

Limit Results on Pattern Entropy

A. Orlitsky, N.P. Santhanam, K. Viswanathan, J. Zhang
ECE Department, UCSD
{alon, nsanthan, kviswana, j6zhang}@ucsd.edu

Abstract — We determine the entropy rate of patterns of *i.i.d.* strings and show that they satisfy an asymptotic equipartition property.

I. 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 large 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 separates the description of the order of the words from the specification of each word’s binary representation.

Results along this line show [10, 11] 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.* sequences in terms of the source entropy and alphabet size.

In this paper we determine the entropy rate of patterns of *i.i.d.* distributions and show that they satisfy an asymptotic equipartition property.

In Section III we prove that for discrete distributions the entropy rate of patterns equals that of the distribution, and

that for distributions with continuous probability q , defined in the next section, the entropy rate of patterns equals that of a modified distribution where the continuous probability is assigned to a new discrete element. One implication of these results is that for discrete distributions the conditional entropy rate of the sequence when its pattern is known is zero.

In Section IV we strengthen these results and show that patterns satisfy an asymptotic equipartition property. Namely, that for any source with pattern entropy rate H , as the blocklength n increases to infinity, random patterns have probability close to 2^{-nH} .

In this abstract we address only distributions with finite entropy. The corresponding infinite-entropy results will be proven in the complete version of the paper.

II. DEFINITIONS

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

$$\Psi(\bar{x}) \stackrel{\text{def}}{=} \iota_{\bar{x}}(x_1)\iota_{\bar{x}}(x_2)\dots\iota_{\bar{x}}(x_n),$$

of all indices. For example, if $\bar{x} = \text{“abracadabra”}$, $\iota_{\bar{x}}(a) = 1$, $\iota_{\bar{x}}(b) = 2$, $\iota_{\bar{x}}(r) = 3$, $\iota_{\bar{x}}(c) = 4$, and $\iota_{\bar{x}}(d) = 5$, hence

$$\Psi(\text{abracadabra}) = 12314151231.$$

We let Ψ^n denote the collection of length- n patterns. For example, $\Psi^1 = \{1\}$, $\Psi^2 = \{11, 12\}$, and $\Psi^3 = \{111, 112, 121, 122, 123\}$.

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

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

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

$$\begin{aligned} p(11) &= p(\{aa, bb\}) = \alpha^2 + \bar{\alpha}^2, \\ p(12) &= p(\{ab, ba\}) = 2\alpha\bar{\alpha}, \end{aligned}$$

¹Supported by the NSF and the Ericsson corporation.

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 pattern by Ψ_1, \dots, Ψ_n . Its *entropy* is

$$H(\Psi_1, \dots, \Psi_n) = \sum_{\bar{\psi} \in \Psi^n} p(\bar{\psi}) \log \frac{1}{p(\bar{\psi})},$$

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

$$H_\Psi = \lim_{n \rightarrow \infty} \frac{1}{n} H(\Psi_1, \dots, \Psi_n).$$

III. THE ENTROPY RATE OF PATTERNS

We determine the entropy rate of patterns of sequences generated by *i.i.d.* distributions. We show that for discrete distributions it equals the entropy rate of the underlying distribution, and that for distributions with continuous probability q , it equals that of a modified distribution where the continuous probability is assigned to a new discrete element.

1. Discrete distributions

It is easy to show that when the alphabet \mathcal{A} is finite the entropy rates of sequences and their patterns coincide. Since the pattern is determined by the sequence and can derive from at most $|\mathcal{A}|!$ sequences,

$$H(X_1, \dots, X_n) - \log(|\mathcal{A}|!) \leq H(\Psi_1, \dots, \Psi_n) \leq H(X_1, \dots, X_n).$$

Hence

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(\Psi_1, \dots, \Psi_n) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) = H(X).$$

When the alphabet is infinite, more needs to be proven. Cesáro's mean theorem states that if a sequence a_1, a_2, \dots tends to a limit then $b_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n a_i$ tends to the same limit. Entropy rates are therefore often, *e.g.*, [15], evaluated via the conditional entropies:

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) = \lim_{n \rightarrow \infty} H(X_n | X_1, \dots, X_{n-1})$$

assuming that the latter limit exists. We will use Cesáro's theorem after proving the next three lemmas.

Let p be a discrete distribution over an alphabet \mathcal{A} . The entropy of a set $A \subseteq \mathcal{A}$ is

$$H_A = \sum_{a \in A} p(a) \log \frac{1}{p(a)}.$$

Clearly,

$$0 \leq H_A \leq H.$$

For $0 \leq \gamma \leq 1$, let

$$H_\gamma \stackrel{\text{def}}{=} \inf \{H_A : p(A) \geq 1 - \gamma\},$$

be the lowest entropy of sets whose missing mass is at most γ .

Lemma 1. For every discrete distribution p ,

$$\lim_{\gamma \rightarrow 0} H_\gamma = H.$$

Proof If the alphabet \mathcal{A} is finite, the result follows easily. When $\mathcal{A} = \{a_1, a_2, \dots\}$ is countably infinite, assume without loss of generality that $p(a_1) \geq p(a_2) \geq \dots$. Let $N(\gamma)$ be the largest integer i such that $p_i > \gamma$, and let

$$A_\gamma = \{a_1, a_2, \dots, a_{N(\gamma)}\}$$

be the set of all letters with probability $> \gamma$.

If a set $A \subseteq \mathcal{A}$ has probability $\geq 1 - \gamma$, then it must contain all elements whose probability is $> \gamma$, hence $A_\gamma \subseteq A$. It follows that

$$\sum_{i=1}^{N(\gamma)} p(a_i) \log \frac{1}{p(a_i)} \leq H_\gamma \leq H.$$

Clearly, as $\gamma \rightarrow 0$, $N(\gamma) \rightarrow \infty$, hence

$$\lim_{\gamma \rightarrow 0} \sum_{i=1}^{N(\gamma)} p(a_i) \log \frac{1}{p(a_i)} = \lim_{N(\gamma) \rightarrow \infty} \sum_{i=1}^{N(\gamma)} p(a_i) \log \frac{1}{p(a_i)} = H,$$

and the lemma follows. \square

Let ν_n be the event that X_n is "new", namely

$$X_n \neq X_i, \quad i = 1, 2, \dots, n-1.$$

Lemma 2. For all discrete distributions p ,

$$p(\nu_n) \leq \frac{H}{\log n}.$$

Proof The lemma clearly holds when the entropy is infinite. Otherwise,

$$\begin{aligned} p(\nu_n) &= \sum_{x \in \mathcal{A}} (1 - p(x))^n p(x) \\ &\leq \frac{1}{\log n} \sum_{x \in \mathcal{A}} p(x) \log \frac{1}{p(x)} \\ &= \frac{H}{\log n}, \end{aligned}$$

where the inequality follows by expanding $\log x$ using the Taylor series in $1 - x$,

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

For a string $\bar{x} = x_1 \dots x_n$, let $\mathcal{A}(\bar{x}) = \{x_1, \dots, x_n\}$ be the set of elements appearing in the string, and let its *missing mass* be

$$\gamma(\bar{x}) = 1 - p(\mathcal{A}(\bar{x}))$$

be the probability of elements that do not appear in \bar{x} . The following lemma shows that as the blocklength increases, with high probability, the missing mass decreases to 0.

Lemma 3. For any discrete distribution with entropy H , for $n > 1$,

$$p\left\{\gamma(X_1, \dots, X_n) \geq \frac{H}{\log \log n}\right\} \leq \frac{\log \log n}{\log n}.$$

Proof The lemma holds for distributions with infinite entropy. For finite entropy,

$$\mathbf{E}\gamma(X_1, \dots, X_n) = \nu_n \leq \frac{H}{\log n},$$

where \mathbf{E} denotes expectation, and the inequality follows from the last lemma. The current lemma then follows from Markov's inequality. \square

We now prove that the pattern entropy rate is H .

Theorem 4. For all discrete distributions,

$$H_\Psi = H.$$

Proof Since X_1, \dots, X_n determine Ψ_1, \dots, Ψ_n ,

$$\frac{1}{n}H(\Psi_1, \dots, \Psi_n) \leq \frac{1}{n}H(X_1, \dots, X_n) = H, \quad (1)$$

hence

$$H_\Psi \leq H.$$

We now prove that

$$\lim_{n \rightarrow \infty} \frac{1}{n}H(\Psi_1, \dots, \Psi_n) \geq H, \quad (2)$$

and the theorem will follow. Let

$$A_n = \left\{\bar{x} : \gamma(\bar{x}) \leq \frac{H}{\log \log n}\right\}.$$

be the set of length- n strings whose missing mass is at most

$$\frac{H}{\log \log n} \stackrel{\text{def}}{=} \epsilon_n.$$

Then

$$\begin{aligned} H(\Psi_n | \Psi_1, \dots, \Psi_{n-1}) &\geq H(\Psi_n | X_1, \dots, X_{n-1}) \\ &\geq \sum_{\bar{x} \in A_n} p(\bar{x}) H(\Psi_n | \bar{x}) \\ &\geq \sum_{\bar{x} \in A_n} p(\bar{x}) \sum_{x \in \mathcal{A}(\bar{x})} p(x) \log \frac{1}{p(x)} \\ &\stackrel{(a)}{\geq} \sum_{\bar{x} \in A_n} p(\bar{x}) H \epsilon_n \\ &\stackrel{(b)}{\geq} \left(1 - \frac{\log \log n}{\log n}\right) H \epsilon_n, \end{aligned}$$

where (a) follows from Lemma 2, and (b) from Lemma 3. By Lemma 1 and a slight variation of Cesàro's Theorem,

$$\frac{1}{n}H(\Psi_1, \dots, \Psi_n) \geq \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{\log \log i}{\log i}\right) H \epsilon_i \rightarrow H,$$

implying (2). \square

2. Mixed distributions

We show that the entropy of patterns rate equals that of the underlying distribution where the continuous probability is assigned to a new discrete element. More precisely, let p be a mixed distribution with continuous probability q . Define p' to be the discrete distribution obtained from p by omitting the continuous part, and adding a single discrete element of probability q . We show that

$$H_\Psi = H(p') = h(q) + (1-q)\tilde{H} \stackrel{\text{def}}{=} H',$$

where \tilde{H} be the entropy of the distribution obtained by considering only the discrete probabilities in p , each normalized by $1-q$.

Let $\tilde{\mathcal{A}}(\bar{x})$ be the collection of discrete elements in the support of p that appear in the string $\bar{x} = x_1, \dots, x_n$, let

$$\tilde{\gamma}(\bar{x}) = 1 - \frac{p(\tilde{\mathcal{A}}(\bar{x}))}{1-q}$$

be the normalized probability of discrete elements not in \bar{x} , and let

$$\tilde{n}(\bar{x}) = \sum_{i=1}^{|\bar{x}|} 1(x_i \in \tilde{\mathcal{A}}(\bar{x}))$$

be the number of elements in \bar{x} from the discrete support.

Lemma 5. $p\{\bar{x} : \tilde{n}(\bar{x}) \leq \frac{n(1-q)}{2}\} \leq e^{-\frac{n(1-q)}{8}}$.

Proof The lemma follows by observing that $\mathbf{E}\tilde{n}(\bar{x}) = n(1-q)$, and using Chernoff bound, e.g., [16]. \square

Lemma 6.

$$p\left\{\tilde{\gamma}(X_1, \dots, X_n) \geq \frac{\tilde{H}}{\log \log \tilde{n}(\bar{x})}\right\} \leq \frac{\log \log \tilde{n}(\bar{x})}{\log \tilde{n}(\bar{x})}. \quad \square$$

We now derive the entropy rate.

Theorem 7. For all distributions,

$$H_\Psi = H'.$$

Proof As in Lemma 4, we first show that the pattern entropy rate is at most H . Let X'_i be X_i if $X_i \in \tilde{\mathcal{A}}(\bar{x})$ and the new discrete value d otherwise. Since X'_1, \dots, X'_n also determine Ψ_1, \dots, Ψ_n ,

$$\frac{1}{n}H(\Psi_1, \dots, \Psi_n) \leq \frac{1}{n}H(X'_1, \dots, X'_n) = H',$$

hence

$$H_\Psi \leq H'.$$

The lower bound follows in a similar fashion to Theorem 4. Let

$$A_n = \left\{\bar{x} : \tilde{\gamma}(\bar{x}) \leq \frac{H}{\log \log \tilde{n}(\bar{x})} \text{ and } \tilde{n}(\bar{x}) \geq \frac{n(1-q)}{2}\right\}.$$

be the set of length- n strings in which the mass of the missing discrete elements is at most

$$(1-q) \frac{H}{\log(\log n + \log(1-q) - 1)} \stackrel{\text{def}}{=} (1-q)\epsilon_n.$$

Then,

$$\begin{aligned}
& H(\Psi_n | \Psi_1, \dots, \Psi_{n-1}) \\
& \geq H(\Psi_n | X_1, \dots, X_{n-1}) \\
& \geq \sum_{\bar{x} \in A_n} p(\bar{x}) H(\Psi_n | \bar{x}) \\
& \geq \sum_{\bar{x} \in A_n} p(\bar{x}) \left(\sum_{x \in \tilde{\mathcal{A}}(\bar{x})} p(x) \log \frac{1}{p(x)} + p(\tilde{\mathcal{A}}(\bar{x})^c) \log \frac{1}{p(\tilde{\mathcal{A}}(\bar{x})^c)} \right) \\
& \geq \sum_{\bar{x} \in A_n} p(\bar{x}) \left((1-q) \tilde{H}_{\epsilon_n} + (1-q) \log \frac{1}{1-q} \right. \\
& \quad \left. + p(\tilde{\mathcal{A}}(\bar{x})^c) \log \frac{1}{p(\tilde{\mathcal{A}}(\bar{x})^c)} \right) \\
& \rightarrow H',
\end{aligned}$$

where the limit follows because Lemmas 5, 6, and the union bound imply

$$p(A_n) \rightarrow 1,$$

Lemma 1 implies

$$\tilde{H}_{\epsilon_n} \rightarrow \tilde{H},$$

and

$$p(\tilde{\mathcal{A}}(\bar{x})^c) \log \frac{1}{p(\tilde{\mathcal{A}}(\bar{x})^c)} \rightarrow q \log \frac{1}{q}.$$

As in Theorem 4, $\frac{1}{n} H(\Psi_1, \dots, \Psi_n)$ is sandwiched between two sequences whose limit is H' , and the result follows. \square

IV. ASYMPTOTIC EQUIPARTITION PROPERTY

Shannon [17] showed that strings generated by *i.i.d.* distributions over finite alphabets satisfy an asymptotic equipartition property. Chung [18] generalized this result to infinite alphabets. We prove an equivalent property for patterns of such strings, specifically, that for any $\delta > 0$, as $n \rightarrow \infty$,

$$p \left\{ \frac{1}{n} \left| \log \frac{1}{p(\bar{\psi})} - \mathbf{E} \log \frac{1}{p(\bar{\psi})} \right| \geq \delta \right\} \rightarrow 0.$$

The proof uses *profiles* of patterns, which we define next.

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

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

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

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

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

$$\bar{\varphi} \stackrel{\text{def}}{=} (\varphi_{|\bar{\psi}|}, \dots, \varphi_1)$$

the vector of prevalences of μ in $\bar{\psi}$ for $1 \leq \mu \leq |\bar{\psi}|$. 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 $\varphi(\psi) = (0, 0, 1, 0, 2)$.

If p is an *i.i.d.* distribution, then all patterns $\bar{\psi} \in \Psi^n$ with profile $\bar{\varphi}$, have the same probability, namely,

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

where

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

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

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

We use McDiarmid's bound to show that $\log N(\bar{\varphi})$ concentrates around its mean.

Lemma 8. [McDiarmid [19]] Let X_1, \dots, X_n be independent random variables and let the function $f(x_1, \dots, x_n)$ be such that any change in a single x_i changes $f(x_1, \dots, x_n)$ by at most η . Then,

$$p \left\{ \left| f(X_1, \dots, X_n) - \mathbf{E} f(X_1, \dots, X_n) \right| > \eta \sqrt{\frac{n \ln \frac{2}{\delta}}{2}} \right\} < \delta. \quad \square$$

Corollary 9. For all $\alpha > 0$,

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

Proof Let $f(x_1, \dots, x_n) = \log N(\bar{\varphi})$. 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 lemma follows by setting $\delta = \frac{2}{e^{2n^\alpha}}$ in Lemma 8. \square

We now show that with high probability, profile code-lengths deviate from their expectation by at most roughly $n^{\frac{1+\alpha}{2}} \log n$.

Lemma 10. For all $\alpha > 0$,

$$\begin{aligned}
p \left\{ \left| \log \frac{1}{p(\bar{\varphi})} - \mathbf{E} \log \frac{1}{p(\bar{\varphi})} \right| \geq \left(\pi \sqrt{\frac{2}{3}} \log e \right) n^{\frac{1+\alpha}{2}} \log n \right\} \\
\leq \frac{\exp \left(\pi \sqrt{\frac{2n}{3}} \right)}{\exp \left(\pi \sqrt{\frac{2n}{3}} n^{\frac{\alpha}{2}} \log n \right)}.
\end{aligned}$$

Proof Let $p(n)$ be the number of profiles of length n patterns. We bound the expectation of profile code-lengths as follows,

$$\mathbf{E} \log \frac{1}{p(\bar{\varphi})} \leq \log p(n) \leq \left(\pi \sqrt{\frac{2}{3}} \log e \right) \sqrt{n}.$$

The first inequality follows from Jensen's inequality, while the second follows as $p(n)$ can be shown [11] to be the number of integer partitions of n , which has been computed in [20]. Let $\ell = \left(\pi \sqrt{\frac{2}{3}} \log e \right) n^{\frac{1+\alpha}{2}} \log n$. It follows that

$$\left| \log \frac{1}{p(\bar{\varphi})} - \mathbf{E} \log \frac{1}{p(\bar{\varphi})} \right| \geq \ell \Rightarrow \log \frac{1}{p(\bar{\varphi})} \geq \ell,$$

hence

$$\begin{aligned}
p \left\{ \left| \log \frac{1}{p(\bar{\varphi})} - \mathbf{E} \log \frac{1}{p(\bar{\varphi})} \right| \geq \ell \right\} & \leq p \left\{ \log \frac{1}{p(\bar{\varphi})} \geq \ell \right\} \\
& \leq \frac{\exp \left(\pi \sqrt{\frac{2n}{3}} \right)}{\exp \left(\pi \sqrt{\frac{2n}{3}} n^{\frac{\alpha}{2}} \log n \right)}.
\end{aligned}$$

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

Corollary 9 and Lemma 10 imply the asymptotic equipartition property.

Theorem 11. For all $\delta > 0$,

$$p\left\{\frac{1}{n}\left|\log\frac{1}{p(\bar{\psi})}-\mathbf{E}\log\frac{1}{p(\bar{\psi})}\right|\geq\delta\right\}=\exp\left(-\Omega\left(\frac{n\delta^2}{\log^2 n}\right)\right).$$

Proof Observe that

$$\begin{aligned} & p\left\{\frac{1}{n}\left|\log\frac{1}{p(\bar{\psi})}-\mathbf{E}\log\frac{1}{p(\bar{\psi})}\right|\leq\delta\right\} \\ & \geq p\left\{\frac{1}{n}\left|\log N(\bar{\varphi})-\mathbf{E}\log N(\bar{\varphi})\right|+\frac{1}{n}\left|\log\frac{1}{p(\bar{\varphi})}-\mathbf{E}\log\frac{1}{p(\bar{\varphi})}\right|\leq\delta\right\} \\ & \stackrel{(a)}{\geq} p\left\{\left\{\frac{1}{n}\left|\log N(\bar{\varphi})-\mathbf{E}\log N(\bar{\varphi})\right|\leq 3n^{\frac{\alpha-1}{2}}\log n\right\}\right. \\ & \quad \left.\cap\left\{\frac{1}{n}\left|\log\frac{1}{p(\bar{\varphi})}-\mathbf{E}\log\frac{1}{p(\bar{\varphi})}\right|\leq\left(\pi\sqrt{\frac{2}{3}}\log e\right)n^{\frac{\alpha-1}{2}}\log n\right\}\right\} \\ & \stackrel{(b)}{\geq} 1-\frac{2}{e^{2n^\alpha}}-\frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}n^{\frac{\alpha}{2}}\log n\right)}. \end{aligned}$$

The inequality (a) holds for $0 < \alpha \leq \alpha_\delta$, where α_δ is the solution of

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

Note that $0 < \alpha_\delta < 1$ for sufficiently large n . The inequality (b) follows by using Lemmas 9 and 10.

Equation (3) implies that

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

and the theorem follows by observing that the $2e^{-2n^\alpha}$ dominates the convergence rate. \square

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. Multi-alphabet coding with separate alphabet description. In *Proceedings of Compression and Complexity of Sequences*, 1997.
- [9] N. Jevtić, A. Orłitsky, and N.P. Santhanam. Universal compression of unknown alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2002.
- [10] A. Orłitsky and N.P. Santhanam. Performance of universal codes over infinite alphabets. In *Proceedings of the Data Compression Conference*, March 2003.
- [11] A. Orłitsky, 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. Orłitsky, 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] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and sons., 1991.
- [16] D. Angluin and L.G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, April 1979.
- [17] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.
- [18] K.L. Chung. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612–614, 1961.
- [19] C. McDiarmid. *Surveys in Combinatorics 1989*, chapter On the method of bounded differences, pages 148–188. Cambridge University Press, 1989.
- [20] G.H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75–115, 1918.