Universal Compression of Memoryless Sources Over Unknown Alphabets

Alon Orlitsky, Narayana P. Santhanam, *Student Member, IEEE*, and Junan Zhang, *Student Member, IEEE*

***Abstract*—It has long been known that the compression redundancy of independent and identically distributed (i.i.d.) strings increases to infinity as the alphabet size grows. It is also apparent that any string can be described by separately conveying its symbols, and its *pattern*—the order in which the symbols appear. Concentrating on the latter, we show that the patterns of i.i.d. strings over all, including infinite and even unknown, alphabets, can be compressed with diminishing redundancy, both in block and sequentially, and that the compression can be performed in linear time.**

To establish these results, we show that the number of patterns is the Bell number, that the number of patterns with a given number of symbols is the Stirling number of the second kind, and that the redundancy of patterns can be bounded using results of Hardy and Ramanujan on the number of integer partitions. The results also imply an asymptotically optimal solution for the Good-Turing probability-estimation problem.

***Index Terms*—Large and unknown alphabets, patterns, set and integer partitions, universal compression.**

I. Introduction

SHANNON showed that every discrete source can be compressed no further than to its entropy, and that if its distribution is known, compression to essentially the entropy can be achieved. However in most applications, the underlying distribution is not known. For example, in text compression, neither the distribution of words nor their dependencies on previous words are known, and change with author, subject, and time. Similarly, in image compression, the distribution and interrelation of pixels are not known and vary from picture to picture.

A common assumption in these applications is that the source distribution belongs to some natural class $\mathcal{P}$, such as the collection of independent and identically distributed (i.i.d.), Markov, or stationary distributions, but that the precise distribution within $\mathcal{P}$ is not known [1], [2]. The objective then is to compress the data almost as well as when the distribution is known in advance, namely, to find a *universal* compression scheme that performs almost optimally by approaching the entropy no matter which distribution in $\mathcal{P}$ generates the data. The following is a brief introduction to universal compression. For an extensive overview, see [3]–[6].

Manuscript received March 29, 2003; revised March 31, 2004. This work was supported by the National Science Foundation under Grant CCR-0313367. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Yokohama, Japan, June/July 2003.

The authors are with the Department of Electrical and Computer Engineering and the Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: alon@ucsd.edu; nsanthan@ucsd.edu; j6zhang@ucsd.edu).

Communicated by W. Szpankowski, J. C. Kieffer, and E.-h. Yang, Guest Editors.

Digital Object Identifier 10.1109/TIT.2004.830761

Let a source X be distributed over a discrete support set $\mathcal{X}$ according to a probability distribution p. An *encoding* of X is a prefix-free 1–1 mapping $\phi : \mathcal{X} \to \{0,1\}^*$. It can be shown that every encoding of X corresponds to a probability assignment q over $\mathcal{X}$ where the number of bits allocated to $x \in \mathcal{X}$ is approximately $\log(1/q(x))$. Roughly speaking, the optimal encoding that is selected based on the distribution $p \in \mathcal{P}$ and achieves its entropy, allocates $\log(1/p(x))$ bits to every $x \in \mathcal{X}$.

The extra number of bits required to encode x when q is used instead of p is, therefore,

$$\log \frac{1}{q(x)} - \log \frac{1}{p(x)} = \log \frac{p(x)}{q(x)}.$$

The *worst case redundancy* of q with respect to the distribution $p \in \mathcal{P}$ is

$$\hat{R}(p,q) \stackrel{\text{def}}{=} \sup_{x \in \mathcal{X}} \log \frac{p(x)}{q(x)},$$

the largest number of extra bits allocated for any possible x. The *worst case redundancy* of q with respect to the collection $\mathcal{P}$ is

$$\hat{R}(\mathcal{P},q) \stackrel{\text{def}}{=} \sup_{p \in \mathcal{P}} \hat{R}(p,q),$$

the number of extra bits used for the worst distribution in $\mathcal{P}$ and worst $x \in \mathcal{X}$. The *worst case redundancy* of $\mathcal{P}$ is

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \inf_q \hat{R}(\mathcal{P},q) = \inf_q \sup_{p \in \mathcal{P}} \sup_{x \in \mathcal{X}} \log \frac{p(x)}{q(x)}, \qquad (1)$$

the lowest number of extra bits required in the worst case by any possible encoder q.

For any pair of distributions p and q, $\hat{R}(p,q)$ is nonnegative, and therefore $\hat{R}(\mathcal{P})$ is always nonnegative. Note that when the redundancy $\hat{R}(\mathcal{P})$ is small, there is an encoding that assigns to every x a probability not much smaller than that assigned to x by the most favorable distribution in $\mathcal{P}$.

In addition to worst case redundancy, one can define the *average-case redundancy* of $\mathcal{P}$

$$\overline{R}(\mathcal{P}) = \inf_q \sup_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

that reflects the lowest expected number of additional bits required by an encoding that does not know the underlying distribution. The average-case redundancy is clearly always lower than the worst case redundancy. Since our primary interest is

showing that low redundancy can be achieved, we concentrate on worst case redundancy.

Most universal-compression results, even those applying to distributions with memory, build on corresponding results for i.i.d. distributions. Consequently, the redundancy of $\mathcal{I}_m^n$, the collection of i.i.d.distributions over sequences of length n drawn from an alphabet of size m, was studied extensively, and a succession of papers [7]–[14] has shown that for any fixed m as n increases

$$\hat{R}(\mathcal{I}_m^n) = \frac{m-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + o_m(1) \qquad (2)$$

where Γ is the gamma function, and the $o_m(1)$ term diminishes with increasing n at a rate determined by m.

For any fixed alphabet size m, this redundancy grows logarithmically with the block length n, hence, as n increases to infinity, the per-symbol redundancy $\hat{R}(\mathcal{I}_m^n)/n$ diminishes, implying that, asymptotically, i.i.d. distributions can be represented essentially optimally even when the underlying probability distribution is unknown.

In many applications, however, the natural alphabet which captures the structure of the data is very large, possibly even infinite. For example, the natural units that capture the structure of English are words, not letters, and clearly not bits. English consists of hundreds of thousands of words, comparable to the number of words in a typical text. The natural symbols in image coding are pixels, which may assume 2^{24} different values, and the natural symbols in facsimile transmission are images of the individual letters.

When, as in the above applications, the alphabet size m is large, $\hat{R}(\mathcal{I}_m^n)$ is high too, and increases to infinity as m grows. Similar conclusions hold also when m is allowed to grow with the block length n [15].

A systematic study of universal compression over arbitrary alphabets was taken by Kieffer [4] who analyzed a slightly less restrictive form of redundancy, related to weak universal compression. He derived a necessary and sufficient condition for weak universality, and used it to show that even under this weaker requirement, i.i.d. distributions over infinite alphabets entail infinite redundancy. Faced with Kieffer's impossibility results, subsequent universal compression work has typically avoided general distributions over large alphabets.

On the theoretical side, researchers constructed compression algorithms for subclasses of i.i.d. distributions that satisfy Kieffer's condition. Elias [16], Györfi, Pali, and Van der Meulen [17], and Foster, Stine, and Wyner [18] considered monotone, namely, $p(i) \geq p(i+1)$, i.i.d. distributions over the natural numbers, Uyematsu and Kanaya [19] studied bounded-moment distributions, and Kieffer and Yang [20] and He and Yang [21] showed that any collection satisfying Kieffer's condition can be universally compressed by grammar codes.

Since actual distributions may not satisfy Kieffer's condition, practical compression algorithms have typically taken a different approach, circumventing some of the problems associated with large alphabets by converting them into smaller ones. For example, common implementations of both the Lempel–Ziv and the context-tree-weighting algorithms do not operate on words. Instead, they convert words into letters, letters into bits, and then compress the resulting sequence of bits. Such compression algorithms risk losing the sequential relation between words. For example, if the probability of a word depends on the preceding four words, then it is determined by the previous 20 or more letters, namely, upwards of 160 bits, yet most programs truncate their memory at significantly fewer bits.

Motivated by language modeling for speech recognition and the discussion above, we recently took a different approach to compression of large, possibly infinite, alphabets [15], [22]. A similar approach was considered by Åberg, Shtarkov, and Smeets [23] who lower-bounded its performance when the underlying alphabet is finite and the sequence length increases to infinity, see Section V-C.

To motivate this approach, consider perhaps the simplest infinite-redundancy collection. Let $\mathcal{C}^n \stackrel{\text{def}}{=} \{p_k^n : k \in \mathbb{N}\}$, where each p_k^n is the constant distribution that assigns probability 1 to the length-n sequence $k,\ldots,k$, and probability 0 to any other (constant or varying) sequence. If the source that generates the sequence, namely, k, is known, no bits are needed to describe the resulting sequence $k,\ldots,k$. Yet a universal compression scheme that knows only the class $\mathcal{C}^n$ of distributions, needs to describe k which, as k grows, requires an unlimited number of bits. Therefore, both the worst case and average redundancy incurred by any universal compression scheme is infinite.

While disappointing in showing that even simple collections may have infinite redundancy, this example also suggests an approach around some of this redundancy. The unboundedly many bits required to convey the sequence $k,\ldots,k$ clearly do not describe the sequence's evolution, or *pattern*, but only the element k it consists of. It is natural to ask if this observation holds in general, namely, whether the patterns of all sequences can be efficiently conveyed, and the infinite redundancy found by Kieffer stems only from describing the elements that occur.

The description of any string, over any alphabet, can be viewed as consisting of two parts: the symbols appearing in the string and the pattern that they form. For example, the string

"abracadabra"

can be described by conveying the *pattern*

"12314151231"

and the *dictionary*

index	1	2	3	4	5
letter	a	b	r	c	d

.

Together, the pattern and dictionary specify that the string "abracadabra" consists of the first letter to appear (a), followed by the second letter to appear (b), then by the third to appear (r), the first that appeared (a again), the fourth (c), the first (a), etc.

Of the pattern and the dictionary parts of describing a string, the former has a greater bearing on many applications [24]–[30]. For example, in language modeling, the pattern reflects the structure of the language while the dictionary reflects the spelling of words. We therefore concentrate on pattern compression.

II. Results

In Section III, we formally define patterns and their redundancy. In Section IV, we derive some useful properties of patterns, including a correspondence between patterns and set partitions. We use this correspondence to show that the number of patterns is the Bell number and that the number of patterns with a given number of symbols is the Stirling number of the second kind.

We are primarily interested in universal codes for the class $\mathcal{I}^n$ of all i.i.d. distributions, over all possible alphabets, finite or infinite. As mentioned earlier, for standard compression, the per-symbol redundancy increases to infinity as the alphabet size grows. Yet in Section V, we show that $\hat{R}(\mathcal{I}_\Psi^n)$, the block redundancy of compressing patterns of i.i.d. distributions over potentially infinite alphabets is bounded by

$$\left(\frac{3}{2}\log e\right)n^{1/3}(1+o(1)) \le \hat{R}(\mathcal{I}_\Psi^n) \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}.$$

Therefore, the per-symbol redundancy of coding patterns diminishes to zero as the block length increases, irrespective of the alphabet size. The proofs use an analogy between patterns and set partitions which allows us to incorporate celebrated results of Hardy and Ramanujan on the number of partitions of an integer.

In Section VI, we consider sequential pattern compression, which is of interest in most practical applications. We first construct a sequential compression algorithm whose redundancy is at most

$$\frac{2}{\sqrt{2}-1}\left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n} = \frac{4\pi\log e}{(2-\sqrt{2})\sqrt{3}}\sqrt{n}.$$

However, this algorithm has high computational complexity, hence, we also derive a linear-complexity sequential algorithm whose redundancy is at most

$$\mathcal{O}(n^{2/3})$$

where the implied constant is less than 10. The proofs of the sequential-compression results are more involved than those of block compression, and further improvement of the proposed algorithms involved is warranted.

Note that for both block and sequential compression, the redundancy grows sublinearly with the block length, hence the per-symbol redundancy $\hat{R}(\mathcal{I}_\Psi^n)/n$ diminishes to zero. These results can be extended to distributions with memory [31]. They can also be adapted to yield an asymptotically optimal solution for the Good-Turing probability-estimation problem [32].

III. Patterns

We formally describe patterns and their redundancy.

Let $\mathcal{A}$ be any alphabet. For $\overline{x} = x_1^n = x_1,\ldots,x_n \in \mathcal{A}^n$

$$\mathcal{A}(\overline{x}) \stackrel{\text{def}}{=} \{x_1,\ldots,x_n\}$$

denotes the set of symbols appearing in $\overline{x}$. The *index of* $x \in \mathcal{A}(\overline{x})$ is

$$\imath_{\overline{x}}(x) \stackrel{\text{def}}{=} \min\{|\mathcal{A}(x_1^i)| : 1 \le i \le n \text{ and } x_i = x\},$$

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 indexes. 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,

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

Let

$$\Psi(\mathcal{A}^n) = \{\Psi(\overline{x}) : \overline{x} \in \mathcal{A}^n\}$$

denote the set of patterns of all strings in $\mathcal{A}^n$. For example, if $\mathcal{A}$ contains two elements, then $\Psi(\mathcal{A}) = \{1\}$, $\Psi(\mathcal{A}^2) = \{11, 12\}$, $\Psi(\mathcal{A}^3) = \{111,112,121,122\}$, etc. Let

$$\Psi^n = \bigcup_{\mathcal{A}} \Psi(\mathcal{A}^n)$$

denote the set of all length-n patterns, and let

$$\Psi^* = \bigcup_{n=0}^{\infty} \Psi^n$$

be the set of all patterns. For example

$$\begin{aligned}
\Psi^0 &= \{\lambda\}\\
\Psi^1 &= \{1\}\\
\Psi^2 &= \{11,12\}\\
\Psi^3 &= \{111,112,121,122,123\}\\
\Psi^* &= \{\lambda,1,11,12,111,112,\ldots\}
\end{aligned}$$

where λ is the empty string. Fig. 1 depicts a tree representation of all patterns of length at most 4.

It is easy to see that a string $\overline{\psi}$ is a pattern iff it consists of positive integers such that no integer $i > 1$ appears before the first occurrence of $i-1$. For example, 1, 12, and 1213 are patterns, while 2, 21, and 131 are not.

Every probability distribution p over $\mathcal{A}^*$ induces a distribution p_Ψ over patterns on Ψ^*, where

$$p_\Psi(\overline{\psi}) \stackrel{\text{def}}{=} p(\{\overline{x} \in \mathcal{A}^* : \Psi(\overline{x}) = \overline{\psi}\})$$

is the probability that a string generated according to p has pattern $\overline{\psi}$. When p_Ψ is used to evaluate a specific pattern probability $p_\Psi(\overline{\psi})$, the subscript Ψ can be inferred, and is hence omitted. For example, let p be a uniform distribution over $\{a,b\}^2$. Then p induces on Ψ^2 the distribution

$$\begin{aligned}
p(11) &= p(\{aa,bb\}) = \frac{1}{2}\\
p(12) &= p(\{ab,ba\}) = \frac{1}{2}.
\end{aligned}$$

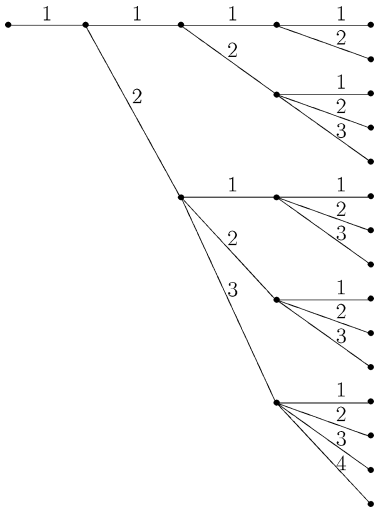

Fig. 1. A tree representation of patterns of $\mathrm{length} \leq 4$.

For a collection $\mathcal{P}$ of distributions over $\mathcal{A}^*$ let

$$\mathcal{P}_\Psi \stackrel{\text{def}}{=} \{p_\Psi : p \in \mathcal{P}\}$$

denote the collection of distributions over Ψ^* induced by probability distributions in $\mathcal{P}$. From the derivations leading to (1), the worst case *pattern redundancy* of $\mathcal{P}$, i.e., the worst case redundancy of patterns generated according to an unknown distribution in $\mathcal{P}_\Psi$ is

$$\hat{R}(\mathcal{P}_\Psi) = \inf_q \sup_{p \in \mathcal{P}_\Psi} \sup_{\overline{\psi} \in \Psi^*} \log \frac{p(\overline{\psi})}{q(\overline{\psi})} \tag{3}$$

where q is any distribution over Ψ^*. In particular, for all $\mathcal{P}$

$$\hat{R}(\mathcal{P}_\Psi) \geq 0.$$

As mentioned earlier, we are mostly interested in $\hat{R}(\mathcal{I}_\Psi^n)$, the pattern redundancy of $\mathcal{I}^n$, the collection of arbitrary i.i.d. distributions over length-n strings. We show that the per-symbol redundancy $\hat{R}(\mathcal{I}_\Psi^n)/n$ diminishes to zero, and that diminishing per-symbol redundancy can be achieved both by block and sequential coding with a constant number of operations per symbol.

IV. Preliminaries

We first establish a correspondence between patterns and set partitions. Set partitions have been studied extensively by a number of well-known researchers [33]–[35], and in this section and in Section V we use their properties to derive the asymptotics of $\left|\Psi^n\right|$ and of the growth rate of $\hat{R}(\mathcal{I}_\Psi^n)$.

A. Set Partitions and Patterns

A *partition* of a set S is a collection of disjoint nonempty subsets of S whose union is S. For $n \geq 0$, let $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$ with $[0] \stackrel{\text{def}}{=} \emptyset$. Let $\mathcal{S}^n$ be the set of all partitions of $[n]$, and let

$$\mathcal{S}^* = \bigcup_{n=0}^{\infty} \mathcal{S}^n$$

be the collection of all partitions of $[n]$ for all $n \in \mathbb{Z}^+$. For example

$$\mathcal{S}^0 = \{\emptyset\}$$
$$\mathcal{S}^1 = \Big\{\{\{1\}\}\Big\}$$
$$\mathcal{S}^2 = \Big\{\{\{1,2\}\}, \{\{1\},\{2\}\}\Big\}.$$

For $0 \leq m \leq n$, let $B(n,m)$ be the number of partitions of $[n]$ into m sets, and let

$$B(n) = \sum_{k=0}^{n} B(n,m)$$

be the number of partitions of $[n]$. For example, $B(0,0) = 1$, $B(1,0) = 0$, $B(1,1) = 1$, $B(2,0) = 0$, $B(2,1) = 1$, $B(2,2) = 1$, and so on, hence,

$$B(0) = 1,\ B(1) = 1,\ B(2) = 2,\ B(3) = 5,\ B(4) = 15, \ldots.$$

Note that $B(n,0)$ is 1 for $n = 0$ and 0 otherwise.

The numbers $B(n,m)$ are called Stirling numbers of the second kind while the numbers $B(n)$, are called *Bell numbers*. Many results are known for both [36]. In particular, it is easy to see that for all $n > 0$, Bell numbers satisfy the recursion

$$B(n+1) = \sum_{i=0}^{n} \binom{n}{i} \cdot B(i).$$

Set partitions are equivalent to patterns. To see that, let the mapping $\mathfrak{f}_\Psi : \Psi^* \to \mathcal{S}^*$ assign to $\overline{\psi} \in \Psi^*$ the set partition

$$\mathfrak{f}_\Psi(\overline{\psi}) \stackrel{\text{def}}{=} \left\{\{i : \psi_i = j\} : 1 \leq j \leq \max_{1 \leq i \leq |\overline{\psi}|} \psi_i\right\}$$

where $|\overline{\psi}|$ denotes the length of $\overline{\psi}$. For example, for the pattern $\overline{\psi} = \psi_1 \ldots \psi_5 = 12131$

$$\begin{aligned}\mathfrak{f}_\Psi(\overline{\psi}) &= \big\{\{i : \psi_i = 1\}, \{i : \psi_i = 2\}, \{i : \psi_i = 3\}\big\} \\ &= \big\{\{1,3,5\}, \{2\}, \{4\}\big\}.\end{aligned}$$

The following follows easily.

Lemma 1: The function $\mathfrak{f}_\Psi : \Psi^* \to \mathcal{S}^*$ is a bijection. Furthermore, for every n

$$\mathfrak{f}_\Psi(\Psi^n) = \mathcal{S}^n. \qquad \square$$

B. Profiles

We classify patterns and set partitions by their *profile*, which will be useful in evaluating the redundancy of i.i.d.-induced distributions.

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

$$\mu_\psi \stackrel{\text{def}}{=} \mu_\psi(\overline{\psi}) \stackrel{\text{def}}{=} \left|\{1 \leq i \leq |\overline{\psi}| : \psi_i = \psi\}\right|,$$

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

$$\varphi_\mu \stackrel{\text{def}}{=} \varphi_\mu(\overline{\psi}) \stackrel{\text{def}}{=} \left|\{\psi : \mu_\psi = \mu\}\right|,$$

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

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

the vector of prevalences of μ in $\overline{\psi}$ for $1 \le \mu \le |\overline{\psi}|$.

Similarly, the *profile* of S is the vector

$$\overline{\varphi}(S) = (\varphi_{|\overline{\psi}|}, \ldots, \varphi_1),$$

where

$$\varphi_\mu = \left|\{s \in S : |s| = \mu\}\right|$$

is the number of sets of cardinality μ in S. The following is easily observed.

Lemma 2: For all $\overline{\psi} \in \Psi^*$,

$$\varphi(\overline{\psi}) = \varphi\big(\mathfrak{f}_\Psi(\overline{\psi})\big). \qquad \square$$

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 $\varphi(\psi) = (0,0,1,0,2)$. On the other hand, we see that

$$\mathfrak{f}_\Psi(\overline{\psi}) = \big\{\{1,3,5\},\{2\},\{4\}\big\}$$

with its profile $\varphi(S) = (0,0,1,0,2)$.

Let

$$\Phi^n = \{\overline{\varphi} : \exists\, \overline{\psi} \in \Psi^n : \varphi(\overline{\psi}) = \overline{\varphi}\}$$

be the set of profiles of all length-n patterns, and let

$$\Phi^* = \bigcup_{n=0}^{\infty} \Phi^n$$

be the set of profiles of all patterns. Clearly, Φ^n and Φ^* are also the set of profiles of all set partitions in $\mathcal{S}^n$ and $\mathcal{S}^*$, respectively, and for all $\overline{\varphi} \in \Phi^n$

$$\sum_{\mu=1}^{n} \mu\varphi_\mu = n.$$

For $\overline{\varphi} \in \Phi^*$, let

$$\Psi_{\overline{\varphi}} \stackrel{\text{def}}{=} \{\overline{\psi} \in \Psi^* : \varphi(\overline{\psi}) = \overline{\varphi}\}$$

be the collection of patterns of profile $\overline{\varphi}$, and, equivalently, let

$$\mathcal{S}_{\overline{\varphi}} \stackrel{\text{def}}{=} \{S \in \mathcal{S}^* : \varphi(S) = \overline{\varphi}\}$$

denote the collection of partitions whose profile is $\overline{\varphi}$. It follows that for all $\overline{\varphi} \in \Phi^*$

$$\mathfrak{f}_\Psi(\Psi_{\overline{\varphi}}) = \mathcal{S}_{\overline{\varphi}}.$$

C. Useful Results

In this subsection, we evaluate the size of $\Psi_{\overline{\varphi}}$ and recall Shtarkov's result for computing the worst case redundancy.

Number of Patterns of a Given Profile

Let

$$N(\overline{\varphi}) \stackrel{\text{def}}{=} \left|\Psi_{\overline{\varphi}}\right| = \left|\mathcal{S}_{\overline{\varphi}}\right|$$

be the number of patterns of profile $\overline{\varphi}$. We get the following.

Lemma 3: For all $n \ge 0$ and $\overline{\varphi} = (\varphi_1, \ldots, \varphi_n) \in \Phi^n$

$$N(\overline{\varphi}) = \frac{n!}{\prod_{\mu=0}^{n} (\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}.$$

Proof: There is only one pattern of length 0, the empty string λ, hence the lemma holds.

There are many ways to derive this result. To see one, let $S \in \mathcal{S}_{\overline{\varphi}}$ be a profile-$\overline{\varphi}$ partition of $[n] = \{1, \ldots, n\}$. For $\mu = 1, \ldots, n$, let S_μ be the collection of elements in sets of size μ. Clearly

$$|S_\mu| = \mu\varphi_\mu$$

hence $[n]$ can be decomposed into the sets $S_1, \ldots, S_n$ in

$$\binom{n}{1\varphi_1, 2\varphi_2, \ldots, n\varphi_n} = \frac{n!}{\prod_{\mu=1}^{n} (\mu\varphi_\mu)!}$$

ways. Each set S_μ can be further decomposed into φ_μ interchangeable sets of size μ in

$$\binom{\mu\varphi_\mu}{\underbrace{\mu, \ldots, \mu}_{\varphi_\mu}} \frac{1}{\varphi_\mu!} = \frac{(\mu\varphi_\mu)!}{(\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}$$

ways. These two decompositions uniquely define the partition, hence, the number of profile-$\overline{\varphi}$ partitions of $[n]$ is

$$\frac{n!}{\prod_{\mu=1}^{n} (\mu\varphi_\mu)!} \cdot \prod_{\mu=1}^{n} \frac{(\mu\varphi_\mu)!}{\mu!^{\varphi_\mu} \cdot \varphi_\mu!} = \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}. \qquad \square$$

Shtarkov's Sum

We will frequently evaluate redundancies using a result by Shtarkov [37] showing that the distribution achieving $\hat{R}(\mathcal{P})$ in (1) is

$$q^*(x) = \frac{\sup_{p \in \mathcal{P}} p(x)}{\sum_{x' \in \mathcal{X}} \sup_{p \in \mathcal{P}} p(x')}.$$

It follows that the redundancy of a collection $\mathcal{P}$ of distributions over $\mathcal{X}$ is determined by *Shtarkov's sum*

$$\hat{R}(\mathcal{P}) = \log \left(\sum_{x \in \mathcal{X}} \sup_{p \in \mathcal{P}} p(x) \right). \tag{4}$$

Approximation of Binomial Coefficients

While finite-alphabet results typically involve binomial coefficients of form $\binom{n}{\alpha n}$ for some constant α, large alphabets often require the calculation of $\binom{n}{o(n)}$. The following lemma provides a convenient approximation.

Lemma 4: When $m \to \infty$ and $m = \mathcal{O}(\sqrt{n})$

$$\binom{n}{m} = \Theta\left(\frac{1}{\sqrt{m}} \left(\frac{e\,n}{m} \right)^m \right)$$

and when, in addition, $m = o(\sqrt{n})$

$$\binom{n}{m} = \frac{1}{\sqrt{2\pi m}}\left(\frac{e\,n}{m}\right)^m (1 + o(1)).$$

Proof: Feller's bounds on Stirling's approximation [38] state that for every $n \geq 1$

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{5}$$

Hence, for all $m \leq n$

$$\frac{e^{-\frac{1}{12}\left(\frac{1}{m}+\frac{1}{(n-m)}\right)}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{n}{m}\right)^m\left(\frac{n}{n-m}\right)^{n-m} \leq \binom{n}{m} \leq \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{n}{m}\right)^m\left(\frac{n}{n-m}\right)^{n-m}.$$

Taking derivatives, it is easy to see that for all $x \geq 0$,

$$e^{x-x^2/2} \leq 1 + x \leq e^x,$$

hence, for all $m \leq n$

$$\left(\frac{n}{n-m}\right)^{n-m} = \left(1 + \frac{m}{n-m}\right)^{n-m} \leq \left(e^{\frac{m}{n-m}}\right)^{n-m} = e^m$$

and

$$\left(\frac{n}{n-m}\right)^{n-m} = \left(1 + \frac{m}{n-m}\right)^{n-m} \geq \exp\left[\left(\frac{m}{n-m} - \frac{1}{2}\left(\frac{m}{n-m}\right)^2\right)(n-m)\right] = \frac{e^m}{e^{\frac{1}{2}\frac{m^2}{n-m}}}.$$

Therefore, for all $m \leq n$

$$\frac{1}{C\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{e\,n}{m}\right)^m \leq \binom{n}{m} \leq \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{e\,n}{m}\right)^m,$$

where

$$C = \exp\left(\frac{1}{12m} + \frac{1}{12(n-m)} + \frac{1}{2}\frac{m^2}{n-m}\right)$$

proving the first part of the lemma. When $m \to \infty$ and $m = o(\sqrt{n})$, $C = 1 + o(1)$, and the second part follows. □

V. Block Compression

We show that for all $n \in \mathbb{Z}^+$

$$\left(\frac{3}{2}\log e\right)\cdot n^{\frac{1}{3}}(1 + o(1)) \leq \hat{R}(\mathcal{I}_\Psi^n) \leq \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}$$

namely, the redundancy of patterns of i.i.d. distributions is sublinear in the block length, implying that the per-symbol redundancy diminishes to zero. To obtain these bounds we rewrite Shtarkov's sum (4) as

$$\hat{R}(\mathcal{I}_\Psi^n) = \log\left(\sum_{\overline{\varphi}\in\Phi^n}\sum_{\overline{\psi}\in\Psi_{\overline{\varphi}}}\sup_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi})\right). \tag{6}$$

In the next subsection, we use this sum to compute $\hat{R}(\mathcal{I}_\Psi^1)$ and $\hat{R}(\mathcal{I}_\Psi^2)$. However, for larger n, exact calculation of the maximum-likelihood probabilities of patterns, namely, $\sup_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi})$, seems difficult [39]. Hence, in Sections V-B and V-C, respectively, we prove upper and lower bounds on the maximum-likelihood probabilities of patterns and use these bounds to upper and lower bound $\hat{R}(\mathcal{I}_\Psi^n)$.

A. The Redundancy of Patterns of Lengths 1 and 2

We determine the redundancies $\hat{R}(\mathcal{I}_\Psi^1)$ and $\hat{R}(\mathcal{I}_\Psi^2)$ of i.i.d.-induced distributions over patterns of lengths 1 and 2, respectively.

There is only one distribution on $\Psi^1 = \{1\}$, hence,

$$\hat{R}(\mathcal{I}_\Psi^1) = 0.$$

For length 2, consider the collection of distributions over

$$\Psi^2 = \{11, 12\}$$

induced by the set $\mathcal{I}^2$ of i.i.d. distributions over strings of length 2. By Shtarkov's sum

$$\hat{R}(\mathcal{I}_\Psi^2) = \log\left(\sum_{\overline{\psi}\in\Psi^2}\sup_{p\in\mathcal{I}_\Psi^2} p(\overline{\psi})\right) = \log\left(\sup_{p\in\mathcal{I}_\Psi^2} p(11) + \sup_{p\in\mathcal{I}_\Psi^2} p(12)\right). \tag{7}$$

Since any constant i.i.d. distribution assigns $p(11) = 1$, the maximum-likelihood probability of 11 is 1, hence,

$$\sup_{p\in\mathcal{I}_\Psi^2} p(11) = 1.$$

Similarly, any continuous distribution over $[0, 1]$ assigns $p(12) = 1$, hence,

$$\sup_{p\in\mathcal{I}_\Psi^2} p(12) = 1.$$

Incorporating into (7), we obtain

$$R(\mathcal{I}_\Psi^2) = \log(1 + 1) = 1.$$

Unfortunately, calculation of maximum-likelihood probabilities for longer patterns seems difficult. Therefore, instead of evaluating the sum in (6) exactly, we bound the maximum-likelihood probabilities of $\overline{\psi} \in \Psi^n$ to obtain bounds on $\hat{R}(\mathcal{I}_\Psi^n)$.

B. Upper Bound

We first describe a general bound on the redundancy of any collection of distributions, and then use it to show that

$$\hat{R}(\mathcal{I}_\Psi^n) \leq \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}.$$

A General Upper Bound Technique

Let $\mathcal{P}$ be a collection of distributions. The following bound on the redundancy of $\mathcal{P}$ is easily obtained.

Lemma 5: For all $\mathcal{P}$

$$\hat{R}(\mathcal{P}) \leq \log |\mathcal{P}|.$$

Proof: The claim is obvious when $\mathcal{P}$ is infinite, and for finite $\mathcal{P}$, Shtarkov's sum implies that

$$\begin{aligned}\hat{R}(\mathcal{P}) &= \log \sum_{x\in\mathcal{X}} \max_{p\in\mathcal{P}} p(x) \\ &\leq \log \sum_{p\in\mathcal{P}} \sum_{\substack{x\in\mathcal{X}: \\ \max_{p'\in\mathcal{P}} p'(x)=p(x)}} p(x) \\ &\leq \log \sum_{p\in\mathcal{P}} 1 \\ &= \log |\mathcal{P}|. \end{aligned}$$ □

Intuitively, for finite $\mathcal{P}$, the lemma corresponds to first identifying the maximum-likelihood distribution of x from all distributions in $\mathcal{P}$ and then describing x using this distribution. If not all distributions are candidates for maximum-likelihood distributions, the above bound can be improved as follows.

A collection $\hat{\mathcal{P}}$ of distributions dominates $\mathcal{P}$ if for all $x \in \mathcal{X}$

$$\sup_{p\in\hat{\mathcal{P}}} p(x) \geq \sup_{p\in\mathcal{P}} p(x),$$

namely, the highest probability of any $x \in \mathcal{X}$ in $\hat{\mathcal{P}}$ is at least as high as that in $\mathcal{P}$. The next lemma then follows immediately from Shtarkov's sum.

Lemma 6: If $\hat{\mathcal{P}}$ dominates $\mathcal{P}$, then

$$\hat{R}(\mathcal{P}) \leq \hat{R}(\hat{\mathcal{P}}).$$ □

The preceding lemmas imply that the redundancy is upper-bounded by the logarithm of the size of $\hat{\mathcal{P}}$.

Corollary 7: If $\hat{\mathcal{P}}$ dominates $\mathcal{P}$, then

$$\hat{R}(\mathcal{P}) \leq \log \left|\hat{\mathcal{P}}\right|.$$ □

To illustrate this bound, we bound the standard redundancy of strings distributed i.i.d. over finite alphabets.

Example 1: Consider the collection $\mathcal{I}_2^n$ of all i.i.d. distributions over alphabets of size 2, which, without loss of generality, we assume to be $\{0,1\}$. Clearly

$$\hat{\mathcal{I}}_2^n = \left\{ p \in \mathcal{I}_2^n : p(0) = \frac{k}{n} \text{ where } 0 \leq k \leq n \text{ and } k \in \mathbb{Z} \right\}$$

dominates $\mathcal{I}_2^n$, hence, from Corollary 7

$$\hat{R}(\mathcal{I}_2^n) \leq \log \left|\hat{\mathcal{I}}_2^n\right| = \log (n+1).$$

Similarly, since

$$\left|\hat{\mathcal{I}}_n^m\right| = \binom{n+m-1}{m-1}$$

it follows that for every m and n

$$\hat{R}(\mathcal{I}_n^m) \leq (m-1)\log\left(e \cdot \frac{n+m-1}{m-1}\right).$$ □

Redundancy

We show that $\hat{R}(\mathcal{I}_\Psi^n)$ is at most the logarithm of the number of profiles. Similar to the correspondence between patterns and set partitions, we obtain a correspondence between profiles of patterns and unordered partitions of positive integers, and use Hardy and Ramanujan's results on the number of unordered partitions of positive integers to bound the number of profiles, and hence the redundancy.

Lemma 8: For all n

$$\hat{R}(\mathcal{I}_\Psi^n) \leq \log |\Phi^n|.$$

Proof: Every induced i.i.d. distribution assigns the same probability to all patterns of the same profile. Hence, the probability assigned to every pattern of a given profile is at most the inverse of the number of patterns of that profile. Let $\hat{\mathcal{I}}_\Psi^n$ consist of $|\Phi^n|$ distributions, one for each profile. The distribution associated with any profile assigns to each pattern of that profile a probability equal to the inverse of the number of patterns of the profile. $\hat{\mathcal{I}}_\Psi^n$ clearly dominates $\mathcal{I}_\Psi^n$ and the lemma follows. □

To count the number of profiles in Φ^n, we observe the following correspondence with unordered partitions of positive integers.

An *unordered partition* of a positive integer n is a multiset of positive integers whose sum is n. An unordered partition can be represented by the vector $\overline{\varphi} = (\varphi_n, \ldots, \varphi_1)$, where φ_μ denotes the number of times μ occurs in the partition. For example, the partition $\{1,1,3\}$ of 5 corresponds to the vector $(0,0,1,0,2)$. Unordered partitions of a positive integer n and profiles of patterns in Ψ^n are equivalent as follows.

Lemma 9: A vector $\overline{\varphi}$ is an unordered partition of n iff $\overline{\varphi} \in \Phi^n$. □

Henceforth, we use the notation developed for profiles of patterns in Section IV-B for unordered partitions as well.

Lemma 10: (Hardy and Ramanujan [40], see also [41]) The number of unordered partitions of n is

$$\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}(1-o(1))\right) \leq |\Phi^n| \leq \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right).$$ □

Lemmas 8 and 10 imply the following upper bound on the pattern redundancy of $\mathcal{I}^n$.

Theorem 11: For all n

$$\hat{R}(\mathcal{I}_\Psi^n) \leq \left(\pi\sqrt{\frac{2}{3}\log e}\right)\sqrt{n}.$$ □

In particular, the pattern redundancy of i.i.d. strings is sublinear in the block length, and hence the per-symbol redundancy diminishes as the number of compressed symbols increases. We note that the number of integer partitions has also been used by

Csiszár and Shields [42] to bound the redundancy of renewal processes.

C. Lower Bound

In the last section, we showed that the redundancy of patterns of i.i.d. strings is $\mathcal{O}(n^{1/2})$. We now show that it is $\Omega(n^{1/3})$. We provide a simple proof of this lower bound and mention a more complex approach that yields the same growth rate, but with a higher multiplicative constant.

Theorem 12: As n increases

$$\hat{R}(\mathcal{I}_\Psi^n) \geq \log\left(\frac{e^{23/12}}{\sqrt{2\pi}}\right) \cdot n^{\frac{1}{3}}(1+o(1)).$$

Proof: Let

$$\Psi_p^{-1}(\overline{\psi}) = \{\overline{x} \in \mathcal{A}^* : \Psi(\overline{x}) = \overline{\psi} \text{ and } p(\overline{x}) > 0\}$$

be the *support* of a pattern $\overline{\psi}$ with respect to a distribution p over an alphabet $\mathcal{A}$. As noted in [23], $\Psi_p^{-1}(\overline{\psi})$ can be partitioned into sets, each with $\prod_\mu \varphi_\mu!$ equi-probable sequences, where $\overline{\varphi} = (\varphi_n, \ldots, \varphi_1)$ is the profile of $\overline{\psi}$. By standard maximum-likelihood arguments, the probability of any sequence with profile $\overline{\varphi}$ is at most $\prod_{\mu=1}^n \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}$, hence,

$$\begin{aligned}\sup_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi}) &= \sup_{p\in\mathcal{I}^n} p\big(\Psi_p^{-1}(\overline{\psi})\big) \\ &\geq \prod_\mu \varphi_\mu! \cdot \max_{p\in\mathcal{I}^n} \max_{\overline{x}\in\Psi_p^{-1}(\overline{\psi})} p(\overline{x}) \\ &\geq \prod_\mu \varphi_\mu! \Big(\frac{\mu}{n}\Big)^{\mu\varphi_\mu}. \end{aligned} \tag{8}$$

From Shtarkov's sum (4)

$$\begin{aligned}\hat{R}(\mathcal{I}_\Psi^n) &= \log\left(\sum_{\overline{\varphi}\in\Phi^n} \sum_{\overline{\psi}\in\Psi_{\overline{\varphi}}} \sup_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi})\right) \\ &\overset{(a)}{\geq} \log\left(\sum_{\overline{\varphi}\in\Phi^n} \frac{n!}{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu}\varphi_\mu!} \cdot \prod_{\mu=1}^n \varphi_\mu! \Big(\frac{\mu}{n}\Big)^{\mu\varphi_\mu}\right) \\ &\overset{(b)}{\geq} \log\left(\frac{e^n n!}{n^n} \sum_{m=1}^n \sum_{\overline{\varphi}\in\Phi_m^n} \frac{1}{\sqrt{\prod_{\mu=1}^n \mu^{\varphi_\mu}}} \Big(\frac{1}{\sqrt{2\pi}}\Big)^m \frac{1}{e^{m/12}}\right) \\ &\overset{(c)}{\geq} \log\left(\sum_{m=1}^n \sum_{\overline{\varphi}\in\Phi_m^n} \Big(\frac{m}{n}\Big)^{m/2} \Big(\frac{1}{\sqrt{2\pi}}\Big)^m \frac{1}{e^{m/12}}\right) \\ &\overset{(d)}{\geq} \log\left(\sum_{m=1}^n \binom{n-1}{m-1} \frac{1}{m!} \cdot \Big(\frac{m}{n}\Big)^{m/2} \Big(\frac{1}{\sqrt{2\pi}}\Big)^m \frac{1}{e^{m/12}}\right) \\ &\geq \log\left(\binom{n-1}{m-1} \cdot \frac{1}{m!} \cdot \Big(\frac{m}{n}\Big)^{m/2} \Big(\frac{1}{\sqrt{2\pi}}\Big)^m \frac{1}{e^{m/12}}\right)\Bigg|_{m=n^{\frac{1}{3}}} \\ &\overset{(e)}{\geq} \log\left(\frac{m}{n} \frac{1+o(1)}{\sqrt{2\pi m}} \Big(\frac{ne}{m}\Big)^m \cdot \frac{1}{\sqrt{2\pi m} e^{\frac{1}{12m}}} \frac{e^m}{m^m} \right. \\ &\qquad \left. \Big(\frac{m}{n}\Big)^{m/2} \Big(\frac{1}{\sqrt{2\pi}}\Big)^m \frac{1}{e^{m/12}}\right)\Bigg|_{m=n^{\frac{1}{3}}} \\ &\geq \log\left(\frac{e^{2n^{1/3}}}{\big(\sqrt{2\pi}\big)^{n^{1/3}} e^{n^{1/3}/12}}\right)(1+o(1)) \\ &= \log\left(\frac{e^{23/12}}{\sqrt{2\pi}}\right) \cdot n^{1/3}(1+o(1)) \end{aligned}$$

where (a) follows from Lemma 3 and (8), (b) from Feller's bounds (5), (c) from the arithmetic-geometric mean inequality, (d) because each unordered partition into m parts can be ordered in at most $m!$ ways, and (e) from Lemma 4. The theorem follows. □

Note that the constant in the bound can be increased by taking

$$m = \left(2\pi e^{-5/6}\right)^{-1/3} \cdot n^{1/3}$$

in the proof, yielding

$$\hat{R}(\mathcal{I}_\Psi^n) \geq \left(2\pi e^{-5/6}\right)^{-1/3} \cdot \frac{3}{2} \log e \cdot n^{1/3}(1+o(1)).$$

Generating functions and Hayman's theorem can be used to evaluate the exact asymptotic growth of

$$\log\left(\sum_{\overline{\varphi}\in\Phi^n} \frac{n!}{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu}\varphi_\mu!} \prod_{\mu=1}^n \Big(\frac{\mu}{n}\Big)^{\mu\varphi_\mu}\right) \tag{9}$$

thereby improving the lower bound to the following.

Theorem 13: [43] (see also [15]) As n increases

$$\hat{R}(\mathcal{I}_\Psi^n) \geq \left(\frac{3}{2}\log e\right) n^{\frac{1}{3}}(1+o(1)).$$ □

These lower bounds should be compared with those in Åberg, Shtarkov, and Smeets [23] who lower-bounded pattern redundancy when the number m of symbols is fixed and finite and the block length n increases to infinity. While it is not clear whether their proof extends to arbitrary m, which may grow with n, the bound they derive may still hold in general. If so, it would yield a lower bound similar to those described here. For a more complete discussion, see [44]. Note also that subsequent to the derivation of Theorem 13, Shamir *et al.* [45], [46] showed that the average-case pattern redundancy is lower-bounded by $(\pi/2)^{1/3}\, 1.5 \log e\, n^{(1-\epsilon)/3}$ for arbitrarily small ϵ.

VI. Sequential Compression

The compression schemes considered so far operated on the whole block of symbols. In many applications the symbols arrive and must be encoded sequentially. Compression schemes for such applications are called *sequential* and associate with every pattern $\psi_1^n \in \Psi^n$, a probability distribution $q(x|\psi_1^n)$ over

$$[\max(\psi_1^n)+1] = \{1, \ldots, \max(\psi_1^n)+1\}$$

representing the probability that the encoder assigns to the possible values of ψ_{n+1} after seeing ψ_1^n. For example

$$q(x|\Lambda) \overset{\text{def}}{=} q(x)$$

is a distribution over $\{1\}$, namely, $q(1)=1$, $q(x|1)$, and $q(x|11)$ are distributions over $\{1,2\}$, while $q(x|12)$ is a distribution over $\{1,2,3\}$.

Let q be a sequential encoder. For each $n \in \mathbb{Z}^+$, q induces a probability distribution over Ψ^n given by

$$q(\psi_1^n) \overset{\text{def}}{=} \prod_{i=1}^n q(\psi_i|\psi_1^{i-1}).$$

Some simple algorithms along the lines of the add-constant rules were analyzed in [15] and shown to have diminishing per-symbol redundancy when the number of distinct symbols is small, but a constant per-symbol redundancy in general. In this section, we describe two sequential encoders with diminishing

per-symbol redundancy. The first encoder, $q_{1/2}$, has worst case redundancy of at most

$$\frac{4\pi \log e}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}$$

only slightly higher than the upper bound on $\hat{R}(\mathcal{I}_\Psi^n)$. However, this encoder has high computational complexity, and in Section VI-B, we consider a sequential encoder $q_{2/3}$ with linear computational complexity and redundancy less than

$$10n^{2/3}$$

which still grows sublinearly with n though not as slowly as the block redundancy.

A. A Low-Redundancy Encoder

We construct an encoder $q_{1/2}$ that for all n, and all patterns ψ_1^n achieves a redundancy

$$\hat{R}(\mathcal{I}_\Psi^n, q_{1/2}) \le \frac{4\pi \log e}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}.$$

The encoder uses distributions that are implicit in the block coding results.

Let

$$\hat{p}_{\psi_1^n}(\psi_1^n) \stackrel{\text{def}}{=} \sup_{p \in \mathcal{I}_\Psi^n} p(\psi_1^n)$$

denote the maximum-likelihood probability assigned to a pattern $\psi_1^n \in \Psi^n$ by any i.i.d. distribution in $\mathcal{I}_\Psi^n$. Recall that $N(\overline{\varphi})$ is the number of patterns with profile $\overline{\varphi}$, and that, as in Lemma 8, every i.i.d. distribution assigns the same probability to all patterns of the same profile. We, therefore, obtain the following upper bound on the maximum pattern probabilities.

Lemma 14: For any pattern $\psi_1^n \in \Psi^n$ of profile $\overline{\varphi} \in \Phi^n$

$$\hat{p}_{\psi_1^n}(\psi_1^n) \le \frac{1}{N(\varphi(\psi_1^n))}.$$ □

Based on this upper bound, we can construct the following distribution over Ψ^n:

$$\tilde{p}(\psi_1^n) \stackrel{\text{def}}{=} \frac{\frac{1}{N(\varphi(\psi_1^n))}}{\sum_{\overline{\psi} \in \Psi^n} \frac{1}{N(\varphi(\overline{\psi}))}} = \frac{1}{N(\varphi(\psi_1^n))\,|\Phi^n|}. \tag{10}$$

For $n \ge 1$, let

$$t_n = 2^{\lceil \log n \rceil}$$

be the smallest power of 2 that is at least n, e.g., $t_1 = 1, t_2 = 2$, and $t_3 = t_4 = 4$. Note that $\frac{t_n}{2} < n \le t_n$.

For every $k \ge n$, and patterns ψ_1^n, let

$$\Psi^k(\psi_1^n) = \{\overline{y} \in \Psi^k : y_1 y_2 \ldots y_n = \psi_1^n\}$$

be the set of all patterns that extend ψ_1^n in Ψ^k, and let

$$\tilde{p}^k(\psi_1^n) \stackrel{\text{def}}{=} \tilde{p}(\Psi^k(\psi_1^n)) = \sum_{\overline{y} \in \Psi^k(\psi_1^n)} \tilde{p}(\overline{y})$$

be the probability of the set $\Psi^k(\psi_1^n)$ under the distribution $\tilde{p}$.

The encoder assigns

$$q_{1/2}(1) = 1,$$

and for all $n > 1$ and $\psi_1^n \in \Psi^n$ it assigns the conditional probability

$$q_{1/2}(\psi_n | \psi_1^{n-1}) = \frac{\tilde{p}^{t_n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^{n-1})}. \tag{11}$$

We now bound the redundancy of $q_{1/2}$.

Theorem 15: For all n

$$\hat{R}(\mathcal{I}_\Psi^n, q_{1/2}) \le \frac{4\pi \log e}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}.$$

Proof: Recall that

$$\hat{R}(\mathcal{I}_\Psi^n, q_{1/2}) = \max_{\psi_1^n} \log \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{q_{1/2}(\psi_1^n)}.$$

The theorem holds trivially for $n = 1$. For $n > 1$, rewrite

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} = \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \cdot \frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)}.$$

For all $\psi_1^n \in \Psi^n$, Lemma 16 shows that

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right)$$

and Lemma 17 that

$$\frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{t_n}}{\sqrt{2}-1}\right)$$

and the theorem follows. □

Lemma 16: For all n and ψ_1^n

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right).$$

Proof: Observe that

$$\begin{aligned}
\hat{p}_{\psi_1^n}(\psi_1^n) &\stackrel{(a)}{=} \sup_{p \in \mathcal{I}_\Psi^n} \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} p(\overline{y}) \\
&\stackrel{(b)}{\le} \sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} \hat{p}_{\overline{y}}(\overline{y}) \\
&\stackrel{(c)}{\le} \left(\sum_{\overline{y} \in \Psi^{t_n}(\psi_1^n)} \tilde{p}(\overline{y})\right) \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right) \\
&= \tilde{p}^{t_n}(\psi_1^n) \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right)
\end{aligned}$$

where (a) follows since for all $k \ge n$ and any i.i.d.-induced distribution

$$\sum_{\overline{y} \in \Psi^k(\psi_1^n)} p(\overline{y}) = p(\psi_1^n),$$

(b) by interchanging sup with the summation, and (c) because (10), together with Lemmas 10 and 14 imply that for all n

$$\tilde{p}(\psi_1^n) = \frac{1}{N(\varphi(\psi_1^n))\,|\Phi^n|} \ge \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right)}.$$

Note that this inequality corresponds to the upper bound on $\hat{R}(\mathcal{I}_\Psi^n)$ in Section V-B. □

Lemma 17: For all $n \geq 2$ and all ψ_1^n

$$\frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{t_n}}{\sqrt{2}-1}\right).$$

Proof: We prove by induction on $i \geq 0$ that for all $2^i < n \leq 2^{i+1}$ and all ψ_1^n

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^{i+1}}}{\sqrt{2}-1}\right). \tag{12}$$

The lemma will follow since for every n, t_n is a power of two.

The basis holds since for $i = 0$, $n = 2$, and all ψ_1^2 satisfy

$$q_{1/2}(\psi_1^2) = \tilde{p}(\psi_1^2) = \frac{1}{2}.$$

To prove the step, note from (11) that for $i \geq 1$, all $2^i < n \leq 2^{i+1}$ and ψ_1^n satisfy

$$q_{1/2}(\psi_1^n) = q_{1/2}(\psi_1^{2^i})\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}$$

hence,

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{1/2}(\psi_1^n)} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} \cdot \frac{\tilde{p}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})}. \tag{13}$$

By the induction hypothesis

$$\frac{\tilde{p}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} = \frac{\tilde{p}^{2^i}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^i}}{\sqrt{2}-1}\right). \tag{14}$$

By definition (10) and Lemma 10

$$\tilde{p}(\psi_1^{2^i}) = \frac{1}{N(\overline{\varphi}(\psi_1^{2^i}))\left|\Phi^{2^i}\right|} \geq \frac{1}{N(\overline{\varphi}(\psi_1^{2^i}))\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}\right)}.$$

On the other hand, distinct patterns $\psi_1^{2^i}$ have disjoint sets $\Psi^{2^{i+1}}(\psi_1^{2^i})$, while patterns of the same profile have the same probability $\tilde{p}^{2^{i+1}}(\psi_1^{2^i})$, hence,

$$N\left(\varphi(\psi_1^{2^i})\right) \cdot \tilde{p}^{2^{i+1}}(\psi_1^{2^i}) \leq \sum_{\overline{y} \in \Psi^{2^{i+1}}(\psi_1^{2^i})} \tilde{p}(\overline{y}) = 1$$

and thus,

$$\tilde{p}^{2^{i+1}}(\psi_1^{2^i}) \leq \frac{1}{N(\overline{\varphi}(\psi_1^{2^i}))}.$$

It follows that

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}\right).$$

Incorporating this inequality and (14) in (13), we get (12). □

B. A Low-Complexity Encoder

The evaluation of $q_{1/2}(\psi_1^n)$ in (11) may take super-polynomial time. We, therefore, present a linear-complexity encoder $q_{2/3}$ whose sequential redundancy is less than $10n^{2/3}$ hence its per-symbol redundancy also diminishes to zero as the block length increases, albeit at the slower rate of $\mathcal{O}(n^{-1/3})$.

For notational convenience, let φ_μ^n denote the prevalence of μ in ψ_1^n, $\varphi_\mu(\psi_1^n)$, and let

$$r \stackrel{\text{def}}{=} \mu_{\psi_n}(\psi_1^{n-1}).$$

For $c \in \mathbb{Z}^+$, define

$$f_c(\varphi) \stackrel{\text{def}}{=} \max(\varphi, c) = \begin{cases} c, & 0 \leq \varphi \leq c-1 \\ \varphi, & \varphi \geq c \end{cases}$$

and let

$$g_c(\varphi) \stackrel{\text{def}}{=} \prod_{i=1}^{\varphi} f_c(i) = \begin{cases} c^\varphi, & 0 \leq \varphi \leq c-1 \\ \frac{c^c}{c!}\varphi!, & \varphi \geq c. \end{cases}$$

Finally, define the sequence

$$c_n = \lceil n^{1/3} \rceil.$$

The encoder assigns

$$q_{2/3}(1) = 1$$

and for all $n \geq 2$ and $\psi_1^n \in \Psi^n$, it assigns the conditional probability

$$q_{2/3}(\psi_n|\psi_1^{n-1}) = \frac{1}{S_{c_n}(\psi_1^{n-1})} \cdot \begin{cases} f_{c_n}(\varphi_1^{n-1}+1), & r = 0 \\ (r+1)\frac{f_{c_n}(\varphi_{r+1}^{n-1}+1)}{f_{c_n}(\varphi_r^{n-1})}, & r > 0 \end{cases} \tag{15}$$

where

$$S_{c_n}(\psi_1^{n-1}) \stackrel{\text{def}}{=} f_{c_n}\left(\varphi_1^{n-1}+1\right) + \sum_{\mu=1}^{n-1} \varphi_\mu^{n-1}(\mu+1)\frac{f_{c_n}\left(\varphi_{\mu+1}^{n-1}+1\right)}{f_{c_n}\left(\varphi_\mu^{n-1}\right)}$$

is a normalization factor. It can be shown by induction that for all $n \geq 2$ and all patterns $\psi_1^n \in \Psi^n$

$$q_{2/3}(\psi_1^n) = \frac{\prod_{\mu=1}^n \left((\mu!)^{\varphi_\mu^n} g_{c_n}(\varphi_\mu^n)\right)}{\prod_{i=2}^n S_{c_i}(\psi_1^{i-1})} \cdot \prod_{i=1}^{n-1}\left(\prod_{\mu=1}^{i}\frac{g_{c_i}(\varphi_\mu^i)}{g_{c_{i+1}}(\varphi_\mu^i)}\right).$$

Theorem 18: For all n

$$\hat{R}(\mathcal{I}_\Psi^n, q_{2/3}) \leq C \cdot n^{2/3}$$

where $C < 10$.

Proof: The theorem holds trivially for $n = 1$. For $n \geq 2$, it can be shown that for all $\psi_1^n \in \Psi^n$

$$\begin{aligned} \log\left(\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{q_{2/3}(\psi_1^n)}\right) &\leq \log\left(\frac{1/N(\varphi(\psi_1^n))}{q_{2/3}(\psi_1^n)}\right) \\ &= \log\left(\prod_{\mu=1}^n \frac{\varphi_\mu^n!}{g_{c_n}(\varphi_\mu^n)}\right) \\ &\quad + \log\left(\prod_{i=1}^{n-1}\left(\prod_{\mu=1}^{i}\frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)}\right)\right) \\ &\quad + \log\left(\frac{\prod_{i=2}^n S_{c_i}(\psi_1^{i-1})}{n!}\right) \\ &\stackrel{\text{def}}{=} T_1(\psi_1^n) + T_2(\psi_1^n) + T_3(\psi_1^n). \end{aligned}$$

Since for all $c \in \mathbb{Z}^+$ and $\varphi \in \mathbb{N}$, $g_c(\varphi) \geq \varphi!$

$$T_1(\psi_1^n) \leq 0.$$

In Lemma 20, we show that for all $n \geq 2$ and $\psi_1^n \in \Psi^n$

$$T_2(\psi_1^n) \leq \sum_{i=1}^{n-1}\sqrt{2ic_{i+1}}\log\left(\frac{c_{i+1}}{c_i}\right).$$

Since only $n^{1/3}$ terms are not nonzero, and using the inequality $\ln(1+x) \le x$ for $x > -1$, we obtain

$$T_2(\psi_1^n) \le \sum_{j=1}^{n^{1/3}} \sqrt{2}\log e \cdot j = \mathcal{O}(n^{2/3}).$$

In Lemma 21, we show that for all $n \ge 2$ and $\psi_1^n \in \Psi^n$

$$T_3(\psi_1^n) \le \sum_{i=1}^{n-1} \log\left(1 + \frac{1}{c_{i+1}} + \sqrt{\frac{2(2c_{i+1}+1)^2}{ic_{i+1}}}\right) + \log\frac{1}{n}.$$

Substituting $c_i = \lceil i^{1/3} \rceil$, we obtain

$$\begin{aligned} T_3(\psi_1^n) &< \sum_{i=1}^{n-1} \log\left(1 + \frac{1}{i^{1/3}} + \sqrt{\frac{8((i+1)^{1/3}+2)^2}{i^{4/3}}}\right) + \log\frac{1}{n} \\ &\le \sum_{i=1}^{n-1} \log\left(1 + \frac{c_1}{i^{1/3}} + \frac{c_2}{i^{2/3}}\right) + \log\frac{1}{n} \\ &\le \mathcal{O}\left(\sum_{i=1}^{n-1} \frac{1}{i^{1/3}}\right) + \log\frac{1}{n} \\ &= \mathcal{O}(n^{2/3}). \end{aligned}$$

The theorem follows by a simple evaluation of the constants involved. □

Theorem 19: The number of operations required to compute all of

$$q_{2/3}(\psi_1),\ q_{2/3}(\psi_2|\psi_1), \ldots, q_{2/3}(\psi_n|\psi_1^{n-1})$$

grows linearly with n.

Proof: The conditional probabilities are calculated in (15). Note that computing $c_1, \ldots, c_n$ requires only $\mathcal{O}(n^{1/3})$ multiplications and $\mathcal{O}(n)$ comparisons. It therefore suffices to evaluate the complexity of calculating $S_{c_1}(\psi_1^1), \ldots, S_{c_{n-1}}(\psi_1^{n-1})$.

Let $\mathcal{Z}^3 \stackrel{\text{def}}{=} \{1^3, 2^3, 3^3, \ldots\}$ be the set of perfect cubes. For all $1 < i \notin \mathcal{Z}^3$, $S_{c_i}(\psi_1^{i-1})$ can be updated from $S_{c_{i-1}}(\psi_1^{i-2})$ in constant time. For all $i \in \mathcal{Z}^3$, $S_{c_i}(\psi_1^{i-1})$ can be calculated in $\mathcal{O}(\sqrt{i})$ time because there are at most $\sqrt{2(i-1)}$ multiplicities μ such that $\varphi_\mu > 0$.

Hence, the evaluation of $S_{c_2}(\psi_1^1), \ldots, S_{c_n}(\psi_1^{n-1})$ can be completed in time

$$\sum_{\substack{1 \le i \le n-1 \\ i \in \mathcal{Z}^3}} \mathcal{O}(\sqrt{i}) + \sum_{\substack{1 \le i \le n-1 \\ i \notin \mathcal{Z}^3}} \mathcal{O}(1) = \mathcal{O}(n^{5/6}) + \mathcal{O}(n) = \mathcal{O}(n).$$

□

Finally, the following two technical lemmas complete the proof of Theorem 18.

Lemma 20: For all $n \ge 2$ and $\psi_1^n \in \Psi^n$

$$\log\left(\prod_{i=1}^{n-1}\left(\prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)}\right)\right) \le \sum_{i=1}^{n-1} \sqrt{2ic_{i+1}} \log\left(\frac{c_{i+1}}{c_i}\right).$$

Proof: Let

$$L_i(\varphi) = \begin{cases} \left(\frac{c_{i+1}}{c_i}\right)^{\varphi}, & 0 \le \varphi < c_{i+1} \\ \left(\frac{c_{i+1}}{c_i}\right)^{c_{i+1}}, & \varphi \ge c_{i+1}\ . \end{cases}$$

Observe that

$$\begin{aligned} \frac{g_{c_{i+1}}(\varphi)}{g_{c_i}(\varphi)} &= \begin{cases} \left(\frac{c_{i+1}}{c_i}\right)^{\varphi}, & \varphi \le c_i \\ \frac{{c_{i+1}}^{\varphi} c_i!}{{c_i}^{c_i} \varphi!}, & c_i < \varphi < c_{i+1} \\ \frac{{c_{i+1}}^{c_{i+1}}}{{c_i}^{c_i}} \frac{c_i!}{c_{i+1}!}, & \varphi \ge c_{i+1} \end{cases} \\ &\le L_i(\varphi) \end{aligned}$$

where we note that for $c_i = i^{1/3}$, the second case will never occur. Hence, for all i and $\psi_1^i \in \Psi^i$

$$\prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)} \le \prod_{\mu=1}^{i} L_i(\varphi_\mu^i). \tag{16}$$

Let $\tilde{\varphi}(\psi_1^i)$ maximize $\prod_{\mu=1}^{i} L_i(\varphi_\mu^i)$. It can be shown that

$$\tilde{\varphi}_1^i = \tilde{\varphi}_2^i = \cdots = \tilde{\varphi}_{\omega-1}^i = c_{i+1},\ \tilde{\varphi}_\omega^i \le c_{i+1}$$

and $\tilde{\varphi}_\mu^i = 0$ for $\mu \ge \omega + 1$ where

$$\omega = \sqrt{\frac{2i}{c_{i+1}} + \left(\frac{\tilde{\varphi}_\omega^i}{c_{i+1}} - \frac{1}{2}\right)^2} - \left(\frac{\tilde{\varphi}_\omega^i}{c_{i+1}} - \frac{1}{2}\right)$$

is the solution of

$$c_{i+1} \cdot \frac{\omega(\omega-1)}{2} + \tilde{\varphi}_\omega^i \cdot \omega = i.$$

Hence the right-hand side of (16) simplifies to

$$\prod_{\mu=1}^{i} L_i(\tilde{\varphi}_\mu^i) \le \left(\frac{c_{i+1}}{c_i}\right)^{(\omega-1)c_{i+1}+\tilde{\varphi}_\omega^i} \le \left(\frac{c_{i+1}}{c_i}\right)^{\sqrt{2ic_{i+1}}}$$

implying

$$\log\left(\prod_{i=1}^{n-1}\left(\prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)}\right)\right) \le \sum_{i=1}^{n-1} \sqrt{2ic_{i+1}} \log\left(\frac{c_{i+1}}{c_i}\right).$$

□

Lemma 21: For all $c \in \mathbb{Z}^+$ and all patterns $\psi_1^n \in \Psi^n$

$$S_c(\psi_1^n) \le (1 + \frac{1}{c})n + \sqrt{\frac{2n(2c+1)^2}{c}}.$$

Proof: As before, write $\varphi_\mu(\psi_1^n)$ as φ_μ, and let

$$T_\mu \stackrel{\text{def}}{=} \varphi_\mu(\mu+1)\frac{f_c(\varphi_{\mu+1}+1)}{f_c(\varphi_\mu)}.$$

Recall that

$$\begin{aligned} S_c(\psi_1^n) &= f_c(\varphi_1+1) + \sum_{\mu=1}^{n} \varphi_\mu \cdot (\mu+1)\frac{f_c(\varphi_{\mu+1}+1)}{f_c(\varphi_\mu)} \\ &= f_c(\varphi_1+1) + \sum_{\mu=1}^{n} T_\mu. \end{aligned}$$

For convenience, let $\varphi_0 = c$, so that

$$S_c(\psi_1^n) = c \cdot 1 \cdot \frac{f_c(\varphi_1+1)}{f_c(c)} + \sum_{\mu=1}^{n} T_\mu = \sum_{\mu=0}^{n} T_\mu,$$

T_μ takes different forms depending on whether $(\varphi_\mu, \varphi_{\mu+1})$ falls into $[0, c-1] \times [0, c-1]$, $[0, c-1] \times [c, \infty)$, $[c, \infty) \times [0, c-1]$, or $[c, \infty) \times [c, \infty)$. We, therefore, partition $\varphi_0(=c), \varphi_1, \ldots, \varphi_n, \varphi_{n+1}(=0)$ into alternating nonempty segments

with prevalences $\varphi_\mu \geq c$ and $\varphi_\mu < c$ which we call *high* and *low segments*, respectively. More precisely, a high segment is an interval $[b, e]$ $(b, e \in \mathbb{N})$ such that

$$\varphi_\mu \begin{cases} < c, & \mu = b-1 \geq 0 \\ \geq c, & b \leq \mu \leq e \\ < c, & \mu = e+1 \end{cases}$$

and a low segment is an interval $[b, e]$ $(b, e \in \mathbb{Z}^+)$ such that

$$\varphi_\mu \begin{cases} \geq c, & \mu = b-1 \\ < c, & b \leq \mu \leq e \\ \geq c, & \mu = e+1. \end{cases}$$

Observe that the first segment is always a high segment and the last is always a low segment, hence, the number of high segments equals the number of low segments. Let J be the number of high (or low) segments, and $[b_j, e_j]$ be the jth high segment. Hence, for $j = 1, \ldots, J-1$, $[e_j + 1, b_{j+1} - 1]$ is the jth low segment and $[e_J + 1, n+1]$ is the Jth low segment.

For all μ in high segments, $f_c(\varphi_\mu) = \varphi_\mu$ and $f_c(\varphi_\mu + 1) = \varphi_\mu + 1$, while for all μ in low segments, $f_c(\varphi_\mu) = f_c(\varphi_\mu + 1) = c$. For $a > b$, let $[a, b] \stackrel{\text{def}}{=} \emptyset$ and $\sum_{i=a}^{b} c_i = 0$. Define

$$\begin{aligned} \mathcal{A} &\stackrel{\text{def}}{=} \cup_{j=1}^{J} [b_j, e_j - 1] \\ \mathcal{B} &\stackrel{\text{def}}{=} \{e_1, e_2, \ldots, e_J\} \\ \mathcal{C} &\stackrel{\text{def}}{=} \cup_{j=2}^{J} [e_{j-1} + 1, b_j - 2] \cup [e_J + 1, n] \\ \mathcal{D} &\stackrel{\text{def}}{=} \{b_2 - 1, b_3 - 1, \ldots, b_J - 1\}. \end{aligned}$$

Consequently

$$\begin{aligned} T_\mu &= \varphi_\mu(\mu+1)\frac{f_c(\varphi_{\mu+1}+1)}{f_c(\varphi_\mu)} \\ &= \begin{cases} (\mu+1)(\varphi_{\mu+1}+1), & \mu \in \mathcal{A} \\ (\mu+1)c, & \mu \in \mathcal{B} \\ \varphi_\mu(\mu+1), & \mu \in \mathcal{C} \\ \varphi_\mu(\mu+1)(\varphi_{\mu+1}+1)/c, & \mu \in \mathcal{D}. \end{cases} \end{aligned}$$

We upper-bound T_μ by

$$T_\mu \leq \begin{cases} (\mu+1)\varphi_{\mu+1} + \frac{\mu\varphi_\mu}{c} + 1, & \mu \in \mathcal{A} \\ cb_j + c(e_j + 1 - b_j), & \mu = e_j \in \mathcal{B} \\ \mu\varphi_\mu + \varphi_\mu, & \mu \in \mathcal{C} \\ \frac{1}{c}\varphi_{b_j-1}\varphi_{b_j} b_j + \frac{1}{c}(b_j - 1)\varphi_{b_j-1} + \frac{c-1}{c}, & \mu = b_j \ \ 1 \in \mathcal{D}. \end{cases} \tag{17}$$

Recall that

$$\begin{aligned} S_c(\psi_1^n) = \sum_{\mu=0}^{n} T_\mu &= \sum_{j=1}^{J} \sum_{\mu=b_j}^{e_j-1} T_\mu + \sum_{j=1}^{J} T_{e_j} \\ &\quad + \sum_{j=2}^{J} \sum_{\mu=e_{j-1}+1}^{b_j-2} T_\mu + \sum_{\mu=e_J+1}^{n} T_\mu + \sum_{j=2}^{J} T_{b_j-1}. \end{aligned}$$

From (17)

$$\begin{aligned} S_c(\psi_1^n) &\leq \sum_{j=1}^{J} \sum_{\mu=b_j}^{e_j-1} \left((\mu+1)\varphi_{\mu+1} + \frac{\mu\varphi_\mu}{c} + 1 \right) \\ &\quad + \sum_{j=1}^{J} (cb_j + c(e_j + 1 - b_j)) \\ &\quad + \sum_{j=2}^{J} \sum_{\mu=e_{j-1}+1}^{b_j-2} (\mu\varphi_\mu + \varphi_\mu) + \sum_{\mu=e_J+1}^{n} (\mu\varphi_\mu + \varphi_\mu) \\ &\quad + \sum_{j=2}^{J} \left(\frac{1}{c}\varphi_{b_j-1}\varphi_{b_j} b_j + \frac{1}{c}(b_j - 1)\varphi_{b_j-1} + \frac{c-1}{c} \right) \\ &\leq \sum_{\mu=0}^{n} \mu\varphi_\mu - \sum_{j=1}^{J} b_j\varphi_{b_j} - \sum_{j=2}^{J} (b_j - 1)\varphi_{b_j-1} \\ &\quad + \frac{1}{c}\left(\sum_{j=1}^{J} \sum_{\mu=b_j}^{e_j-1} \mu\varphi_\mu + \sum_{j=2}^{J} (b_j - 1)\varphi_{b_j-1} \right) \\ &\quad + \sum_{j=1}^{J} cb_j + \sum_{j=2}^{J} \frac{1}{c}\varphi_{b_j-1}\varphi_{b_j} b_j \\ &\quad + (c+1)\sum_{j=1}^{J} (e_j + 1 - b_j) + \sum_{\varphi_\mu < c} \varphi_\mu + J - 1 \\ &\leq n + \frac{1}{c}n \\ &\quad + \sum_{j=2}^{J} \left(cb_j + \frac{1}{c}\varphi_{b_j-1}\varphi_{b_j} b_j - b_j\varphi_{b_j-1} - b_j\varphi_{b_j} \right) \\ &\quad + (c+1)\sum_{j=1}^{J} (e_j + 1 - b_j) + \sum_{\varphi_\mu < c} \varphi_\mu + c(J-1) \\ &\leq (1 + \frac{1}{c})n + \sum_{j=2}^{J} b_j(\varphi_{b_j-1} - c)(\frac{\varphi_{b_j}}{c} - 1) \\ &\quad + (c+1)\sum_{j=1}^{J} (e_j + 1 - b_j) + \sum_{\varphi_\mu < c} \varphi_\mu + c(J-1). \end{aligned} \tag{18}$$

To simplify the above expression, observe that for $2 \leq j \leq J$, $\varphi_{b_j-1} < c$ and $\varphi_{b_j} \geq c$, hence,

$$\sum_{j=2}^{J} b_j(\varphi_{b_j-1} - c)(\frac{\varphi_{b_j}}{c} - 1) \leq 0 \tag{19}$$

and that

$$\begin{aligned} (c+1)\sum_{j=1}^{J} (e_j + 1 - b_j) + \sum_{\varphi_\mu < c} \varphi_\mu + c(J-1) &\leq \sum_{\varphi_\mu \geq c} (2c+1) + \sum_{\varphi_\mu < c} \varphi_\mu \\ &\leq \sqrt{\frac{2n(2c+1)^2}{c}} \end{aligned} \tag{20}$$

where the last inequality follows because the profile maximizing

$$\sum_{\varphi_\mu \geq c} (2c+1) + \sum_{\varphi_\mu < c} \varphi_\mu$$

has $\varphi_1 = \varphi_2 = \cdots = \varphi_{\omega-1} = c$, $\varphi_\omega \leq c$, and $\varphi_\mu = 0$ for $\mu > \omega$ where

$$\omega = \sqrt{\frac{2n}{c} + \left(\frac{\varphi_\omega}{c} - \frac{1}{2} \right)^2} - \left(\frac{\varphi_\omega}{c} - \frac{1}{2} \right).$$

Incorporating (19) and (20) into (18), we obtain

$$S_c(\psi_1^n) \le (1+\frac{1}{c})n + \sqrt{\frac{2n(2c+1)^2}{c}}. \qquad \square$$

ACKNOWLEDGMENT

We thank N. Alon, N. Jevtić, G. Shamir, W. Szpankowski, and K. Viswanathan for helpful discussions.

REFERENCES

[1] B. Fittingoff, "Universal methods of coding for the case of unknown statistics," in *Proc. 5th Symp. Information Theory*, Moscow-Gorky, U.S.S.R., 1972, pp. 129–135.

[2] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[3] J. Shtarkov, "Coding of discrete sources with unknown statistics," in *Topics in Information Theory (Coll. Math. Soc. J. Bolyai, no. 16)*, I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland, 1977, pp. 559–574.

[4] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.

[5] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[6] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, Oct. 1998.

[7] R. Krichevsky and V. Trofimov, "The preformance of universal coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.

[8] T. Cover, "Universal portfolios," *Math. Finance*, vol. 1, no. 1, pp. 1–29, Jan. 1991.

[9] Y. Shtarkov, T. Tjalkens, and F. Willems, "Multialphabet universal coding of memoryless sources," *Probl. Inform. Transm.*, vol. 31, no. 2, pp. 114–127, 1995.

[10] T. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, pp. 348–363, Mar. 1996.

[11] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.

[12] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Probl. Inform. Transm.*, vol. 34, no. 2, pp. 142–146, 1998.

[13] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE Trans. Inform. Theory*, vol. 46, pp. 431–445, Mar. 2000.

[14] M. Drmota and W. Szpankowski, "The precise minimax redundancy," in *Proc. IEEE Symp. Inform. Theory*, Lausanne, Switzerland, June/July 2002, p. 35.

[15] A. Orlitsky and N. Santhanam, "Performance of universal codes over infinite alphabets," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2003, pp. 402–410.

[16] P. Elias, "Universal codeword sets and representations of integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, Mar. 1975.

[17] L. Györfi, I. Pali, and E. V. der Meulen, "On universal noiseless source coding for infinite source alphabets," *Europ. Trans. Telecommun. and Related Technologies*, vol. 4, pp. 125–132, 1993.

[18] D. Foster, R. Stine, and A. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1713–1720, June 2002.

[19] T. Uyematsu and F. Kanaya, "Asymptotic optimality of two variations of Lempel-Ziv codes for sources with countably infinite alphabet," in *Proc. IEEE Symp. Information Theory*, Lausanne, Switzerland, June/July 2002, p. 122.

[20] J. Kieffer and E. Yang, "Grammar based codes: A new class of universal lossless source codes," *IEEE Trans. Inform. Theory*, vol. 46, pp. 737–754, May 2000.

[21] D. He and E. Yang, "On the universality of grammar-based codes for sources with countably infinite alphabets," in *Proc. IEEE Symp. Information Theory*, Yokohama, Japan, June/July 2003.

[22] N. Jevtić, A. Orlitsky, and N. Santhanam, "Universal compression of unknown alphabets," in *Proc. IEEE Symp. Information Theory*, Lausanne, Switzerland, 2002, p. 320.

[23] J. Åberg, Y. Shtarkov, and B. Smeets, "Multialphabet coding with separate alphabet description," in *Proc. Compression and Complexity of Sequences*, Salerno, Italy, 1997, pp. 56–65.

[24] K. Church and W. Gale, "Probability scoring for spelling correction," *Statist. and Comput.*, vol. 1, pp. 93–103, 1991.

[25] W. Gale, K. Church, and D. Yarowsky, "A method for disambiguating word senses," *Comput. and Humanities*, vol. 26, pp. 415–419, 1993.

[26] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. 34th Annu. Meeting Association for Computational Linguistics*, San Francisco, CA, 1996, pp. 310–318.

[27] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its application to learning," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1424–1439, July 1998.

[28] V. Vovk, "A game of prediction with expert advice," *J. Comput. Syst. Sci.*, vol. 56, no. 2, pp. 153–173, 1998.

[29] N. Cesa-Bianchi and G. Lugosi, "Minimax regret under log loss for general classes of experts," in *Proc. 12th Annu. Conf. Computational Learning Theory*, Santa Cruz, CA, 1999, pp. 12–18.

[30] F. Song and W. Croft, "A general language model for information retrieval (poster abstract)," in *Proc. Conf. Research and Development in Information Retrieval*, Berkeley, CA, 1999, pp. 279–280.

[31] A. Dhulipala and A. Orlitsky, "On the redundancy of hmm patterns," in *Proc. 2004 Proc. IEEE Int. Symp. Information Theory*, to be published.

[32] A. Orlitsky, N. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, Oct. 2003.

[33] E. Bell, "Exponential numbers," *Amer. Math. Monthly*, vol. 41, pp. 411–419, 1934.

[34] G. Rota, "The number of partitions of a set," *Amer. Math. Monthly*, vol. 71, pp. 498–504, 1964.

[35] N. Sloane. *Online Encyclopedia of Integer Sequences* [Online]. Available: http://www.research.att.com/~njas/sequences/

[36] J. van Lint and R. Wilson, *A Course in Combinatorics*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[37] Y. Shtarkov, "Universal sequential coding of single messages," *Probl. Inform. Transm.*, vol. 23, no. 3, pp. 3–17, 1987.

[38] W. Feller, *An Introduction to Probability Theory*. New York: Wiley, 1968.

[39] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," manuscript, submitted for publication.

[40] G. Hardy and S. Ramanujan, "Asymptotic formulae in combinatory analysis," *Proc. London Math. Soc.*, vol. 17, no. 2, pp. 75–115, 1918.

[41] G. Hardy and E. Wright, *An Introduction to the Theory of Numbers*. Oxford, U.K.: Oxford Univ. Press, 1985.

[42] I. Csiszár and P. Shields, "Redundancy rates for renewal and other processes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2065–2072, Nov. 1996.

[43] N. Jevtić, A. Orlitsky, and N. Santhanam, "A Lower bound on compression of unknown alphabets," manuscript, submitted for publication.

[44] A. Orlitsky and N. Santhanam, "Speaking of infinity," manuscript, submitted for publication.

[45] G. Shamir and L. Song, "On the entropy of patterns of iid sequences," in *Proc. 41st Annu. Allerton Conf. Communication, Control, and Computing*, Allerton, IL, Oct. 2003, pp. 160–170.

[46] G. Shamir, "Universal lossless compression with unknown alphabets—The average case," *IEEE Trans. Inform. Theory*, submitted for publication.