Limit Results on Pattern Entropy

A. Orlitsky N.P. Santhanam K. Viswanathan J. Zhang

ECE Department, UCSD

{alon,nsanthan,kviswana,j6zhang}@ucsd.edu

December 27, 2004

Abstract

We determine the entropy rate of patterns of certain random processes, bound the speed at which the per-symbol pattern entropy converges to this rate, and show that patterns satisfy an asymptotic equipartition property. To derive some of these results we upper bound the probability that the n'th variable in a random process differs from all preceding ones.

1 Introduction

Most universal-compression applications involve sources, such as text, speech, or image, whose alphabets are very large, possibly even infinite. Yet as observed already by Davisson [1], as the alphabet size increases to infinity, so does the redundancy, the number of bits over the entropy, required because the distribution is not known in advance. In analyzing this phenomenon, Kieffer [2] showed that even *i.i.d.* distributions over infinite alphabets entail an infinite per-symbol redundancy and established a necessary and sufficient condition for a collection of sources to have a diminishing per-symbol redundancy.

Two approaches have addressed the high redundancy associated with large alphabets. One line of work [3]–[5] follows Elias [6] and considers compression of collections that satisfy Kieffer's condition. Results in this genre typically describe universal algorithms for such collections or find bounds on their redundancy. The most recent results [7] show that all collections satisfying Kieffer's condition can be compressed with diminishing per-symbol redundancy using grammar-based codes.

A second direction [8, 9] separates the description of strings over large alphabets into two parts: description of the symbols appearing in the string, and of their *pattern*, the order in which the symbols appear. For example, in text compression, this approach may separate the description of the order of the words from the specification of each word's binary representation.

Results along this line [10, 11] show that patterns of strings generated by *i.i.d.* distributions over any alphabet, even infinite or unknown, can be compressed with diminishing per-symbol redundancy. These results have also been used [12] to derive asymptotically-optimal solutions for the Good-Turing probability estimation problem. Related average case results have subsequently been proven [13].

It is therefore natural to consider the entropy of patterns, the number of bits needed to compress them when the underlying distribution is known. Shamir and Song [14], bounded the entropy of patterns of *i.i.d.* distributions in terms of the source entropy and alphabet size.

In this paper we determine the entropy rate of patterns of certain processes, bound the speed at which the per-symbol pattern entropy converges to this rate, and show that patterns satisfy an

Supported by the National Science Foundation and the Ericsson corporation.

asymptotic equipartition property. To derive some of these results, we upper bound the probability that the n'th variable in a random process differs from all preceding ones. We note that related entropy-rate results were independently derived by Gemelos and Weissman [15, 16].

The next section defines patterns and their entropy. Section 3 outlines the results and some of their implications. The proofs are provided in Sections 4 to 7.

2 Definitions

Let $\overline{x} = x_1, \ldots, x_n \in \mathcal{A}^n$ be a sequence of elements. The *index* $\imath(x)$ of x is one more than the number of distinct symbols preceding x's first appearance in $\overline{x}$. The *pattern of* $\overline{x}$ is the concatenation

$$\Psi(\overline{x}) \stackrel{\text{def}}{=} \imath_{\overline{x}}(x_1)\imath_{\overline{x}}(x_2)\ldots\imath_{\overline{x}}(x_n),$$

of all indices. For example, if $\overline{x} =$ "*abracadabra*", $\imath_{\overline{x}}(a) = 1$, $\imath_{\overline{x}}(b) = 2$, $\imath_{\overline{x}}(r) = 3$, $\imath_{\overline{x}}(c) = 4$, and $\imath_{\overline{x}}(d) = 5$, hence

$$\Psi(abracadabra) = 12314151231.$$

A distribution can be *discrete*, defined by a probability mass function, *continuous*, defined by a probability density function, or *mixed*, consisting of discrete and continuous parts. We allow for all three types of distributions and let q denote the total *continuous probability*. For example, in the mixed distribution where the value a occurs with probability $1/3$ and with the remaining $2/3$ probability a random value in the interval $[0, 1]$ occurs, say uniformly, the continuous probability is $q = 2/3$.

Every distribution p induces a distribution over patterns where

$$p(\overline{\psi}) \stackrel{\text{def}}{=} p(\{\overline{x} : \Psi(\overline{x}) = \overline{\psi}\}),$$

is the probability that a sequence generated according to p has pattern $\overline{\psi}$. For example, the *i.i.d.* distribution over $\{a, b\}$ where $p(a) = \alpha$ and $p(b) = \overline{\alpha}$ induces on length 2 patterns the distribution

$$p(11) = p(\{aa, bb\}) = \alpha^2 + \overline{\alpha}^2,$$
$$p(12) = p(\{ab, ba\}) = 2\alpha\overline{\alpha},$$

whereas the mixed distribution described above induces $p(11) = p(\{aa\}) = 1/9$ and $p(12) = p(\{xy : x \neq y\}) = 8/9$.

We denote a random n-symbol sequence by $\overline{X} = X_1, \ldots, X_n$ and its pattern by $\overline{\Psi} = \Psi_1, \ldots, \Psi_n$. The entropy of the sequence is

$$H(\overline{X}) = \sum_{\overline{x}} p(\overline{x}) \log \frac{1}{p(\overline{x})},$$

and its entropy rate is the asymptotic per-symbol entropy

$$\mathcal{H}_X = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}).$$

Similarly, the *pattern entropy* is

$$H(\overline{\Psi}) = \sum_{\overline{\psi}} p(\overline{\psi}) \log \frac{1}{p(\overline{\psi})},$$

and the *pattern entropy rate* is the asymptotic per-symbol entropy

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}).$$

These concepts are illustrated by the following examples.

Example 1. Consider the process $X_1, X_2, \ldots$ where $X_1 = 1$ and for $n = 2, 3, \ldots$, X_n is distributed uniformly over $\{X_{n-1} + 1, \ldots, X_{n-1} + n\}$. For example, 1,2,3,4 and 1,3,6,10 are two equally-likely realizations of $X_1, \ldots, X_4$. Since $X_1, \ldots, X_n$ can assume $n!$ equally likely realizations, the sequence entropy is

$$H(\overline{X}) = \log n!,$$

and its entropy rate is

$$\mathcal{H}_X = \lim_{n\to\infty} \frac{1}{n} \log n! = \infty.$$

On the other hand, $X_n > X_i$ for all $i = 1, \ldots, n-1$, hence the pattern is always $\overline{\Psi} = 12 \ldots n$, implying zero pattern entropy rate,

$$\mathcal{H}_\Psi = 0. \qquad \square$$

Example 2. Consider independent Bernoulli-half trials $X_1, X_2, \ldots$. As with all *i.i.d.* distributions,

$$H(\overline{X}) = nH(X_1),$$

hence the sequence entropy rate is

$$\mathcal{H}_X = H(X_1) = 1.$$

It is easy to verify that the resulting patterns are all 2^{n-1} sequences over $\{1, 2\}$ starting with 1. Each pattern corresponds to two possible trial sequences hence has probability $2^{-(n-1)}$. If follows that

$$H(\overline{\Psi}) = n - 1,$$

and the pattern entropy rate is

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{n-1}{n} = 1. \qquad \square$$

Note that in the last example,

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

One of the results we prove shows that this equality holds for all discrete *i.i.d.* distributions.

To place the obtained results in context, we briefly mention some existing sequence- and pattern-redundancy results. We use $\mathcal{R}_X$ to denote the *redundancy rate*, the limit of the per-symbol redundancy of distributions in a given class. For more formal redundancy definitions see, *e.g.,* [1].

As mentioned in the introduction, Kieffer [2] showed that *i.i.d.* distributions over infinite alphabets entail an infinite per-symbol redundancy,

$$\mathcal{R}_X = \infty, \tag{1}$$

while, as shown in [11], the patterns of such processes incur zero per-symbol redundancy,

$$\mathcal{R}_\Psi = 0. \tag{2}$$

These results suggest conveying a sequence $\overline{X}$ by first describing its pattern $\overline{\Psi}$ and then the dictionary Δ that maps $\{1, \ldots, \Psi_n\}$ to $\{X_1, \ldots, X_n\}$. For example, if $\overline{X}$ ="*abracadabra*", we can convey the pattern $\overline{\Psi} = 12314151231$ and the dictionary $\Delta(1) = a$, $\Delta(2) = b$, $\Delta(3) = r$, $\Delta(4) = c$, and $\Delta(5) = d$.

Since the pattern and dictionary determine the sequence, it is easy to see that

$$\mathcal{R}_\Psi + \mathcal{R}_{\Delta|\Psi} \geq \mathcal{R}_X.$$

Hence

$$\mathcal{R}_{\Delta|\Psi} = \infty. \tag{3}$$

Together, these results imply that for *i.i.d.* distributions over arbitrary alphabets, not knowing the underlying distribution results in infinite redundancy (1). Yet all the redundancy is associated with describing the dictionary (3), and none with the pattern (2). As the current results, outlined next, show, very different conclusions hold when the underlying distribution is known.

3 Results

In Section 4 we consider the *probability of innovation*, the probability ν_n that the nth element in a random process differs from all previous elements. We show that for all discrete stationary processes,

$$\nu_n \to 0$$

and that if, additionally, the process has finite marginal entropy H then

$$\nu_n \leq \frac{H}{\log n}(1 + o(1)).$$

While this bound is not tight for all stationary distributions, for example for an independent Bernoulli-half source, $\nu_n = 1/2^{n-1}$, we show that it is tight in the sense that for every positive H and ϵ there is an (*i.i.d.*) distribution with entropy H for which

$$\nu_n \geq \Omega\left(\frac{H}{(\log n)^{1+\epsilon}}\right).$$

In Section 5 we use these results to determine the pattern entropy rate of several distribution classes. We show that, as in Example 2, for all discrete *i.i.d.* processes and all discrete finite-entropy stationary processes, the pattern- and sequence-entropy rates coincide,

$$\mathcal{H}_\Psi = \mathcal{H}_X. \tag{4}$$

For *i.i.d.* distributions with continuous probability q, we show that the pattern entropy rate equals the entropy rate of a modified distribution,

$$\mathcal{H}_\Psi = \mathcal{H}_{\tilde{X}},$$

where $\tilde{X}$ is obtained from X by mapping all elements in the continuous part of the support to a single new discrete element. We note that similar entropy-rate results were independently derived by Gemelos and Weissman [15, 16].

In Section 6 we consider the speed

$$\rho_{X,n} \stackrel{\text{def}}{=} \left|\frac{1}{n}H(\overline{\Psi}) - \mathcal{H}_\Psi\right|$$

at which the per-symbol pattern entropy converges to its rate. For simplicity we consider only discrete *i.i.d.* distributions. We show that $\rho_{X,n}$ does not diminish uniformly for all processes or even for all distributions with a given entropy. We then bound $\rho_{X,n}$ in terms of

$$\sigma^2 \stackrel{\text{def}}{=} \sum p_i \log^2 p_i,$$

the second moment of the self information, showing that

$$\rho_{X,n} \leq \mathcal{O}\left(\frac{\sigma^4}{\log n}\right)^{1/3}.$$

In Section 7 we show that, like the original sequences, patterns of *i.i.d.* sequences satisfy an asymptotic equipartition property in that as the blocklength n increases, their probability tends to $2^{-n\mathcal{H}_\Psi}$. More precisely, we prove the convergence in probability,

$$\frac{1}{n}\log\frac{1}{p(\overline{\Psi})} \stackrel{\text{p}}{\to} \mathcal{H}_\Psi.$$

Two comparisons between these and existing pattern-compression results are in order. For simplicity, we describe them using discrete *i.i.d.* distributions.

First, while the original sequence and its pattern have the same (asymptotic per-symbol) entropy (4), the (asymptotic per-symbol) redundancy of the sequence is infinite (1) whereas that of the pattern is zero (2). Hence, when the distribution is known, describing the sequence and its pattern require the same number of bits, but when the distribution is not known, the sequence may require infinitely many additional bits whereas the pattern requires none.

Additionally, since (a) $\overline{X}$ determines $\overline{\Psi}$, and (b) given $\overline{\Psi}$ there is a 1-1 correspondence between $\overline{X}$ and Δ, we obtain

$$H(\overline{X}) \overset{(a)}{=} H(\overline{\Psi}) + H(\overline{X}|\overline{\Psi}) \overset{(b)}{=} H(\overline{\Psi}) + H(\Delta|\overline{\Psi}).$$

It follows that

$$\mathcal{H}_{\Delta|\Psi} \overset{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} H(\Delta|\overline{\Psi}) = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}) - \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}) = \mathcal{H}_X - \mathcal{H}_\Psi = 0. \tag{5}$$

Hence, while when the distribution is not known, essentially all the redundancy in describing a sequence derives from describing the dictionary (3) and none from the pattern (2), when the distribution is known, essentially all the bits go towards describing the pattern (4), and none towards the dictionary (5).

4 The probability of innovation

The essential difference between a sequence and its pattern is that the latter groups all hitherto unseen symbols into a single *new* element. For symbols that have been observed, the symbols and their indices in the pattern have 1-1 correspondence given the past sequence. To relate sequence and pattern entropy, we therefore show that for any discrete stationary distribution the probability of observing new elements decreases to zero with time. We begin with some definitions.

For $n \geq 1$, let $x^{n-1} = x_1, \ldots, x_{n-1}$ and let $\mathcal{A}(x^{n-1}) = \{x_1, \ldots, x_{n-1}\}$ be the set of elements observed in x^{n-1}. For a random process $X_1, X_2, \ldots$, let

$$I_n = \begin{cases} 1 & X_n \notin \mathcal{A}(X^{n-1}), \\ 0 & \text{otherwise.} \end{cases}$$

indicate whether the n'th symbol is new, and let

$$M_n \overset{\text{def}}{=} |\mathcal{A}(X^n)| = \sum_{i=1}^{n} I_i,$$

be the number of distinct symbols in $X_1, \ldots, X_n$. Finally. the *innovation probability* of the process at time n is

$$\nu_n \overset{\text{def}}{=} p(I_n = 1) = EI_n,$$

the probability that the nth symbol differs from all previous ones.

Since this section concerns only discrete distributions, assume without loss of generality that these strings are drawn from $\mathbb{N} = \{1, 2, \ldots\}$. For a stationary distribution, let $p_j \overset{\text{def}}{=} p(X_n = j)$ denote the marginal probability that the nth random variable is j. The distribution's *marginal entropy*,

$$H \overset{\text{def}}{=} \sum_{j=1}^{\infty} p_j \log \frac{1}{p_j}$$

is the entropy of each X_n.

The next lemma shows that for any stationary distribution the expected number of symbols grows sublinearly with n and provides a stronger bound for distributions with finite marginal entropy.

Lemma 1. For all discrete stationary distributions,

$$EM_n = o(n)$$

and if, in addition, the distribution has finite marginal entropy H, then,

$$EM_n \leq \frac{nH}{\log n}(1 + o(1)).$$

Proof For $j \in \mathbb{N}$, let

$$I_{n,j} = \begin{cases} 1 & X_n = j \notin \mathcal{A}(X^{n-1}) \\ 0 & \text{else,} \end{cases}$$

indicate whether X_n is new and equals j. Then,

$$I_n = \sum_{j=1}^{\infty} I_{n,j}.$$

For any function k_n of n,

$$M_n = \sum_{i=1}^{n} \sum_{j=1}^{k_n} I_{i,j} + \sum_{i=1}^{n} \sum_{j=k_n+1}^{\infty} I_{i,j}.$$

Since any element j can be new at most once,

$$\sum_{i=1}^{n} \sum_{j=1}^{k_n} I_{i,j} = \sum_{j=1}^{k_n} \sum_{i=1}^{n} I_{i,j} \leq \sum_{j=1}^{k_n} 1 = k_n,$$

and, since p_j denotes the probability that $X_n = j$,

$$E\left(\sum_{i=1}^{n} \sum_{j=k_n}^{\infty} I_{i,j}\right) = \sum_{i=1}^{n} \sum_{j=k_n+1}^{\infty} p(X_n = j, I_n = 1) \leq \sum_{i=1}^{n} \sum_{j=k_n+1}^{\infty} p_j = n \cdot \sum_{j=k_n+1}^{\infty} p_j.$$

Letting k_n increase to infinity as $o(n)$, we obtain

$$EM_n \leq k_n + n \cdot \sum_{j=k_n+1}^{\infty} p_j = o(n),$$

where the equality follows since $\sum_{j=k_n+1}^{\infty} p_j = o(1)$.

To prove the second part of the lemma, assume without loss of generality that the probabilities p_j are non-increasing. Then $p_j \leq \frac{1}{j}$ for all $j \geq 1$, and

$$\sum_{j=k_n+1}^{\infty} p_j < \frac{1}{\log k_n} \cdot \sum_{j=k_n+1}^{\infty} p_j \log j \leq \frac{1}{\log k_n} \cdot \sum_{j=k_n+1}^{\infty} p_j \log \frac{1}{p_j} \leq \frac{H}{\log k_n}.$$

Hence

$$EM_n \leq k_n + n \cdot \frac{H}{\log k_n},$$

and the lemma follows by letting

$$k_n = \frac{nH}{\log^2(nH)}.$$

□

In Corollary 3, we apply this lemma to show that the innovation probability of any stationary distribution diminishes with time, a result used in the next section to determine the entropy rate of patterns. We first show that the innovation probability of any stationary process decreases monotonically.

Lemma 2. For any stationary process,

$$\nu_n \geq \nu_{n+1}.$$

Proof For every stationary process and every n,

$$\begin{aligned}\nu_n = p\Big(X_n \notin \mathcal{A}(\{X_1,\ldots,X_{n-1}\}\Big) &= p\Big(X_{n+1} \notin \mathcal{A}(\{X_2,\ldots,X_n\})\Big) \\ &\geq p\Big(X_{n+1} \notin \mathcal{A}(\{X_1,\ldots,X_n\})\Big) = \nu_{n+1}.\end{aligned}$$

□

Corollary 3. For any discrete stationary process,

$$\lim_{n\to\infty} \nu_n = 0,$$

and if, in addition, the distribution has finite marginal entropy H, then for all n,

$$\nu_n \leq \frac{H}{\log n}(1 + o(1)).$$

Proof From Lemmas 1 and 2,

$$n\nu_n \leq \sum_{i=1}^{n} \nu_i = \sum_{i=1}^{n} EI_i = E\sum_{i=1}^{n} I_i = EM_n = o(n),$$

and if the distribution has finite marginal entropy H, then

$$n\nu_n \leq EM_n \leq \frac{nH}{\log n}(1 + o(1)).$$

□

For *i.i.d.* distributions, the last bound can be slightly improved.

Lemma 4. For all discrete *i.i.d.* distributions with finite entropy H and all n,

$$\nu_n \leq \frac{H}{\log n}.$$

Proof For every $0 < p < 1$ and $n \geq 1$, the Taylor series expansion of $\ln(1-x)$ yields

$$\ln\frac{1}{p} = -\ln(1-(1-p)) \geq \sum_{i=1}^{n-1} \frac{(1-p)^i}{i} \geq (1-p)^{n-1}\sum_{i=1}^{n-1}\frac{1}{i} \geq (1-p)^{n-1}\ln n.$$

Therefore,

$$\nu_n = \sum_{x\in\mathcal{A}} p(x)(1-p(x))^{n-1} \leq \frac{1}{\log n}\sum_{x\in\mathcal{A}} p(x)\log\frac{1}{p(x)} = \frac{H}{\log n}.$$

□

Note that while this bound is not tight for all *i.i.d.* distributions, for example the independent Bernoulli-half process has $\nu_n = H/2^{n-1}$, it is tight in the following sense.

Lemma 5. For all positive H and ϵ there is a distribution with entropy H and innovation probability

$$\nu_n = \Omega\Big(\frac{1}{(\log n)^{1+\epsilon}}\Big).$$

Proof We first show that for any $\epsilon > 0$, there is a finite entropy distribution with $\nu_n = \Omega\Big(\frac{1}{(\log n)^{1+\epsilon}}\Big)$. Define the probability distribution $(p_2, p_3, \ldots)$ by

$$p_i = \frac{1}{S\, i\,(\log i)^{2+\epsilon}},$$

where

$$S = \sum_{i \geq 2} \frac{1}{i(\log i)^{2+\epsilon}} < \infty \tag{6}$$

is a normalization factor. The distribution's entropy is

$$H^\epsilon = \sum_{i \geq 2} \frac{\log i + (2+\epsilon)\log\log i + \log S}{S\, i(\log i)^{2+\epsilon}} < \infty, \tag{7}$$

and, observing that $S > 1/2$, we obtain that for all $n \geq 4$,

$$\begin{aligned}
\nu_n &> \sum_{i \geq 2} (1-(n-1)p_i)p_i \\
&= \sum_{i \geq n} \frac{1}{S\, i(\log i)^{2+\epsilon}} - \sum_{i \geq n} \frac{n-1}{S^2\, i^2 (\log i)^{4+2\epsilon}} \\
&> \sum_{i \geq n} \frac{1}{S\, i(\log i)^{2+\epsilon}} - \sum_{i \geq n} \frac{1}{2S\, i(\log i)^{2+\epsilon}} \\
&= \Theta\Big(\frac{1}{(\log n)^{1+\epsilon}} \Big).
\end{aligned}$$

The distribution therefore has the desired innovation probability, and we now modify it to also have the required entropy H. The modification depends on whether H is larger or smaller than H^ϵ.

If $H > H^\epsilon$, consider the distribution $(p'_2, p'_3, \ldots)$ defined by

$$p'_i = \frac{1}{S\, i\, (\log i)^{2+\delta}}$$

where $0 < \delta \leq \epsilon$ and S is a normalization factor defined as before. Its entropy can be made arbitrarily large by decreasing δ and its innovation is

$$\nu_n = \Omega\Big(\frac{1}{(\log n)^{1+\delta}} \Big) = \Omega\Big(\frac{1}{(\log n)^{1+\epsilon}} \Big).$$

If $H < H^\epsilon$, consider the distribution $(p''_1, p''_2, \ldots)$ with $p''_1 = 1-q$ for some $0 < q < 1$ and

$$p''_i = \frac{q}{S\, i\, (\log i)^{2+\epsilon}}$$

for $i \geq 2$, where S is defined by (6). Its entropy is

$$h(q) + qH^\epsilon,$$

where H^ϵ is defined in (7). This entropy can be made equal to any value $0 < H < H^\epsilon$ by an appropriate choice of q. Clearly the new distribution also satisfies

$$\nu_n = \Omega\Big(\frac{1}{(\log n)^{1+\epsilon}} \Big). \qquad \square$$

5 The entropy rate of patterns

We determine the entropy rate of patterns of certain processes. We observe that when the alphabet is finite, the entropy rates of the process and its pattern coincide, and extend this result to all discrete processes that are either *i.i.d.*, or finite-entropy stationary. For *i.i.d.* distributions with a continuous component we show that the pattern entropy rate equals that of a modified process

where the continuous probability is assigned to a new discrete element. We note that similar results were independently obtained by Gemelos and Weissman [15, 16].

It is easy to see that whenever the alphabet $\mathcal{A}$ is finite, the process and pattern entropy rates coincide. Observe that

$$H(\overline{X}) - \log|\mathcal{A}|! \le H(\overline{\Psi}) \le H(\overline{X}), \tag{8}$$

where the upper bound follows as the sequence determines the pattern, and the lower bound follows as, for the same reason,

$$H(\overline{X}) = H(\overline{\Psi}) + H(\overline{X}|\overline{\Psi}) \tag{9}$$

and every pattern can derive from at most $|\mathcal{A}|!$ sequences, hence

$$H(\overline{X}|\overline{\Psi}) \le \log|\mathcal{A}|!.$$

Taking limits in (8) we see that for all distributions over finite alphabets,

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}) = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}) = \mathcal{H}_X. \tag{10}$$

Note that for *i.i.d.* processes, bounds similar to (8) appeared in [14]. The lower bound therein is somewhat weaker than that in (8) while the upper bound is somewhat stronger when $|\mathcal{A}| = o(n)$ and all probabilities are at least $1/n^{1+\epsilon}$ for some $\epsilon > 0$. The upper bound was further improved in [17].

The rest of the section extends (10) to distributions over infinite alphabets. We use the following lemma relating conditional pattern entropy and the pattern entropy rate.

Lemma 6. For any process, if

$$H(\Psi_n|\Psi^{n-1}) \ge h_n.$$

and

$$\lim_{n\to\infty} h_n = \mathcal{H}_X,$$

then

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof Since X^n determines Ψ^n and $H(\Psi^n) = \sum_{i=1}^n H(\Psi_i|\Psi^{i-1})$,

$$\frac{1}{n} H(X^n) \ge \frac{1}{n} H(\Psi^n) \ge \frac{1}{n} \sum_{i=1}^n h_i.$$

Taking limits as $n \to \infty$, the lemma follows because Cesáro's mean theorem implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n h_j = \mathcal{H}_X. \qquad \square$$

We begin with *i.i.d.* distributions, and among them start with those over discrete alphabets. We show that a random sequence is likely to contain all high-probability elements, and that when this happens, the conditional entropy of the pattern approaches that of the sequence.

As in Section 4 we assume without loss of generality that the alphabet is $\mathbb{N} = \{1, 2, \ldots\}$ and let $p_i \stackrel{\text{def}}{=} p(X_n = i)$. For $\epsilon \ge 0$, we let

$$A_\epsilon \stackrel{\text{def}}{=} \{i : p_i > \epsilon\}$$

be the set of all elements whose probability exceeds ϵ.

Theorem 7. For all discrete *i.i.d.* distributions,

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof We first show that a random sequence is likely to contain all elements of sufficiently high probability. More precisely, recall that $\mathcal{A}(X^n)$ is the set of all elements in X^n, and that $A_{\frac{\ln n}{n}}$ is the set of all elements whose probability exceeds $\frac{\ln n}{n}$. Clearly, $|A_{\frac{\ln n}{n}}| \leq \frac{n}{\ln n}$, hence

$$p\Big(A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(X^n)\Big) > 1 - \frac{n}{\ln n}\left(1 - \frac{\ln n}{n}\right)^n > 1 - \frac{1}{\ln n}.$$

Let

$$J_n = \begin{cases} 1 & A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(X^n) \\ 0 & \text{otherwise} \end{cases}$$

indicate whether X^n contains all high-probability elements. Then

$$\begin{aligned} H(\Psi_{n+1}|\Psi^n) &\geq H(\Psi_{n+1}|X^n) \\ &\geq H(\Psi_{n+1}|X^n, J_n) \\ &\geq p(J_n = 1) H(\Psi_{n+1}|X^n, J_n = 1) \\ &\geq \left(1 - \frac{1}{\ln n}\right) \sum_{i \in A_{\frac{\ln n}{n}}} p_i \log \frac{1}{p_i} \\ &\stackrel{\text{def}}{=} \left(1 - \frac{1}{\ln n}\right) H(A_{\frac{\ln n}{n}}). \end{aligned}$$

The theorem follows from Lemma 6 as

$$\lim_{n\to\infty} H(A_{\frac{\ln n}{n}}) = \mathcal{H}_X. \qquad \square$$

For mixed *i.i.d.* distributions we show that the entropy rate of the pattern equals that of a slightly modified process. Let X be a random variable drawn from a mixed distribution p with discrete support A_0 and continuous probability q. Define $\tilde{X}$ to be the discrete random variable obtained from X by replacing all elements in the continuous support with a single new discrete element. Then

$$\mathcal{H}_{\tilde{X}} = \sum_{i \in A_0} p_i \log \frac{1}{p_i} + q \log \frac{1}{q} = H(A_0) + q \log \frac{1}{q}.$$

Theorem 8. For all *i.i.d.* distributions,

$$\mathcal{H}_\Psi = \mathcal{H}_{\tilde{X}}.$$

Proof Since $\tilde{X}^n$ too determines Ψ^n, we proceed as in Theorem 7. Recall the definitions of J_n, $A_{\frac{\ln n}{n}}$, and $H(A_{\frac{\ln n}{n}})$, and let $\mathcal{A}(\overline{x})^c$ denote the set of symbols not in $\overline{x}$,

$$\begin{aligned} H(\Psi_{n+1}|\Psi^n) &\geq p(J_n = 1) H(\Psi_{n+1}|\tilde{X}^n, J_n = 1) \\ &\geq \left(1 - \frac{1}{\ln n}\right)\left(H(A_{\frac{\ln n}{n}}) + \min_{\overline{x}: A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(\overline{x})} p(\mathcal{A}(\overline{x})^c) \log \frac{1}{p(\mathcal{A}(\overline{x})^c)}\right). \end{aligned}$$

The theorem follows by applying Lemma 6 to $\tilde{X}^n$ as

$$\lim_{n\to\infty} H(A_{\frac{\ln n}{n}}) = H(A_0),$$

and

$$\lim_{n\to\infty} \min_{\overline{x}: A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(\overline{x})} p(\mathcal{A}(\overline{x})^c) = q. \qquad \square$$

We now address stationary processes. Note that while Theorem 7 shows that $\mathcal{H}_\Psi = \mathcal{H}_X$ for all discrete *i.i.d.* processes, even those with infinite entropy, as the next example indicates, this equality cannot hold for all discrete stationary processes with infinite entropy.

Example 3. Consider the constant stationary process $X_1 = X_2 = \dots$ defined by

$$p_j = p(X_n = j) = \frac{1}{S}\frac{1}{j \log^2 j},$$

where S is a normalization factor. Then,

$$H(X_1) = \sum_{j=1}^{\infty} p_j \log \frac{1}{p_j} = \infty,$$

hence

$$\mathcal{H}_X = \infty.$$

On the other hand, the pattern is always $11\dots1$, hence

$$\mathcal{H}_\Psi = 0. \qquad \square$$

To prove that $\mathcal{H}_\Psi = \mathcal{H}_X$ for all discrete stationary processes with finite entropy, we use the innovation results of Section 4. We show that the probability that X_n is new, hence more "informative" than Ψ_n, is low for likely X^{n-1}, and that when X_n is not new, the conditional entropy of the pattern is roughly $\mathcal{H}_X$.

Theorem 9. For all finite-entropy discrete stationary processes,

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof As before, we lower bound the conditional pattern entropy with a term that approaches $\mathcal{H}_X$. We show that

$$H(\Psi_n|\Psi^{n-1}) \geq H(X_n|X^{n-1}) - o(1),$$

and the theorem will follow from Lemma 6 as for all finite-entropy stationary processes,

$$\lim_{n\to\infty} H(X_n|X^{n-1}) = \mathcal{H}_X.$$

Recall that for $n \geq 1$, I_n indicates whether X_n is new, hence

$$\begin{aligned}
H(X_n|X^{n-1}) &= H(X_n, I_n|X^{n-1}) \\
&= H(I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n) \\
&= H(I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 0)p(I_n = 0) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(I_n|X^{n-1}) + H(\Psi_n|X^{n-1}, I_n) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(\Psi_n, I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(\Psi_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&\leq H(\Psi_n|\Psi^{n-1}) + H(X_n|I_n = 1)p(I_n = 1)
\end{aligned}$$

We now use Corollary 3 to show that

$$H(X_n|I_n = 1)p(I_n = 1) = o(1).$$

Recall that $p_j = p(X_n = j)$, that

$$A_{\nu_n} = \{j : p_j > \nu_n\}$$

is the set of all elements whose probability exceeds ν_n, that $\nu_n = p(I_n = 1)$, and that $I_{n,j}$ indicates whether X_n is new and equals j. Define

$$\nu_{n,j} \stackrel{\text{def}}{=} p(I_{n,j} = 1) = \sum_{x^{n-1}: j \notin x^{n-1}} p(x^{n-1}, x_n = j),$$

so that $\nu_n = \sum_{j=1}^{\infty} \nu_{n,j}$. Then

$$\begin{aligned}
H(X_n|I_n = 1)p(I_n = 1) &= \sum_{j=1}^{\infty} \nu_{n,j} \log \frac{\nu_n}{\nu_{n,j}} \\
&= \nu_n \log \nu_n + \sum_{j \in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} + \sum_{j \notin A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} \\
&\stackrel{(a)}{\le} \nu_n \log (|A_{\nu_n}| + 1) + \sum_{j \notin A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} \\
&\stackrel{(b)}{\le} \nu_n \log \left(\frac{1}{\nu_n} + 1 \right) + \sum_{j \notin A_{\nu_n}} p_j \log \frac{1}{p_j} \\
&\stackrel{(c)}{=} o(1),
\end{aligned}$$

where (a) follows because

$$\begin{aligned}
\sum_{j \in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} &\le \sum_{j \in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} + (\nu_n - \sum_{j \in A_{\nu_n}} \nu_{n,j}) \log \frac{1}{\nu_n - \sum_{j \in A_{\nu_n}} \nu_{n,j}} \\
&\le \nu_n \log (|A_{\nu_n}| + 1) + \nu_n \log \frac{1}{\nu_n},
\end{aligned}$$

(b) follows as $|A_{\nu_n}| < \frac{1}{\nu_n}$ and $\nu_{n,j} \le p_j \le \nu_n$, which for sufficiently large n is smaller than $\frac{1}{e}$, and (c) follows as Corollary 3 implies that $\nu_n \to 0$, and $\mathcal{H}_X < \infty$ implies that the marginal entropy is finite, hence

$$\lim_{n \to \infty} \sum_{j \notin A_{\nu_n}} p_j \log \frac{1}{p_j} = 0. \qquad \square$$

6 The rate of convergence

In the previous section we determined the pattern entropy rate—the limit of the per-symbol pattern entropy—of *i.i.d.* and certain related distributions. We now address the *convergence rate*

$$\rho_{X,n} \stackrel{\text{def}}{=} \left| \frac{1}{n} H(\overline{\Psi}) - \mathcal{H}_\Psi \right|$$

at which this limit is attained. For simplicity we consider only discrete *i.i.d.* distributions. Then

$$\rho_{X,n} = \frac{1}{n}(H(\overline{X}) - H(\overline{\Psi})) = \frac{1}{n} H(\overline{X}|\overline{\Psi}),$$

where the first equality follows from Theorem 7, and the second from (9).

We first show that $\rho_{X,n}$ does not diminish uniformly for all distributions, or even for all distributions with a given entropy. We then bound $\rho_{X,n}$ in terms of the second moment of the self information.

To show that $\rho_{X,n}$ does not diminish uniformly, the next example shows that it can be made arbitrarily high for all n.

Example 4. The *i.i.d.* process $X_1, X_2, \ldots$, where each X_i is distributed uniformly over $\{1, \ldots, k\}$, has

$$\rho_{X,n} = \frac{1}{n} H(\overline{X}|\overline{\Psi}) \geq \frac{1}{n} H(X_1|\overline{\Psi}) = \frac{\log k}{n},$$

which can be made arbitrarily high by choosing a sufficiently large k. □

While the example shows that $\rho_{X,n}$ does not diminish uniformly for all *i.i.d.* distributions (and in fact is unbounded), the processes it uses have unbounded entropy themselves. It is natural to ask whether $\rho_{X,n}$ diminishes uniformly for all *i.i.d.* processes with a given entropy. The next example answers this question in the negative, showing that for all n, $\rho_{X,n}$ can be made arbitrarily close to the process entropy.

Example 5. For every k, let $p^k = (1-q, q/k, \ldots, q/k)$ be the distribution over $k+1$ elements where one element has probability $1-q$ and each of the remaining k elements has probability q/k. Given $H > 0$, for any $k \geq 2^H$ there is a q such that the $H(p^k) = h(q) + q \log k = H$, and as k increases to infinity, q tends to zero, hence $q \log k \to H$.

Let each of $X_1, X_2, \ldots$ be distributed independently according to p^k. For any fixed blocklength n, as k tends to infinity, the probability that any element of probability q/k appears more than once decreases to 0, hence with high probability there is a 1-1 correspondence between the pattern and the set of locations where the element of probability $1-q$ appears. Consequently $H(\overline{\Psi}) \to nh(q)$.

It follows that for any fixed n, as k increases,

$$\rho_{X,n} = \frac{1}{n} H(\overline{X}|\overline{\Psi}) = \frac{1}{n}\big(H(\overline{X}) - H(\overline{\Psi})\big) \to q \log k \to H.$$ □

In the preceding examples we increased $\rho_{X,n}$ by constructing successively flatter distributions, raising the possibility that $\rho_{X,n}$ will diminish when the distribution $p_1, p_2, \ldots$ diminishes to 0 sufficiently quickly. In Theorem 11 and Corollary 12 we bound $\rho_{X,n}$ in terms of the second moment of the self information

$$\sigma^2 \stackrel{\text{def}}{=} \sum_{i \geq 1} p_i \log^2 \frac{1}{p_i}.$$

To do so, we first prove the following technical lemma.

Lemma 10. For any discrete distribution p, and all $I \geq 1$,

$$\sum_{i \geq I} p_i \log \frac{1}{p_i} \leq \sigma \sqrt{\sum_{i \geq I} p_i}.$$

Proof Using the Cauchy-Schwartz Inequality,

$$\sum_{i \geq I} p_i \log \frac{1}{p_i} = \sum_{i \geq I} \sqrt{p_i \cdot p_i \log^2 \frac{1}{p_i}} \leq \left(\sum_{i \geq I} p_i\right)^{1/2} \left(\sum_{i \geq I} p_i \log^2 \frac{1}{p_i}\right)^{1/2}.$$ □

Theorem 11. For all discrete *i.i.d.* distributions with entropy H,

$$H\left(1 - \Theta\left(\frac{\sigma^2}{H \log n}\right)^{1/3}\right) \leq \frac{1}{n} H(\overline{\Psi}) \leq H.$$

Proof Let

$$\epsilon_n \stackrel{\text{def}}{=} \left(\frac{2H^2}{\sigma \log n}\right)^{2/3}$$

and

$$T_n \stackrel{\text{def}}{=} \left\{ x^{n-1} : p(I_n = 1|x^{n-1}) = \sum_{x \notin \mathcal{A}(\overline{x})} p(x) \le \epsilon_n \right\},$$

be the set of strings whose missing mass is at most ϵ_n. From Lemma 10,

$$\min_{x^{n-1} \in T_n} H(\Psi_n|x^{n-1}) \ge H - \sigma\sqrt{\epsilon_n} = H\left(1 - 2\left(\frac{\sigma^2}{4H\log n}\right)^{1/3}\right) \stackrel{\text{def}}{=} H(1 - 2\delta_n).$$

Note that $E_{X^{n-1}} p(I_n = 1|X^{n-1}) = p(I_n = 1) = \nu_n$, hence from Markov's inequality and Lemma 4,

$$p(T_n) \ge 1 - \frac{H}{\epsilon_n \log n} = 1 - \left(\frac{\sigma^2}{4H\log n}\right)^{1/3} = 1 - \delta_n.$$

It follows that for $n \ge 2$,

$$H(\Psi_n|\Psi^{n-1}) \ge H(\Psi_n|X^{n-1}) \ge \sum_{x^{n-1} \in T_n} p(x^{n-1}) H(\Psi_n|x^{n-1}) \ge H(1 - 3\delta_n).$$

Hence

$$\frac{1}{n} H(\Psi^n) \ge \frac{n-1}{n} H - \frac{3}{n} H \sum_{i=2}^{n} \delta_i = H - \Theta(H\delta_n) = H - \Theta\left(\frac{\sigma^2 H^2}{\log n}\right)^{1/3}. \qquad \square$$

Lemma 10 implies that

$$H \le \sigma,$$

hence the rate of convergence of pattern entropy can be bounded as follows.

Corollary 12. For all discrete *i.i.d.* distributions,

$$\rho_{X,n} \le \mathcal{O}\left(\frac{\sigma^4}{\log n}\right)^{1/3}. \qquad \square$$

7 Asymptotic equipartition of patterns

Shannon [18] showed that strings generated by *i.i.d.* distributions over finite alphabets satisfy an asymptotic equipartition property. Chung [19] generalized this result to infinite alphabets. We prove an equivalent property for patterns of such strings. Specifically, we show that

$$\frac{1}{n} \log \frac{1}{p(\overline{\Psi})} \stackrel{\text{p}}{\to} \frac{1}{n} E \log \frac{1}{p(\overline{\Psi})}, \tag{11}$$

where the convergence is in probability, uniformly over all *i.i.d.* distributions, see Theorem 16. Since by definition,

$$\frac{1}{n} E \log \frac{1}{p(\overline{\Psi})} \;\to\; \mathcal{H}_\Psi,$$

we obtain

$$\frac{1}{n} \log \frac{1}{p(\overline{\Psi})} \stackrel{\text{p}}{\to} \mathcal{H}_\Psi,$$

though here, the results of the last section show that we cannot have uniform convergence over all *i.i.d.* distributions. To prove (11) we use *profiles* of patterns, defined next.

The *multiplicity* of $\psi \in \mathbb{Z}^+$ in a pattern $\overline{\psi}$ is

$$\mu_\psi \stackrel{\text{def}}{=} |\{1 \le i \le |\overline{\psi}| : \psi_i = \psi\}|,$$

the number of times ψ appears in $\overline{\psi}$. The *prevalence* of a multiplicity $\mu \in \mathbb{N}$ in $\overline{\psi}$ is

$$\varphi_\mu \stackrel{\text{def}}{=} |\{\psi : \mu_\psi = \mu\}|,$$

the number of symbols appearing μ times in $\overline{\psi}$. The *profile* of $\overline{\psi}$ is

$$\overline{\varphi} \stackrel{\text{def}}{=} \left(\varphi_1, \ldots, \varphi_{|\overline{\psi}|}\right)$$

the vector of prevalences of all possible multiplicities for $1 \le \mu \le |\overline{\psi}|$. For example, the pattern $\psi = 12131$ has multiplicities $\mu_1 = 3$, $\mu_2 = \mu_3 = 1$, and $\mu_\psi = 0$ for all other $\psi \in \mathbb{Z}^+$. Hence its prevalences are $\varphi_1 = 2$, $\varphi_2 = 0$, $\varphi_3 = 1$, $\varphi_4 = \varphi_5 = 0$, and its profile is $\varphi(\psi) = (2, 0, 1, 0, 0)$.

If p is an *i.i.d.* distribution, then all length-n patterns $\overline{\psi}$ with profile $\overline{\varphi}$, have the same probability,

$$p(\overline{\psi}) = \frac{p(\overline{\varphi})}{N(\overline{\varphi})},$$

where

$$N(\overline{\varphi}) = \frac{n!}{\prod_\mu \mu!^{\varphi_\mu} \varphi_\mu!}$$

is the number of patterns with profile $\overline{\varphi}$. Therefore

$$\log \frac{1}{p(\overline{\psi})} = \log \frac{1}{p(\overline{\varphi})} + \log N(\overline{\varphi}).$$

Let $\overline{\Phi}$ denote the profile of a random sequence $\mathcal{X}$. The following bound by McDiarmid can be used to show that $\log N(\overline{\Phi})$ concentrates around its mean.

Lemma 13. [McDiarmid [20]] Let $\overline{X} = X_1, \ldots, X_n$ be independent random variables and let the function $f(x_1, \ldots, x_n)$ be such that any change in a single x_i changes $f(x_1, \ldots, x_n)$ by at most η. Then,

$$p\left\{ \left| f(\overline{X}) - Ef(\overline{X}) \right| > \eta \sqrt{\frac{n \ln \frac{2}{\delta}}{2}} \right\} < \delta. \qquad \square$$

Corollary 14. For all $\alpha > 0$,

$$p\Big\{ \left| \log N(\overline{\Phi}) - E \log N(\overline{\Phi}) \right| > 3n^{\frac{1+\alpha}{2}} \log n \Big\} < \frac{2}{e^{2n^\alpha}}.$$

Proof Let $f(x_1, \ldots, x_n) = \log N(\overline{\varphi})$. A change in x_i can change $\log \prod \varphi_\mu!$ by at most $2 \log n$, and $\log \prod \mu!^{\varphi_\mu}$ by at most $\log n$. The corollary follows by setting $\delta = \frac{2}{e^{2n^\alpha}}$ in Lemma 13. $\square$

We now show that with high probability, the profile self-information deviates from its expectation by at most roughly $n^{\frac{1+\alpha}{2}} \log n$.

Lemma 15. For all $\alpha > 0$,

$$p\left\{ \left| \log \frac{1}{p(\overline{\Phi})} - H(\overline{\Phi}) \right| \ge \left(\pi \sqrt{\frac{2}{3}} \log e \right) n^{\frac{1+\alpha}{2}} \log n \right\} \le \frac{\exp\left(\pi \sqrt{\frac{2n}{3}} \right)}{\exp\left(\pi \sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}} \log n \right)}.$$

Proof Let $\rho(n)$ be the number of profiles of length-n patterns. Then the entropy of $\overline{\Phi}$ can be bounded by

$$E \log \frac{1}{p(\overline{\Phi})} = H(\overline{\Phi}) \le \log \rho(n) \le \left(\pi \sqrt{\frac{2}{3}} \log e \right) \sqrt{n}.$$

where the second inequality follows as $\rho(n)$ is, see *e.g.,* [11], the number of integer partitions of n, which has been computed by Hardy and Ramanujan [21].

Let $\ell = \left(\pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{1+\alpha}{2}} \log n$. Since $\ell \geq H(\overline{\Phi})$,

$$\left|\log\frac{1}{p(\overline{\Phi})} - H(\overline{\Phi})\right| \geq \ell \Rightarrow \log\frac{1}{p(\overline{\Phi})} \geq \ell,$$

hence

$$p\left\{\left|\log\frac{1}{p(\overline{\Phi})} - H(\overline{\Phi})\right| \geq \ell\right\} \leq p\left\{\log\frac{1}{p(\overline{\Phi})} \geq \ell\right\} \leq \frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}} \log n\right)},$$

where the last inequality follows as the probability of any profile with self-information $\geq \ell$ is at most $2^{-\ell}$ and there can be at most $\rho(n) \leq \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$ such profiles. □

Corollary 14 and Lemma 15 imply the asymptotic equipartition property. Note that the convergence bound is uniform for all *i.i.d.* distributions.

Theorem 16. For all $\delta > 0$,

$$p\left\{\frac{1}{n}\left|\log\frac{1}{p(\overline{\Psi})} - H(\overline{\Psi})\right| \geq \delta\right\} = \exp\left(-\Omega\left(\frac{n\delta^2}{\log^2 n}\right)\right).$$

Proof Observe that $H(\overline{\Psi}) = E\log\frac{1}{p(\overline{\Psi})}$, and that

$$\begin{aligned}
&p\left\{\frac{1}{n}\left|\log\frac{1}{p(\overline{\Psi})} - E\log\frac{1}{p(\overline{\Psi})}\right| \leq \delta\right\} \\
&\qquad \geq p\left\{\frac{1}{n}\left|\log N(\overline{\Phi}) - E\log N(\overline{\Phi})\right| + \frac{1}{n}\left|\log\frac{1}{p(\overline{\Phi})} - E\log\frac{1}{p(\overline{\Phi})}\right| \leq \delta\right\} \\
&\qquad \geq p\left\{\left\{\frac{1}{n}\left|\log N(\overline{\Phi}) - E\log N(\overline{\Phi})\right| \leq 3n^{\frac{\alpha-1}{2}}\log n\right\}\right. \\
&\qquad\qquad \left.\bigcap\left\{\frac{1}{n}\left|\log\frac{1}{p(\overline{\Phi})} - E\log\frac{1}{p(\overline{\Phi})}\right| \leq \left(\pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{\alpha-1}{2}}\log n\right\}\right\} \\
&\qquad \geq 1 - \frac{2}{e^{2n^\alpha}} - \frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}}\log n\right)},
\end{aligned}$$

where for sufficiently large n, $0 < \alpha \leq 1$, is the solution of

$$\left(3 + \pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{\alpha-1}{2}}\log n = \delta.$$

The last inequality follows from Lemmas 14 and 15. Clearly,

$$n^\alpha = \frac{n\,\delta^2}{\left(3 + \pi\sqrt{\frac{2}{3}}\log e\right)^2 \log^2 n},$$

and the theorem follows by observing that the $2e^{-2n^\alpha}$ term dominates. □

Acknowlegdements

We thank Ian Abramson and Tsachy Weissman for helpful discussions.

References

[1] L.D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783—795, November 1973.

[2] J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674—682, November 1978.

[3] L. Györfi, I. Pali, and E.C. Van der Meulen. On universal noiseless source coding for infinite source alphabets. *European Transactions on Telecommunications and Related Technologies*, 4:125—132, 1993.

[4] D.P. Foster, R.A. Stine, and A.J. Wyner. Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Transactions on Information Theory*, 48(6):1713—1720, June 2002.

[5] T. Uyematsu and F. Kanaya. Asymptotic optimality of two variations of Lempel-Ziv codes for sources with countably infinite alphabet. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

[6] P. Elias. Universal codeword sets and representations of integers. *IEEE Transactions on Information Theory*, 21(2):194—203, March 1975.

[7] D. He and E Yang. On the universality of grammar-based codes for sources with countably infinite alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2003.

[8] J. Åberg, Y.M. Shtarkov, and B.J.M. Smeets. Multialphabet coding with separate alphabet description. In *Proceedings of Compression and Complexity of Sequences*, 1997.

[9] N. Jevtić, A. Orlitsky, and N.P. Santhanam. Universal compression of unknown alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

[10] A. Orlitsky and N.P. Santhanam. Performance of universal codes over infinite alphabets. In *Proceedings of the Data Compression Conference*, March 2003.

[11] A. Orlitsky, N.P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469—1481, July 2004.

[12] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427—431, October 17 2003. See also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003.

[13] G. Shamir. Universal lossless compression with unknown alphabets—the average case. Submitted for publication, IEEE Transactions on Information Theory, 2003.

[14] G. Shamir and L. Song. On the entropy of patterns of *i.i.d.* sequences. In *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, pages 160—170, October 2003.

[15] G. Gemelos and T. Weissman. On the entropy rate of pattern processes. Technical Report HPL-2004-159, HP Labs, September 2004.

[16] G. Gemelos and T. Weissman. Submitted for publication, Data Compression Conference, November 2004.

[17] G. Shamir. Sequence patterns, entropy, and infinite alphabets. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, October 2004.

[18] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379—423, 623—656, 1948.

[19] K.L. Chung. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612—614, 1961.

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

[21] G.H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75—115, 1918.