof of this result. In the sequel,, we show that a slightly modified match length estimator $\hat{H'}_{ML}(n)$ satisfies

$$\hat{H'}_{ML}(n) = H(P) + O_P(\frac{1}{\log n}).$$

This will allow us to prove that the LZ algorithm is an optimal data compression algorithm for $\mathcal{P}$ (the simplest class of non-finite probability distributions).

Our analysis follows the usual construction. We begin by considering the renewal process whose i^{th} increment is $-\log P(X_i)$. The time of the k^{th} renewal is given by

$$S_k = \sum_{i=1}^{k} -\log P(X_i).$$

Let $R(a)$ be the number of renewals that occur before $a > 0$ with $T_a = R(a) + 1$. If $a = \log n$ then $T_{\log n}$ is the length k of shortest prefix X_1^k that has $P(X_1^k) < \frac{1}{n}$. This is a well studied random variable for which it is easy to show (using Walds theorem) that

$$ES_{T_a} = H(X)ET_a.$$

Since we desire notational simplicity let $T(n) = T_{\log n}$. Observe that we can write

$$ES_{T(n)} = \log n + E[S_{T(n)} - \log n].$$

The second term on the right hand side is the mean excess which asymptotically equals the second moment of the interarrival time divided by twice the mean (see Ross). Thus we have

$$ES_{T(n)} = \log n + \frac{E(\log P(X))^2}{2H(X)} + O(1).$$

Equating our expressions for $ES_{T(n)}$ leads to

$$H(X) = \frac{\log n}{ET(n)} + \frac{E(\log P(X))^2}{2H(X)ET(n)} + O(\frac{1}{ET(n)}). \tag{12}$$

This suggests that a reasonable estimate for the entropy $H(X)$ could be obtained by plugging in any estimate of $ET(n)$ that is at least $o_P(\frac{1}{\log n})$ accurate. Specifically, assume that we can find an estimate $\bar{T}_n$ for $ET(n)$ satisfying

$$|\bar{T}_n - ET(n)| = o_P(\frac{1}{\log n}).$$

Then $\hat{H}_n = \frac{\log n}{\bar{T}_n}$ satisfies

$$|\hat{H}_n - H(x)| = O_P(\frac{1}{\log n}).$$

Our first attempt at estimating $ET(n)$ is with the random variable $L(n)$ defined above. Plugging L into 12 in place of $ET(n)$ recovers $\hat{H}_{ML}(n)$. We shall show that $L(n))$ is in fact a good estimate for $ET(n)$, although we can improve it. Consider the event the set of sequences x_1^n for which $L(n) - T(n) > k$ for some integer $k > 0$. Let N_I be the (random) number of times the (random length) prefix X_1^I occurs in $X)1^n$. Observe that

$$\{L(n) - T(n) > k\} = \{N_{T(n)+k} > 0\}. \tag{13}$$

In plain language, this implies that the set of sequences for which the longest matching prefix is k greater than $T(n)$ is equal to the set of sequences for which prefix $X_1^{T(n)+k}$ has occurred at least once in X_1^n. The distribution of $\Delta(n) = L(n) - T(n)$ is well known at least for finite alphabet stationary mixing sources and of course Markov sources. For i.i.d. sequences on a countable space the same results hold. Specifically, there exists $0 < \beta < 1$ (dependent on P) such that for all n

$$\begin{aligned}
\Pr\{\Delta(n) > k\} &\leq \beta^k & (14)\\
\Pr\{\Delta(n) < -k\} &\leq Const. \exp(-\beta^k) & (15)\\
E\Delta(n) &= O(1). & (16)
\end{aligned}$$

We include a simple argument that proves the first inequality. It is easy to show using indicators that

$$EN_{T(n)+k} = nP(X_1^{T(n)+k}).$$

Conditioning on $X_1^{T(n)+k} = x_1^{t+k}$ and we have

$$\begin{aligned}
E[N_{T(n)+k}|X_1^{T(n)+k} = x_1^{t+k}] &= nP(x_1^t)P(x_{t+1}^{t+k}) & (17)\\
&\leq n\frac{1}{n}P(x_{t+1}^{t+k}) & (18)\\
&\leq \beta^k. & (19)
\end{aligned}$$

Equation 18 follows from the definition of $T(n)$ which implies that $P(x_1^t) \leq \frac{1}{n}$. Equation 19 follows by defining

$$\beta = \max_{a \in \mathcal{A}} P(a).$$

Equation 14 follows by applying the Markov inequality to 13.

What now remains is to bound the left tail of $\Delta(n)$ and its expectation. A simple argument will reveal the path of proof, for details we refer the reader to published work ([8]). As before we, observe that

$$E[N_{T(n)-k}|X_1^{T(n)+k} = x_1^{t-k}] = nP(x_1^{t-k}) = \frac{nP(x_1^{t-1})}{P(x_{t-1}^{t-k+1})}.$$

This easily leads to the upper bound on the expected number of occurrences of $X_1^{T(n)-k}$:

$$\begin{aligned} E[N_{T(n)-k}] &= EE[N_{T(n)-k}|X_1^{T(n)+k} = x_1^{t-k}] \\ &\leq \min_{a \in \mathcal{A}} \frac{1}{p(a)^{k-1}} \\ &= \beta^{-(k-1)}. \end{aligned}$$

This proves that if the prefixes that are k shorter than $T(n)$ are expected to occur an exponential number of times. It is not hard to show that this implies (15). Finally, since the right and left tails of $\Delta(n)$ vanish rapidly, it is follows that $E\Delta(n)$ is $O(1)$.

How good an estimate of $ET(n)$ is $L(n)$? We know that $L(n)$ is $O(1)$ close to $T(n)$. However, $T(n)$ is normal with mean $O(\log n)$ and standard deviation $O(\sqrt{\log n})$. Thus, plugging in $L(n)$ into (12) proves that $\hat{H}_{ML(n)}$ is only $O_P(1/\sqrt{\log n})$ accurate. To improve this we need only average. The obvious way to do this is to define

$$L_i(n) = \max\{k : X_i^{i+k-1} = X_j^{j+k-1} \text{ for some } j \in [1, n-k]\}$$

and then let

$$\hat{H'}_{ML}(n) = \frac{n \log n}{\sum_{i=1}^{n} L_i(n)}.$$

Because the sequence $L_i(n)$ is dependent it is tedious to attempt a variance calculation. An easier approach, is to break up X_1^n into $\sqrt{n}$ non-overlapping sequences. Then define $L_i(\sqrt{n})$ to be the longest matching prefix into into the i^{th} segment. Then define

$$\hat{H'}_{ML}(n) = \frac{\sqrt{n} \log \sqrt{n}}{\sum_{i=1}^{\sqrt{n}} L_i(\sqrt{n})}$$

Since $L_i(\sqrt{n})$ are independent the average is a $O(1)$ estimate for $ET(\sqrt{n})$. The resulting normalized estimate (since it uses $\sqrt{n}$ costs $\frac{2}{\log n}$. But the resulting estimate is $O_P(\frac{1}{\log n})$

accurate, which proves the following:

Theorem C: For any distribution $P \in \mathcal{P}$ and random variables $X_1, X_2, \ldots$ drawn i.i.d according to P the match length estimator $\hat{H'}_{ML}(n)$ defined above satisfies:

$$\hat{H'}_{ML}(n) = H(P) + O_P(\frac{1}{\log n}).$$

4 Optimality of the Plug in Estimator

We have shown that a match length estimator can achieve the lower bound of Theorem B to first order. It would be a strange universe indeed if the plug-in estimate was worse than the match length estimator for memoryless sources. In this section, we establish that the plug-in approach is as good as the match length approach. The basic idea is that by constraining the entropy and the entropy variance the tail probability cannot contribute to much to the estimate.

Our approach considers the performance of $H(\hat{F}_n)$ (the plug-in estimate) to the performance of an oracle estimator defined as follows. Let $\epsilon > 0$ be arbitrary let $\hat{p}(i)$ be the empirical estimate of $p(i)$ and let

$$\hat{H}(\epsilon) = -\sum_{i:p(i)>\epsilon} \hat{p}(i) \log \hat{p}(i). \tag{20}$$

Notice that $\hat{H}(\epsilon)$ can be computed only by an oracle that knows $p(i)$ for every i. In effect, the oracle estimator computes the plug-in estimate without the contribution of the rare symbols. Thus we have the elementary relation

$$H(\hat{F}_n) \geq \hat{H}(\epsilon)$$

for every n and every ϵ. Now our oracle estimate converges very nicely as $n \to \infty$ provided that ϵ does not tend to 0 too quickly. Again using our oracle, define

$$\underline{H}(\epsilon) = -\sum_{i:p(i)<\epsilon} p(i) \log p(i)$$

$$\bar{H}(\epsilon) = -\sum_{i:p(i)\geq\epsilon} p(i) \log p(i).$$

Clearly, $H = \underline{H}(\epsilon) + \bar{H}(\epsilon)$ for every ϵ. If we choose ϵ tending to zero not too quickly in n, say $\epsilon_n = \frac{1}{\sqrt{n}}$ then it follows easily that

$$\hat{H}(\epsilon_n) - \bar{H}(\epsilon_n) \to 0$$

uniformly over over P with a rate of convergence of $O_P(\frac{1}{\sqrt{n}})$.

Theorem D: For any distribution $P \in \mathcal{P}$ and random variables $X_1, X_2, \ldots$ drawn i.i.d according to P the plug in estimator $H(\hat{F}_n)$ defined above satisfies:

$$H(\hat{F}_n) = H(P) + O_P(\frac{1}{\log n}).$$

Our proof requires two very simple Markov inequalities.

Lemma D: Given any distribution P on the non-negative integers with finite entropy and entropy variance. Then for any $\epsilon > 0$ the following hold

$$\sum_{i:p_i<\epsilon} p_i \log \frac{1}{p_i} \leq \frac{Var(P)}{\log \frac{1}{\epsilon}} \tag{21}$$

$$\sum_{i:p_i<\epsilon} p_i \leq \frac{Var(P)}{(\log \frac{1}{\epsilon})^2} \tag{22}$$

Proof: First observe that

$$\sum_{i:p_i<\epsilon} p_i (\log \frac{1}{p_i})^2 \geq \sum_{i:p_i<\epsilon} p_i (\log \frac{1}{p_i})(\log \frac{1}{\epsilon}).$$

Inequality (21) now follows by noting that the left hand side above is upper bounded by $Var(P)$.

To prove (22) observe that

$$Var(P) \geq \sum_{i:p_i<\epsilon} p_i (\log \frac{1}{p_i})^2 \geq \sum_{i:p_i<\epsilon} p_i (\log \frac{1}{\epsilon})^2.$$

This proves Lemma D.

Proof of Theorem D: Now, if $\epsilon = 1/\sqrt{n}$ say, then

$$\sum_{i:p_i<\epsilon} p_i \log \frac{1}{p_i} \leq \frac{2Var(P)}{\log n} \tag{23}$$

$$\sum_{i:p_i<\epsilon} p_i \leq \frac{4Var(P)}{(\log n)^2} \tag{24}$$

By combining inequalities above we deduce that

$$H(\hat{F}_n) \geq \hat{H}(\epsilon_n) \geq H(P) - \frac{2Var(P)}{\log n} - O(n^{-\frac{1}{2}}).$$

To establish an upper bound, we note that $p_i \geq \frac{1}{n}$ for every symbol i and all n. Thus,

$$H(\hat{F}_n) \leq \sum_{i:p_i \geq \epsilon} \hat{p}_i \log \hat{p}_i + \sum_{i:p_i < \epsilon} \hat{p}_i \log n.$$

As before, with $\epsilon_n = n^{-\frac{1}{2}}$ it follows that

$$\sum_{i:p_i \geq \epsilon} \hat{p}_i \log \hat{p}_i \leq H(P) + O(n^{-\frac{1}{2}}).$$

Applying (24) and we have that

$$\sum_{i:p_i < \epsilon} \hat{p}_i \log n \leq \log n \frac{4Var(P)}{(\log n)^2} = \frac{4Var(P)}{\log n}.$$

Putting it together and we have the upper bound

$$H(\hat{F}_n) \leq H(P) + \frac{4Var(P)}{\log n} + O(n^{-\frac{1}{2}}).$$

This proves the theorem.

References

[1] Antos, A. and Kontoiannis, I. " Convergence Properties of Functional Estimates of Discrete Distributions", *Random Structures and Algorithms*, 2002.

[2] A.D. Wyner., J. Ziv and A.J. Wyner, "On The Role of Pattern Matching in Information Theory", IEEE Transactions on Information Theory, Vol. 44, no. 6, pp. 2045-2056, October, 1998.

[3] Kontoyiannis, I. , P.H. Algoet, Yu. M. Suhov and A.J. Wyner. "Non-parametric entropy estimation for stationary processes and random fields, with applications to English text" .IEEE Transactions Information Theory. Vol. IT-44, pp. 1319 - 1327, May, 1998.

[4] Foster, D., Stine, R. and Wyner, A.J. "Universal codes for finite sequences of integers frawn from a monotone distribution". *IEEE Trans. Inform. Theory,* IT-48:1713-1720, June 2002.

[5] Kieffer, J.C. and Yang, E.-H. " Grammar Based Codes: A new class of universal source codes".*IEEE Trans. Inform. Theory,* I=T-46: 737-754, May, 2000.

[6] Ziv, J,. and Lempel, A, "A Universal algorithm for sequential data compression". *IEEE Trans. Inform. Theory,*, IT-23: 337-343, 1977.

[7] Gyorfi and I. Pali. " There is no universal code for an infinite source alphabet". *IEEE Trans. Inform. Theory,* IT-40: 267-271, Janurary, 1994.

[8] Wyner, A.J. "More on Recurrence and Waiting Times ". The Annals of Applied Probability, Vol. 9, No. 3, pp. 780-796, 1999.

[9] A.D. Wyner and A.J. Wyner, "Improved Redundancy of a Version of the Lempel-Ziv Algorithm". *IEEE Transactions on Information Theory* , Vol. IT-41, pp. 723 731, May, 1995.

[10] Louchard, G. and Szpankowski, W., "On the Average Redundancy of the Lempel-Ziv Code",*IEEE Transactions on Information Theory*, Vol. 43, No. 1, 1-7, 1997.

[11] "Redundancy of the Lempel-Ziv Incremental Parsing Rule", Savari, S. , "Redundancy of the Lempel-Ziv Incremental Parsing Rule", *IEEE Transactions on Information Theory*, Vol. 43, No. 1, 1997.