Universal compression of Markov and related sources over arbitrary alphabets

Anand K. Dhulipala[1] *Student Member, IEEE,* Alon Orlitsky[1,2] *Member, IEEE*

[1]ECE and [2]CSE Departments,
University of California, San Diego

{adhulipala, alon}@ucsd.edu

***Abstract*— Recent work has considered encoding a string by separately conveying its symbols and its pattern—the order in which the symbols appear. It was shown that the patterns of *i.i.d.* strings can be losslessly compressed with diminishing per-symbol redundancy. In this paper the pattern redundancy of distributions with memory is considered. Close lower and upper bounds are established on the pattern redundancy of strings generated by Hidden Markov Models with a small number of states, showing in particular that their per-symbol pattern redundancy diminishes with increasing string length. The upper bounds are obtained by analyzing the growth rate of the number of multi-dimensional integer partitions, and the lower bounds, using Hayman's Theorem.**

***Index Terms*— Hidden Markov Models, integer partitions, large alphabets, multi-dimensional partitions, patterns, redundancy, universal compression.**

I. Introduction

A WIDE variety of applications such as information storage, file compression, and data transmission involve compression of data. Shannon showed that every discrete source can be compressed no further than to its entropy, and that if the underlying distribution is known, this bound can be asymptotically achieved.

Yet in many applications, including those mentioned above, the data to be compressed is generated by an unknown source. To address this problem, the underlying distribution, though unknown, is assumed to belong to a known natural class $\mathcal{P}$ of distributions such as iid, Markov, or stationary distributions, but that the precise distribution within $\mathcal{P}$ is not known.

Compression is then performed using a *universal* compression algorithm designed with knowledge of only the class $\mathcal{P}$. The performance of the compression algorithm is often measured in terms of its *worst-case redundancy*, the largest extra number of bits it uses to encode any output of any distribution in $\mathcal{P}$. The worst-case redundancy $\hat{R}(\mathcal{P})$ of $\mathcal{P}$ is then defined as the lowest worst-case redundancy achieved by any universal compression algorithm for $\mathcal{P}$.

For the collection $\mathcal{I}_m^n$ of *i.i.d.* distributions over length-n strings from an alphabet of size m it was shown with increasing degrees of precision [1]–[7] that as n increases,

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

where Γ is the gamma function and the $o_m(1)$ term diminishes with increasing n at a rate determined by m. The redundancy therefore grows sub-linearly in the blocklength n, and the per-symbol redundancy $\hat{R}(\mathcal{I}_m^n)/n$ diminishes to zero, which implies that strings generated by an unknown *i.i.d.* distribution can be asymptotically compressed essentially as well as when the distribution is known.

However, in practice, for example in text, speech, or video compression, the alphabet size m is large compared to the blocklength n, and the resulting redundancy in (1) is high. In the limit, as m increases to infinity, Kieffer [8] showed that the per-symbol redundancy is unbounded.

Faced with Kieffer's result, researchers have either avoided direct compression of sequences over large or infinite alphabets or have circumvented the problem by considering more restricted collections of distributions.

An alternative encoding method that does not restrict the collection of distributions was considered in [9]–[11]. Instead of compressing the whole string at once, it conveys separately the string's dictionary—which reflects the symbols that appear and its pattern—which indicates the order in which the symbols appear.

Of the pattern and the dictionary parts describing a string the former has greater bearing on the problem. It was shown that the entropy rates of stationary [12] and stationary ergodic [13] processes are essentially equal to the corresponding entropy rates of the patterns they generate. It was further shown experimentally that for English text, the number of bits required to compress

This work was supported by grants from the National Science Foundation and The Ericsson Corporation.

the dictionary, and hence its redundancy, is negligible for large texts [14]. We therefore concentrate on pattern redundancy.

Patterns of strings generated by *i.i.d.* distributions have been studied in [15], [16]. It was shown that the redundancy of compressing patterns of *i.i.d.* strings over potentially infinite alphabets $\hat{R}(\mathcal{I}_\Psi^n)$ is bounded by

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

Therefore, the per-symbol redundancy of encoding patterns diminishes to zero as the blocklength increases.

In practice, the probability distributions underlying common sources are non-*i.i.d.* For example, in English text the probability of the current word depends on the preceding ones; in image coding, the value of the current pixel depends on the previous pixels; and the same is true with facsimile transmission as well. We therefore concentrate on compressing patterns of strings generated by distributions with memory.

After formally defining patterns in Section II, we investigate unconstrained Markov distributions in Section III. We show that the patterns of the strings they generate have unbounded per-symbol redundancy. Markov distributions defined over an unknown arbitrary alphabet have potentially infinite number of states. As the Markov patterns have unbounded per-symbol redundancy, we next consider sources with restricted number of states. In Section IV, we consider memory-constrained Markov distributions, and in particular *Hidden Markov Models*. Using a correspondence between patterns and multi-dimensional integer partitions, we upper and lower bound the pattern redundancy of d-state, order-k, HMMs over potentially infinite alphabets, showing that

$$\Omega\left(n^{\frac{d}{d+2}}+d\log n\right)\le\hat{R}(\mathcal{H}_\Psi^n)\le\mathcal{O}\left(n^{\frac{d}{d+1}}+d^{k+1}\log n\right).$$

Therefore the per-symbol redundancy of HMM patterns diminishes to zero whenever the number of states and the order of the HMMs are fixed or grow sufficiently slowly with n, irrespective of the alphabet size. The proof for the upper bound involves bounding the number of d-dimensional integer partitions of n by constructing its generating function while the proof for the lower bound uses Hayman's analysis.

These results can also be applied to the extensively studied [17]–[22] problem of estimating probability distribution from a sample of data. Recent work [23] defined the *attenuation* of an estimator to be the largest possible ratio of the probability assigned by the underlying distribution to that assigned by the estimator, normalized by the number of symbols. It was shown that for *i.i.d.* distributions there are block and sequential estimators whose attenuation approaches one; namely, as the length of the sequence increases, the per-symbol probability assigned by the estimator is as high as that assigned by the underlying distribution. The bounds we obtain can be used to generalize these results to distributions with memory. They show that for HMMs of order k with d states, the attenuation of any probability estimator is at least $2^{\Omega\left(n^{-\frac{2}{d+2}}\right)}$ and there exist block estimators whose attenuation is at most $2^{\mathcal{O}\left(n^{-\frac{1}{d+1}}\right)}$.

Note that unlike *i.i.d.* distributions, where efficient sequential algorithms that estimate pattern probability with diminishing redundancy have been derived [15], for HMMs we do not know of efficient sequential, or even block, algorithms. A natural approach to this problem is to use an Expectation-Maximization, or Baum-Welch, type algorithm. These algorithms are traditionally used to maximize the probability of sequences and can potentially be modified to maximize pattern probabilities. However, such algorithms are guaranteed to converge only to a local, and not necessarily to a global, maximum.

II. Definitions

We formally define redundancy and patterns.

Let a source X be distributed over a support set $\mathcal{X}$ according to a probability distribution p which is unknown except that it belongs to a given set of distributions $\mathcal{P}$. 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 $\log\frac{1}{q(x)}$, and that an optimal encoding that knows the distribution $p \in \mathcal{P}$ and achieves its entropy, allocates to x approximately $\log\frac{1}{p(x)}$ bits. The extra number of bits required to encode x when q is used instead of p is

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

We would like to find a q that requires the lowest number of extra bits, no matter which distribution $p \in \mathcal{P}$ generates the data. In other words, we want a q that achieves the *worst-case redundancy* defined as

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_q \max_{p\in\mathcal{P}} \max_{x\in\mathcal{X}} \log\frac{p(x)}{q(x)}. \tag{2}$$

In addition to worst-case redundancy, one can also define the *average-case redundancy* of $\mathcal{P}$, the lowest expected number of additional bits required by any encoder q. It is easy to show [15] that the average-case redundancy is always lower than the worst-case redundancy. Since our primary interest is to show that low redundancy can be achieved, we concentrate on evaluating the worst-case redundancy. Though it might be of interest to evaluate the average-case redundancy,

the techniques we use do not allow us to that at this time.

To evaluate redundancies, we frequently take recourse to a result by Shtarkov [24] that the distribution achieving $\hat{R}(\mathcal{P})$ is

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

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

$$\hat{R}(\mathcal{P}) = \log\left(\sum_{x\in\mathcal{X}} \sup_{p\in\mathcal{P}} p(x)\right). \tag{3}$$

In this paper we consider the redundancy of patterns, which are formally defined next. Let $\mathcal{A}$ be any alphabet. For $\overline{x} = x_1x_2\ldots x_n \in \mathcal{A}^n$, let

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

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

$$\imath_{\overline{x}}(x) \stackrel{\text{def}}{=} \inf\{|\mathcal{A}(\overline{x}_1^i)| : 1 \le i \le n \text{ and } x_i = x\},$$

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

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

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

$$\begin{aligned}
&\Psi(\mathcal{A}) = \{1\},\\
&\Psi(\mathcal{A}^2) = \{11, 12\},\\
&\Psi(\mathcal{A}^3) = \{111, 112, 121, 122\}, \text{ etc.}
\end{aligned}$$

Let

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

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

$$\Psi^* = \bigcup_{n=1}^{\infty} \Psi^n$$

be the set of all patterns. For example,

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

It is easy to see that an integer string $\overline{\psi}$ is a pattern iff every integer $i > 1$ appears only after the integer $i-1$ appeared at least once. For example, 1, 12, 1212 are patterns while 2, 13, 23 are not patterns.

Every distribution p over $\mathcal{A}^*$ induces the distribution p_Ψ over patterns in 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 a pattern $\overline{\psi} \in \Psi^*$. 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 (2), it follows that the worst-case pattern redundancy of $\mathcal{P}$, *i.e.*, the worst case redundancy of universally coding patterns generated according to an unknown distribution in $\mathcal{P}_\Psi$ is

$$\hat{R}(\mathcal{P}_\Psi) = \inf_{q_\Psi} \sup_{p\in\mathcal{P}} \sup_{\overline{\psi}\in\Psi^*} \log \frac{p_\Psi(\overline{\psi})}{q_\Psi(\overline{\psi})},$$

where q_Ψ is any distribution over Ψ^*.

III. Redundancy of Markov patterns

Recall that the per-symbol pattern redundancy of *i.i.d.* distributions tends to zero asymptotically. In this section we first show that the per-symbol pattern redundancy of first-order Markov distributions tends to infinity with the block length. It follows that the same also holds for higher order Markov distributions.

Let $\mathcal{M}^n$ denote the collection of *first-order Markov* distributions over length-n strings drawn from any alphabet, and let $\mathcal{M}^n_\Psi$ denote the collection of distributions over Ψ^n induced by distributions in $\mathcal{M}^n$. From Shtarkov's sum (3), the pattern redundancy of $\mathcal{M}^n$ is

$$\hat{R}(\mathcal{M}^n_\Psi) = \log\left(\sum_{\overline{\psi}\in\Psi^n} \sup_{p\in\mathcal{M}^n} p_\Psi(\overline{\psi})\right). \tag{4}$$

We show that the above sum is super-exponential in n and hence that the per-symbol pattern redundancy of $\mathcal{M}^n$ tends to infinity.

For simplicity we assume that n is even. Let $\tilde{\Psi}^n \subseteq \Psi^n$ be the collection of patterns of the form $\overline{\psi} =$

$12\ldots\frac{n}{2}\psi_1\psi_2\ldots\psi_{\frac{n}{2}}$ where $\psi_1\psi_2\ldots\psi_{\frac{n}{2}}$ is a permutation of $12\ldots\frac{n}{2}$. For example,

$$\tilde{\Psi}^2 = \{11\},$$
$$\tilde{\Psi}^4 = \{1212, 1221\},$$

and

$$\tilde{\Psi}^6 = \{123123, 123132, 123213, 123231, 123312, 123321\}.$$

Clearly,

$$|\tilde{\Psi}^n| = \left(\frac{n}{2}\right)!.$$

For every pattern $\overline{\psi} \in \tilde{\Psi}^n$ consider the distribution $\tilde{p} \in \mathcal{M}^n$ over $\{1, 2, \ldots, \frac{n}{2}\}$. Due to our choice of the support set for $\tilde{p}$ note that there exists a string whose pattern is identical to the string itself. Define

$$\tilde{p}(1 \mid \Lambda) = 1,$$
$$\tilde{p}(y \mid x) = \frac{c(y|x)}{2},$$

where Λ is the empty set and $c(y|x)$ is the number of times y occurs immediately following x in $\overline{\psi}$. Note that $c(y|x) \in \{0, 1, 2\}$, as every integer $x \in \{1, 2, \ldots \frac{n}{2}\}$ appears exactly twice in $\overline{\psi}$. Consequently, $\tilde{p}(y|x) \in \{0, \frac{1}{2}, 1\}$.

It follows that for all $\overline{\psi} \in \tilde{\Psi}$,

$$\sup_{p \in \mathcal{M}^n} p_\Psi(\overline{\psi}) \geq \tilde{p}_\Psi(\overline{\psi}) \geq \tilde{p}(\overline{\psi}) \geq \left(\frac{1}{2}\right)^n. \tag{5}$$

From (4) and (5),

$$\hat{R}(\mathcal{M}^n_\Psi) \geq \log\left(\left(\frac{n}{2}\right)!\left(\frac{1}{2}\right)^n\right) \geq \log\left(\frac{n}{8e}\right)^{\frac{n}{2}}.$$

Hence,

$$\frac{\hat{R}(\mathcal{M}^n_\Psi)}{n} \geq \frac{1}{2}\log\left(\frac{n}{8e}\right),$$

and the per-symbol pattern redundancy tends to infinity.

In the next section we consider Hidden Markov Models which have a smaller set of states unlike the Markov distributions.

IV. Redundancy of HMM patterns

In the rest of the paper we concentrate on deriving bounds for the redundancy of compressing HMM patterns. We first develop an analogy between HMM patterns and multi-dimensional integer partitions in the context of the upper bound, and use this analogy to derive the upper and lower bounds.

A. HMM patterns and multi-dimensional partitions

Let $\mathcal{H}^n$ denote the collection of HMMs of order k and d states over length n strings drawn from any alphabet $\mathcal{A}$ and let $\mathcal{H}^n_\Psi$ be the collection of distributions over Ψ^n induced by distributions in $\mathcal{H}^n$. Denote by $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d\}$ the set of states of the HMMs. Let $\overline{\psi} = \psi_1\psi_2\ldots\psi_n \in \Psi^n$ be any pattern and let $\overline{s} = s_1 s_2 \ldots s_n \in \mathcal{S}^n$ be any hidden state sequence. Using (3), the worst-case redundancy of $\mathcal{H}^n_\Psi$ is given by

$$\hat{R}(\mathcal{H}^n_\Psi) = \log\left(\sum_{\overline{\psi} \in \Psi^n} \sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi})\right), \tag{6}$$

where

$$p_\Psi(\overline{\psi}) = \sum_{\overline{s} \in \mathcal{S}^n} p_\Psi(\overline{\psi} \mid \overline{s}) p(\overline{s}) \tag{7}$$

is the probability of the pattern, and

$$p(\overline{s}) = p(s_1 s_2 \ldots s_k) \prod_{i=k+1}^{n} p(s_i \mid s_{i-1} s_{i-2} \ldots s_{i-k})$$

is the probability of the state sequence.

Using (7)

$$\begin{aligned}
&\sum_{\overline{\psi} \in \Psi^n} \sup_{p \in \mathcal{H}^n} \left(\sum_{\overline{s} \in \mathcal{S}^n} p(\overline{s}) p_\Psi(\overline{\psi} \mid \overline{s})\right) \\
&\quad \leq \sum_{\overline{\psi} \in \Psi^n} \sum_{\overline{s} \in \mathcal{S}^n} \left(\sup_{p \in \mathcal{H}^n} p(\overline{s})\right)\left(\sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi} \mid \overline{s})\right) \\
&\quad = \sum_{\overline{s} \in \mathcal{S}^n} \sup_{p \in \mathcal{H}^n} p(\overline{s}) \sum_{\overline{\psi} \in \Psi^n} \sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi} \mid \overline{s}).
\end{aligned} \tag{8}$$

Hence,

$$\hat{R}(\mathcal{H}^n_\Psi) \leq \log\left(\sum_{\overline{s} \in \mathcal{S}^n} \sup_{p \in \mathcal{H}^n} p(\overline{s}) \sum_{\overline{\psi} \in \Psi^n} \sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi} \mid \overline{s})\right). \tag{9}$$

To show that redundancy grows sub-linearly with n, we group together the terms in Shtarkov's sum with same probability, and hence the same maximizing distribution, and show that the number of such groups is sub-exponential in n. We enumerate these groups for HMMs in subsection IV-B.

Define

$$m = \sup_i \psi_i$$

to be the number of distinct symbols in $\overline{\psi}$, and

$$\pi_j = |\{i : s_i = \mathcal{S}_j\}| \tag{10}$$

to be the number of times state $\mathcal{S}_j$ appears in $\overline{s}$.

For $0 \leq i \leq m$ and $0 \leq j \leq d$, define the *multiplicity*

$$\mu_{i,j} = |\{l : \psi_l = i \text{ and } s_l = \mathcal{S}_j\}|$$

to be the number of times state $\mathcal{S}_j$ generated the ith new symbol, and let

$$\Phi(\overline{\psi}, \overline{s}) \stackrel{\text{def}}{=} \{(\mu_{1,1}, \ldots, \mu_{1,d}), \ldots, (\mu_{m,1}, \ldots, \mu_{m,d})\}$$

be the set of *multiplicity vectors* of the pattern $\overline{\psi}$ and state sequence $\overline{s}$.

Since

$$\begin{aligned} p_\Psi(\overline{\psi} \mid \overline{s}) &= p(\{\overline{x} \in \mathcal{A}^n : \Psi(\overline{x}) = \overline{\psi}\} \mid \overline{s}) \\ &= \sum_{\substack{x_1 x_2 \ldots x_m \in \mathcal{A} \\ x_i \neq x_j \text{ for } i \neq j}} \prod_{j=1}^{d} \prod_{i=1}^{m} p^{\mu_{i,j}}(x_i \mid \mathcal{S}_j), \end{aligned}$$

it is apparent that given a state sequence $\overline{s}$, any two patterns with the same set of multiplicity vectors have the same conditional probability.

For example, let $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2\}$ and $\overline{s} = \mathcal{S}_1\mathcal{S}_2\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2$. Then

$$\Phi(11122, \overline{s}) = \Phi(12221, \overline{s}) = \{(1,1), (1,2)\},$$

and indeed,

$$\begin{aligned} &p_\Psi(11122 \mid \mathcal{S}_1\mathcal{S}_2\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2) \\ &\quad = \sum_{x_1 \neq x_2 \in \mathcal{A}} p(x_1 \mid \mathcal{S}_1) p(x_2 \mid \mathcal{S}_1) p^2(x_1 \mid \mathcal{S}_2) p(x_2 \mid \mathcal{S}_2) \\ &\quad = \sum_{x_1 \neq x_2 \in \mathcal{A}} p(x_2 \mid \mathcal{S}_1) p(x_1 \mid \mathcal{S}_1) p^2(x_2 \mid \mathcal{S}_2) p(x_1 \mid \mathcal{S}_2) \\ &\quad = p_\Psi(12221 \mid \mathcal{S}_1\mathcal{S}_2\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2). \end{aligned}$$

Each set of multiplicity vectors can be viewed as an *unordered d-dimensional partition* of n, a collection

$$\{(\phi_{1,1}, \phi_{1,2}, \ldots, \phi_{1,d}), \ldots, (\phi_{m,1}, \phi_{m,2}, \ldots, \phi_{m,d})\},$$

of non-zero d-dimensional vectors such that $\phi_{i,j} \geq 0$ and

$$\sum_{i=1}^{m} \sum_{j=1}^{d} \phi_{i,j} = n.$$

For example, $\{(2,0)\}$, and $\{(1,0),(0,1)\}$ are both 2-dimensional partitions of 2. Note that both $\{(1,0),(0,1)\}$ and $\{(0,1),(1,0)\}$ represent the same 2-dimensional partition of 2 as the partitions are unordered collections of vectors.

A d-dimensional partition ϕ is *compatible* with a state sequence $\overline{s} \in \mathcal{S}^n$ if for all $0 \leq j \leq d$,

$$\pi_j = \sum_{i=1}^{m} \phi_{i,j},$$

where π_j was defined in (10) to be the number of times state $\mathcal{S}_j$ appeared. Let $\Phi_d(\overline{s})$ denote the set of d-dimensional partitions of n compatible with $\overline{s}$ and let

$$\Phi_d^n = \bigcup_{\overline{s} \in \mathcal{S}^n} \Phi_d(\overline{s})$$

denote the set of all d-dimensional partitions of n. For example,

$$\begin{aligned} \Phi_2(\mathcal{S}_1\mathcal{S}_2) &= \{\{(1,1)\}, \{(1,0),(0,1)\}\} \\ \Phi_2(\mathcal{S}_1\mathcal{S}_2\mathcal{S}_2) &= \{\{(1,2)\}, \{(1,1),(0,1)\}, \{(1,0), \\ &\qquad (0,2)\}, \{(1,0),(0,1),(0,1)\}\}, \end{aligned}$$

and

$$\begin{aligned} \Phi_2^1 &= \{\{(1,0)\}, \{(0,1)\}\}, \\ \Phi_2^2 &= \{\{(2,0)\}, \{(1,1)\}, \{(0,2)\}, \{(1,0), \\ &\qquad (1,0)\}, \{(1,0),(0,1)\}, \{(0,1),(0,1)\}\}. \end{aligned}$$

Note the distinction between Φ_d^n, $\Phi_d(\overline{s})$, $\Phi(\overline{\psi}, \overline{s})$ and $\Phi_d^n(\overline{\varphi})$ (defined later in the section). We use similar notation as they all denote the set of d-dimensional partitions of n compatible with the variables they are defined over.

For a fixed $\overline{s}$ and a compatible $\phi \in \Phi_d(\overline{s})$, consider the collection

$$\Psi(\overline{s}, \phi) \stackrel{\text{def}}{=} \left\{\overline{\psi} \in \Psi^n : \Phi(\overline{\psi}, \overline{s}) = \phi\right\}$$

of all patterns whose set of multiplicity vectors with respect to $\overline{s}$ is ϕ.

It is clear from the earlier discussion that all the patterns $\overline{\psi} \in \Psi(\overline{s}, \phi)$ have the same conditional probability $p(\overline{\psi} \mid \overline{s})$. For example, let $\overline{s} = \mathcal{S}_1\mathcal{S}_2\mathcal{S}_2\mathcal{S}_1\mathcal{S}_2$ and $\phi = \{(1,1),(1,2)\}$. Then,

$$\Psi(\overline{s}, \phi) = \{11122, 11221, 11222, 12121, 12122, 12221\},$$

and

$$p(11122 \mid \overline{s}) = p(12221 \mid \overline{s}) = p(11222 \mid \overline{s}) = \ldots.$$

Define

$$p_\Psi(\phi \mid \overline{s}) = \sum_{\overline{\psi} \in \Psi(\overline{s}, \phi)} p_\Psi(\overline{\psi} \mid \overline{s}).$$

Since all patterns in $\Psi(\overline{s}, \phi)$ have the same conditional probability, they have the same maximizing probability distribution as well. Therefore,

$$\sup_{p \in \mathcal{H}^n} p_\Psi(\phi \mid \overline{s}) = \sum_{\overline{\psi} \in \Psi(\overline{s}, \phi)} \sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi} \mid \overline{s}),$$

and

$$\sum_{\overline{\psi} \in \Psi^n} \sup_{p \in \mathcal{H}^n} p_\Psi(\overline{\psi} \mid \overline{s}) = \sum_{\phi \in \Phi_d^n(\overline{s})} \sup_{p \in \mathcal{H}^n} p_\Psi(\phi \mid \overline{s}).$$

Hence, from (8)

$$\begin{aligned}\hat{R}(\mathcal{H}_\Psi^n) &\le \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})\sum_{\phi\in\Phi_d^n(\overline{s})}\sup_{p\in\mathcal{H}^n} p_\Psi(\phi\mid\overline{s})\right)\\ &\le \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})\sum_{\phi\in\Phi_d^n(\overline{s})} 1\right)\\ &= \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})|\Phi_d^n(\overline{s})|\right)\\ &\le \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})|\Phi_d^n|\right)\\ &= \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})\right)+\log|\Phi_d^n|\\ &= \mathcal{I}_1+\mathcal{I}_2, \end{aligned}\tag{11}$$

where

$$\mathcal{I}_1 \stackrel{\text{def}}{=} \log\left(\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s})\right),$$
$$\mathcal{I}_2 \stackrel{\text{def}}{=} \log|\Phi_d^n|.$$

B. Upper bound

We first upper bound $\mathcal{I}_1$ by showing that the probability terms within the logarithm can be arranged into subexponential number of groups, with the terms in each group adding up to at most 1 . We then upper bound $\mathcal{I}_2$ by showing that the number of d-dimensional partitions of n is subexponential. In literature, the set of equiprobable strings is called a type set and the empirical distribution of the strings in a type set is called its type. For a review on the method of types see [25] and the references there in for some results on Markov types.

Note that

$$\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{H}^n} p(\overline{s}) = \sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{M}^n} p(\overline{s}),\tag{12}$$

where $\mathcal{M}^n$ is the collection of Markov distributions of order k over length-n strings drawn from $\mathcal{S} = \{\mathcal{S}_1,\ldots,\mathcal{S}_d\}$, the set of states of the HMMs, $\mathcal{H}^n$.

For the sake of convenience we assume that the context of s_1, the k symbols preceding s_1 are known. Let $c_{\mathcal{S}_i|\overline{\mathcal{S}}}$ denote the number times $\mathcal{S}_i$ appears in $\overline{s}$ following $\overline{\mathcal{S}}\in\mathcal{S}^k$. Observe that

$$p(\overline{s}) = \prod_{\overline{\mathcal{S}}\in\mathcal{S}^k}\prod_{i=1}^{d} p^{c_{\mathcal{S}_i|\overline{\mathcal{S}}}}(\mathcal{S}_i\mid\overline{\mathcal{S}})$$

depends only on the counts $c_{\mathcal{S}_i|\overline{\mathcal{S}}}$ which are totally d^{k+1} in number. Let $\overline{c}$ denote the vector whose elements are these d^{k+1} numbers in lexicographic order. Note that the total number of such vectors is the number of ways n can be divided into d^{k+1} non-negative integers which equals $\binom{n+d^{k+1}-1}{n}$.

Define the probability of the vector $\overline{c}$, $p(\overline{c})$, as the sum of the probabilities of all the state sequences $\overline{s}$ which have $\overline{c}$ as their count vector. Since all these sequences with the same count vector $\overline{c}$ have the same probability, they have the same maximizing probability as well. Therefore,

$$\begin{aligned}\sum_{\overline{s}\in\mathcal{S}^n}\sup_{p\in\mathcal{M}^n} p(\overline{s}) &= \sum_{\overline{c}}\sup_{p\in\mathcal{M}^n} p(\overline{c})\\ &\le \sum_{\overline{c}} 1\\ &= \binom{n+d^{k+1}-1}{n}.\end{aligned}$$

Hence,

$$\mathcal{I}_1 \le \log\binom{n+d^{k+1}-1}{n} \le d^{k+1}\log n.\tag{13}$$

To upper bound $\mathcal{I}_2$ we enumerate the number of d-dimensional partitions of n. For the sake of convenience and brevity, we use a different notation for d-dimensional partitions when $d=1$. Let $\overline{\varphi} = (\varphi_1,\varphi_2,\ldots,\varphi_n)$ be a one-dimensional unordered partition of n such that

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

Observe that $\overline{\varphi}$ can also be represented as

$$\{\underbrace{1,1,\ldots,1}_{\varphi_1},\underbrace{2,2,\ldots,2}_{\varphi_2},\ldots,\underbrace{n,n,\ldots,n}_{\varphi_n}\},$$

which is a 1-dimensional partition of n by definition. Let Φ^n denote the set of all 1-dimensional partitions of n. For example,

$$\begin{aligned}\Phi^1 &= \{1\},\\ \Phi^2 &= \{(0,1),(2,0)\},\\ \Phi^3 &= \{(0,0,1),(1,1,0),(3,0,0)\}.\end{aligned}$$

A d-dimensional partition ϕ is said to be of type $\overline{\varphi}$ if

$$|\{i:\sum_{j=1}^{d}\phi_{i,j}=\mu\}| = \varphi_\mu,$$

i.e., the number of vectors in ϕ whose elements sum to μ is φ_μ. Denote by $\Phi_d^n(\overline{\varphi})$ the set of d-dimensional partitions of n whose type is $\overline{\varphi}$. For example,

$$\begin{aligned}\Phi_2^2(\overline{\varphi}=(0,1)) &= \{\{(1,1)\},\{(2,0)\},\{(0,2)\}\},\\ \Phi_2^2(\overline{\varphi}=(2,0)) &= \{\{(1,0),(1,0)\},\{(1,0),(0,1)\},\\ &\qquad \{(0,1),(0,1)\}\}.\end{aligned}$$

Number of distinct d-dimensional vectors whose elements (which are non-negative integers) sum to μ is

$\binom{\mu+d-1}{\mu}$ and from these φ_μ vectors, not necessarily distinct, can be is chosen in $\binom{\varphi_\mu+\binom{\mu+d-1}{\mu}-1}{\varphi_\mu}$ ways. So, the number of d-dimensional partitions ϕ which are of type $\overline{\varphi}$ is

$$\Phi_d^n(\overline{\varphi}) = \prod_{\mu=1}^{n} \binom{\varphi_\mu + \binom{\mu+d-1}{\mu} - 1}{\varphi_\mu}.$$

Observe that

$$|\Phi_d^n| = \sum_{\overline{\varphi}\in\Phi^n} |\Phi_d^n(\overline{\varphi})|.$$

Hence, the total number of d-dimensional partitions is

$$|\Phi_d^n| = \sum_{\overline{\varphi}\in\Phi^n} \prod_{\mu=1}^{n} \binom{\varphi_\mu + \binom{\mu+d-1}{\mu} - 1}{\varphi_\mu}. \tag{14}$$

Let $P_d(x)$ denote the generating function of $|\Phi_d^n|$ such that

$$P_d(x) = \sum_{n=0}^{\infty} P_d^n\, x^n,$$

where

$$P_d^n \stackrel{\text{def}}{=} |\Phi_d^n|.$$

Using (14), for $0 < x < 1$,

$$\begin{aligned} P_d(x) &= \sum_{n=1}^{\infty} \sum_{\overline{\varphi}\in\Phi^n} \prod_{\mu=1}^{n} \left(\binom{\varphi_\mu + \binom{\mu+d-1}{\mu} - 1}{\varphi_\mu} x^{\mu\,\varphi_\mu} \right) \\ &= \prod_{\mu=1}^{\infty} \left(\sum_{\varphi_\mu=0}^{\infty} \binom{\varphi_\mu + \binom{\mu+d-1}{\mu} - 1}{\varphi_\mu} x^{\mu\,\varphi_\mu} \right) \\ &= \prod_{\mu=1}^{\infty} (1-x^\mu)^{-\binom{\mu+d-1}{\mu}}, \end{aligned} \tag{15}$$

where the last equality follows due to the *generalized binomial theorem*. Since P_d^n is monotonically increasing in n,

$$\begin{aligned} P_d(x) &= \sum_{n=0}^{\infty} P_d^n\, x^n \\ &> P_d^n\,(x^n + x^{n+1} + \cdots) \\ &= P_d^n \frac{x^n}{(1-x)}. \end{aligned} \tag{16}$$

Also note that for $\mu \geq 1$,

$$\binom{\mu+d-1}{\mu} \leq d \binom{\mu+d-2}{\mu-1} \tag{17}$$

From (15), (16) and (17) we get

$$\begin{aligned} \log\left(P_d^n \frac{x^n}{(1-x)}\right) &< \sum_{\mu=1}^{\infty} \binom{\mu+d-1}{\mu} \log(1-x^\mu)^{-1} \\ &= \sum_{\mu=1}^{\infty} \binom{\mu+d-1}{\mu} \sum_{j=1}^{\infty} \frac{x^{\mu\,j}}{j} \\ &= \sum_{j=1}^{\infty} \frac{1}{j} \sum_{\mu=1}^{\infty} \binom{\mu+d-1}{\mu} x^{\mu\,j} \\ &\leq \sum_{j=1}^{\infty} \frac{1}{j} \sum_{\mu=1}^{\infty} d \binom{\mu+d-2}{\mu-1} x^{\mu\,j} \\ &= \sum_{j=1}^{\infty} \frac{1}{j} \sum_{\mu=0}^{\infty} d x^j \binom{\mu+d-1}{\mu} x^{\mu\,j} \\ &= \sum_{j=1}^{\infty} \frac{d}{j} x^j (1-x^j)^{-d}. \end{aligned} \tag{18}$$

Observe that

$$\begin{aligned} (1-x^j) &= (1-x)(1+x+\cdots+x^{j-1}) \\ &\geq j\,(1-x)x^{j-1}. \end{aligned}$$

Rewriting,

$$(1-x^j)^{-1} \leq \frac{1}{j}\,(1-x)^{-1} x^{1-j}. \tag{19}$$

From (18) and (19) we get

$$\begin{aligned} \log\left(P_d^n \frac{x^n}{1-x}\right) &\leq \sum_{j=1}^{\infty} \frac{d\,x(1-x^j)^{-(d-1)}}{j^2} (1-x)^{-1} \\ &< d\,(1-x)^{-d} \sum_{j=1}^{\infty} \frac{1}{j^2} \\ &= \frac{\pi^2}{6} d\,(1-x)^{-d}. \end{aligned}$$

Substituting $x = \frac{(\frac{6\,n}{\pi^2 d^2})^{\frac{1}{d+1}} - 1}{(\frac{6\,n}{\pi^2 d^2})^{\frac{1}{d+1}}}$ for sufficiently large n and for $d = o(\log(n))$ we have,

$$\log(P_d^n) \leq \frac{d+1}{d} \left(\frac{\pi^2 d^2}{6}\right)^{\frac{1}{d+1}} n^{\frac{d}{d+1}} + \mathcal{O}(n^{\frac{d-1}{d+1}}). \tag{20}$$

Therefore, from (11), (13) and (20),

$$\hat{R}(\mathcal{H}^n) \leq 3n^{\frac{d}{d+1}} + d^{k+1} \log n + \mathcal{O}(n^{\frac{d-1}{d+1}}).$$

which is sublinear in n for k that grows sufficiently slowly with n.

C. Lower Bound

We derive the lower bound by bounding the maximum likelihood probabilities of patterns and incorporating it into Shtarkov's sum.

Consider any state sequence $\overline{s}$. Using (7), for every $\overline{\psi} \in \Psi$

$$\begin{aligned}\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi}) &= \sup_{p\in\mathcal{H}^n}\left(\sum_{\tilde{\overline{s}}\in\mathcal{S}^n} p_\Psi(\overline{\psi}\mid\tilde{\overline{s}})p(\tilde{\overline{s}})\right)\\ &\geq \sup_{p\in\mathcal{H}^n}\left(p_\Psi(\overline{\psi}\mid\overline{s})p(\overline{s})\right)\\ &= \left(\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi}\mid\overline{s})\right)\left(\sup_{p\in\mathcal{H}^n} p(\overline{s})\right)\\ &= \left(\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi}\mid\overline{s})\right)\left(\sup_{p\in\mathcal{M}^n} p(\overline{s})\right). \quad (21)\end{aligned}$$

Note that the last two equalities hold due to (12) and due to the fact that $p_\Psi(\overline{\psi}\mid\overline{s})$ and $p(\overline{s})$ are functions of disjoint sets of variables and hence can be optimized separately.

Also,

$$\begin{aligned}\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi}\mid\overline{s}) &= \sup_{p\in\mathcal{H}^n} p(\{\overline{x}\in\mathcal{A}^n : \Psi(\overline{x})=\overline{\psi}\}\mid\overline{s})\\ &\geq \sup_{p\in\mathcal{H}^n}\sup_{\{\overline{x}\in\mathcal{A}^n:\Psi(\overline{x})=\overline{\psi}\}} p(\overline{x}\mid\overline{s})\\ &= \prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}. \quad (22)\end{aligned}$$

Therefore, from (21) and (22) we get,

$$\begin{aligned}\sum_{\overline{\psi}\in\Psi^n}\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi}) &\geq \sum_{\overline{\psi}\in\Psi^n}\prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}\sup_{p\in\mathcal{M}^n} p(\overline{s})\\ &= \sum_{\phi\in\Phi_d^n(\overline{s})}\sum_{\overline{\psi}\in\Psi^n(\overline{s},\phi)}\prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}\sup_{p\in\mathcal{H}^n} p(\overline{s}). \quad (23)\end{aligned}$$

Let $\overline{\pi} = (\pi_1, \pi_2, \ldots, \pi_d)$ be an ordered integer partition of n into d parts, *i.e.,*

$$\sum_{j=1}^{d}\pi_j = n.$$

Define

$$\overline{s}(\overline{\pi}) \stackrel{\text{def}}{=} \underbrace{\mathcal{S}_1\mathcal{S}_1\ldots\mathcal{S}_1}_{\pi_1}\underbrace{\mathcal{S}_2\mathcal{S}_2\ldots\mathcal{S}_2}_{\pi_2}\ldots\underbrace{\mathcal{S}_d\mathcal{S}_d\ldots\mathcal{S}_d}_{\pi_d}.$$

Note that the sets $\Phi_d^n(\overline{s}(\overline{\pi}))$ of d-dimensional partitions of n compatible with the state sequence $\overline{s}(\overline{\pi})$ are disjoint. Also, the total number of d-dimensional partitions of n can be expressed as

$$\Phi_d^n = \bigcup_{\overline{\pi}}\Phi_d^n(\overline{s}(\overline{\pi})) \quad (24)$$

and the number of ordered partitions of n into d parts is given by

$$|\{\overline{\pi} : \sum_{j=1}^{d}\pi_j = n\}| = \binom{n+d-1}{n}. \quad (25)$$

Since (23) is true for all $\overline{s}$, it is also true for the state sequences $\overline{s}(\overline{\pi})$ in particular. Replacing right hand side of (23) by its average over all $\overline{s}(\overline{\pi})$ and using (24) and (25) and recalling that $\Psi^n(\overline{s}(\overline{\pi}),\phi)$ represents the collection of all patterns whose set of multiplicity vectors with respect to $\overline{s}(\overline{\pi})$ is ϕ, we get

$$\begin{aligned}&\sum_{\overline{\psi}\in\Psi^n}\sup_{p\in\mathcal{H}^n} p_\Psi(\overline{\psi})\\ &\geq \frac{1}{\binom{n+d-1}{n}}\sum_{\phi\in\Phi_d^n}\sum_{\overline{\psi}\in\Psi^n(\overline{s}(\overline{\pi}),\phi)}\prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}\inf_{\overline{\pi}}\sup_{p\in\mathcal{M}^n} p(\overline{s}(\overline{\pi}))\\ &= \frac{1}{\binom{n+d-1}{n}}\sum_{\phi\in\Phi_d^n}|\{\Psi^n(\overline{s}(\overline{\pi}),\phi)\}|\prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}\inf_{\overline{\pi}}\sup_{p\in\mathcal{M}^n} p(\overline{s}(\overline{\pi})), \quad (26)\end{aligned}$$

Note that ϕ uniquely determines $\overline{s}(\overline{\pi})$.

We can show that

$$|\{\Psi^n(\overline{s}(\overline{\pi}),\phi)\}| = \prod_{j=1}^{d}\frac{\pi_j!}{\prod_{i=1}^{m}\mu_{i,j}!}\,\frac{1}{\prod_{p=1}^{q}r_p!}, \quad (27)$$

where the q distinct vectors of ϕ occur $r_1, r_2, \ldots, r_q$ number of times respectively. It can be shown that

$$\log\binom{n+d-1}{n} \leq d\log n, \quad (28)$$

and (see Appendix I for proof)

$$\inf_{\overline{\pi}}\sup_{p\in\mathcal{M}^n} p(\overline{s}(\overline{\pi})) \geq \Omega(d^{d-k}e^{-d}n^{-d}) \quad (29)$$

Using (27) and the inequality

$$\prod_{j=1}^{d}\frac{\pi_j!}{\pi_j^{\pi_j}} \geq \frac{n!}{n^n}$$

we get

$$\begin{aligned}&\sum_{\phi\in\Phi_d^n}|\{\Psi^n(\overline{s}(\overline{\pi}),\phi)\}|\prod_{j=1}^{d}\prod_{i=1}^{m}\left(\frac{\mu_{i,j}}{\pi_j}\right)^{\mu_{i,j}}\\ &\geq \frac{n!}{n^n}\sum_{\phi\in\Phi_d^n}\prod_{j=1}^{d}\prod_{i=1}^{m}\frac{\mu_{i,j}^{\mu_{i,j}}}{\mu_{i,j}!}\frac{1}{\prod_{p=1}^{q}r_p!}\\ &= \frac{n!}{n^n}\sum_{\overline{\varphi}\in\Phi^n}\sum_{\phi\in\Phi_d^n(\overline{\varphi})}\prod_{j=1}^{d}\prod_{i=1}^{m}\frac{\mu_{i,j}^{\mu_{i,j}}}{\mu_{i,j}!}\frac{1}{\prod_{p=1}^{q}r_p!}\\ &\stackrel{\text{def}}{=} \frac{n!}{n^n}g[n]\end{aligned}$$

where $g[n]$ is the coefficient of x^n in

$$g(x) = \sum_{i=0}^{\infty} g[n]x^n.$$

Using simple combinatorial arguments it can be shown that

$$g[n] = \sum_{\overline{\varphi}=(\varphi_1,\varphi_2,\ldots,\varphi_n)\in\Phi^n} \prod_{\mu=1}^{n} \frac{(h[\mu])^{\varphi_\mu}}{\varphi_\mu!},$$

where $h[\mu]$ is the coefficient of x^μ in

$$h(x) = \left(\sum_{i=0}^{\infty} \frac{i^i}{i!}\, x^i\right)^d.$$

Therefore,

$$\begin{aligned} g(x) &= \sum_{r=1}^{\infty} \sum_{\overline{\varphi}=(\varphi_1,\varphi_2,\ldots,\varphi_r)\in\Phi^r} \prod_{\mu=1}^{r} \frac{(h[\mu]x^\mu)^{\varphi_\mu}}{\varphi_\mu!} \\ &= \prod_{\mu=1}^{\infty} \left(\sum_{\varphi_\mu=0}^{\infty} \frac{(h[\mu]x^\mu)^{\varphi_\mu}}{\varphi_\mu!}\right) \\ &= \prod_{\mu=1}^{\infty} \exp(h[\mu]x^\mu) \\ &= \exp\left(\sum_{\mu=1}^{\infty} h[\mu]x^\mu\right) \\ &= \exp(h(x)-1) \end{aligned}$$

Hayman [26] developed a technique to compute the asymptotics of coefficients of the power series of functions which satisfy certain admissibility criteria. For a discussion of similar techniques in analysis, see [27]. It can be shown that $g(x)$ is admissible for Hayman's theorem. Using Hayman's analysis we show that (see Appendix II)

$$\log(\frac{n!}{n^n} g[n]) = \frac{d+2}{2} \log e \Big(\frac{n}{d}\Big)^{\frac{d}{d+2}} + \mathcal{O}\Big(n^{\frac{d-1}{d+2}}\Big). \quad (30)$$

Using (6), (26), (28), (29), (30), for fixed or sufficiently slowly growing d and k,

$$\hat{R}(\mathcal{H}_\Psi^n) \geq \Omega\Big(n^{\frac{d}{d+2}} + d\log n\Big).$$

We note that the class of processes that are Markov renewal of order $d-1$ has similar redundancy rate bounds [28].

Appendix I

Recall that

$$\overline{\pi} = (\pi_1, \pi_2, \ldots, \pi_d)$$

is an ordered partition of n and

$$\begin{aligned} \overline{s}(\overline{\pi}) &= s_1 s_2 \ldots s_n \\ &= \underbrace{\mathcal{S}_1 \ldots \mathcal{S}_1}_{n_1} \underbrace{\mathcal{S}_2 \ldots \mathcal{S}_2}_{n_2} \ldots \underbrace{\mathcal{S}_d \ldots \mathcal{S}_d}_{n_d} \end{aligned}$$

is a sequence of states generated by a Markov process of order k.

It is easy to see that a context, a k-tuple $s_j, \ldots, s_{j+k-1}$, occurs more than once if only if $n_i > k$ and $s_j, \ldots, s_{j+k-1} = \mathcal{S}_i \ldots \mathcal{S}_i$ for some i, in which case it is followed by $\mathcal{S}_i$ $n_i - k$ times and $\mathcal{S}_l$ once where

$$l \stackrel{\text{def}}{=} \min_{r>i,\, n_r>0} r$$

Consider the Markov distribution $\tilde{p}$ such that

$$\tilde{p}(s_1, \ldots s_k) = \frac{1}{d^k}$$

and if $n_i > k$,

$$\begin{aligned} \tilde{p}(\mathcal{S}_i | \underbrace{\mathcal{S}_i \ldots \mathcal{S}_i}_{k}) &= \frac{n_i - k}{n_i - k + 1} \\ \tilde{p}(\mathcal{S}_l | \underbrace{\mathcal{S}_i \ldots \mathcal{S}_i}_{k}) &= \frac{1}{n_i - k + 1} \end{aligned}$$

and for a context $s_j \ldots s_{j+k-1}$ which appears only once,

$$\tilde{p}(s_{j+k} | s_j \ldots s_{j+k-1}) = 1$$

Note that $\tilde{p}$ maximizes $p(\overline{s}(\overline{\pi}))$ over all $p \in \mathcal{M}^n$. Therefore,

$$\begin{aligned} \sup_{p\in\mathcal{M}^n} p(\overline{s}(\overline{\pi})) &= \tilde{p}(\overline{s}(\overline{\pi})) \\ &= \frac{1}{d^k} \prod_{i,\, n_i>k} \frac{(n_i-k)^{n_i-k}}{(n_i-k+1)^{n_i-k+1}} \\ &= \frac{1}{d^k} 2^{-\sum_{i,\, n_i>k} (n_i-k+1) H(\frac{n_i-k}{n_i-k+1})} \\ &\geq \frac{1}{d^k} 2^{-\tilde{n} H(\frac{\tilde{d}}{\tilde{n}})} \\ &= \frac{1}{d^k} \left(\frac{\tilde{d}}{\tilde{n}}\right)^{\tilde{d}} \left(1 - \frac{\tilde{d}}{\tilde{n}}\right)^{\tilde{n}-\tilde{d}} \end{aligned}$$

where $\tilde{n} = \sum_{i,\, n_i>k} (n_i - k + 1)$, $\tilde{d} = \sum_{i,\, n_i>k} 1$ and the inequality holds since the binomial entropy function is concave. Hence, for fixed or sufficiently slowly increasing d and k,

$$\inf_{\overline{\pi}} \sup_{p\in\mathcal{M}^n} p(\overline{s}(\overline{\pi})) \geq \Omega(d^{d-k} e^{-d} n^{-d})$$

APPENDIX II

Let

$$g(z) = \exp\left(\left(\sum_{i=0}^{\infty} \frac{i^i}{i!} z^i\right)^d - 1\right).$$

Define

$$a(z) \stackrel{\text{def}}{=} \frac{d \log g(z)}{d \log z}, \quad \text{and} \quad b(z) \stackrel{\text{def}}{=} z \frac{da(z)}{dz}$$

and let x_n be a solution of $a(x_n) = n$. If g satisfies certain admissibility criteria [27] then,

$$g[n] \sim \frac{g(x_n)}{x_n^n \sqrt{2\pi b(x_n)}}. \tag{31}$$

We evaluate $a(z)$ using the related "tree function" [27]

$$T(z) = \sum_{k=1}^{\infty} \frac{k^{k-1} z^k}{k!},$$

which satisfies

$$T(z) = z e^{T(z)}. \tag{32}$$

Therefore,

$$\sum_{k=0}^{\infty} \frac{k^k z^k}{k!} = \frac{1}{1 - T(z)}. \tag{33}$$

By differentiating (32) and (33) and using the absolute convergence of the series, we obtain

$$a(z) = \frac{dT(z)}{(1 - T(z))^{d+2}}$$

and

$$b(z) = \frac{dT(z) + d(d+1)T^2(z)}{(1 - T(z))^{d+4}}.$$

At $z = \frac{1}{e}$ we have the following singular expansion [27]

$$\frac{1}{1 - T(z)} = \frac{1}{\sqrt{2(1 - ez)}} + \frac{1}{3} - \frac{\sqrt{2}}{24}\sqrt{1 - ez} + \mathcal{O}(1 - ez)$$

Consequently,

$$a(z) = \frac{d}{(2(1 - ez))^{\frac{d+2}{2}}} + \mathcal{O}\left(\frac{1}{(2(1 - ez))^{\frac{d+1}{2}}}\right)$$

and the solution to $a(x_n) = n$ is

$$x_n = \frac{1}{e}\left(1 - \frac{1}{2\left(\frac{n}{d}\right)^{\frac{2}{d+2}}}\right) + \mathcal{O}\left(\frac{1}{n^{\frac{3}{d+2}}}\right) \tag{34}$$

Therefore, using (31) and (34)

$$\log\left(\frac{n!}{n^n} g[n]\right) = \frac{d+2}{2} \log e \left(\frac{n}{d}\right)^{\frac{d}{d+2}} + \mathcal{O}\left(n^{\frac{d-1}{d+2}}\right).$$

ACKNOWLEDGMENTS

We thank N.Jevtić, N.P. Santhanam, W. Szpankowski and J. Ziv for helpful discussions and advice.

REFERENCES

[1] R.E. Krichevsky and V.K. Trofimov. The performance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199–207, March 1981.

[2] T.M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991.

[3] T.M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, March 1996.

[4] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.

[5] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34(2):142–146, 1998.

[6] Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, March 2000.

[7] M. Drmota and W. Szpankowski. The precise minimax redundancy. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

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

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

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

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

[12] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, To appear, July 2006.

[13] G.M. Gemelos and T. Weissman. On the entropy rate of pattern processes. In *Proceedings of the Data Compression Conference*, March 2005.

[14] A.K. Dhulipala and A. Orlitsky. Experiments on English compression. In preparation.

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

[16] N. Jevtić, A. Orlitsky, and N.P. Santhanam. A lower bound on compression of unknown alphabets. Submitted for publication, February 2004.

[17] I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, December 1953.

[18] A. Nadas. On Turing's formula for word probabilities. *itassp*, ASSP-33(6):1414–1416, Dec 1985.

[19] A. Nadas. Good, Jelinek, Mercer, and Robins on Turing's estimate of probabilities. *ajmms*, 11:93–103, 1991.

[20] S.F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[21] I.J. Good. Turing's anticipation of Empirical Bayes in connection with the cryptanalysis of the Naval Enigma. *Journal of Statistics Computation and Simulation*, 66:101–111, 2000.

[22] D. McAllester and R. Schapire. On the convergence rate of Good Turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[23] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *Proceedings of the 44th Annual Symposium on Foundations of Computer Science*, October 2003.

[24] Y.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

[25] I. Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, October 1988.
[26] W.K. Hayman. A generalization of Stirling's formula. *Journal für die reine und angewandte Mathematik*, 196:67–95, 1956.
[27] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, 2001.
[28] I. Csiszár and P.C. Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information Theory*, 42(6):2065–2072, November 1996.

Anand K. Dhulipala Anand Dhulipala received B.Tech degree in electrical engineering from the Indian Institute of Technology, Madras in 2001 and M.S. degree in electrical and computer engineering in 2004 from the University of California, San Diego where he is currently a PhD candidate.

Alon Orlitsky Alon Orlitsky received B.Sc. degrees in Mathematics and Electrical Engineering from Ben Gurion University in 1980 and 1981, and M.Sc. and Ph.D. degrees in Electrical Engineering from Stanford University in 1982 and 1986.

From 1986 to 1996 he was with the Communications Analysis Research Department of Bell Laboratories. He spent the following year as a quantitative analyst at D.E. Shaw and Company, an investment firm in New York city. In 1997 he joined the University of California, San Diego, where he is currently a professor of Electrical and Computer Engineering and of Computer Science and Engineering.

Alon's research concerns information theory, learning, and speech recognition. He is a recipient of the 1981 ITT International Fellowship and of the 1992 IEEE W.R.G. Baker Paper Award