ISRAEL JOURNAL OF MATHEMATICS 111 (1999), 29–51

CODE LENGTH BETWEEN MARKOV PROCESSES

BY

A. IWANIK* AND J. SERAFIN*

Institute of Mathematics, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
e-mail: iwanik@im.pwr.wroc.pl and serafin@im.pwr.wroc.pl

ABSTRACT

Let X_1 and X_2 be two mixing Markov shifts over finite alphabet. If the entropy of X_1 is strictly larger than the entropy of X_2, then there exists a finitary homomorphism $\phi : X_1 \to X_2$ such that the code length is an L^p random variable for all $p < 4/3$. In particular, the expected length of the code ϕ is finite.

1. Introduction

In 1977 Keane and Smorodinsky [5] constructed a finitary homomorphism (or coding) from any Bernoulli process X to another Bernoulli process $\bar{X}$ of a strictly lower entropy. In the same paper they announced that the expected code length should be finite. This, however, is by no means evident and the proof was only given much more recently in [6].

The result of [5] was extended to Markov shifts by Akcoglu, del Junco and Rahe [1]. They construct a finitary coding between X and $\bar{X}$ under a sole assumption that X is ergodic Markov and $\bar{X}$ is mixing Markov of a lower entropy. Their construction is similar to [5], an essential role being played by a low entropy marker process. The presence of markers makes it possible to represent almost every source sequence $x \in X$ as a consistently ascending nested family of words, which fill longer and longer "skeletons" determined by the marker process. Fillers of sufficiently large rank are encoded to corresponding fillers in $\bar{X}$, thus eventually

* Research supported by KBN grant 2 P03A 039 15 1998–2001.
Received May 14, 1998

defining the required finitary coding $\phi\colon X \to \bar{X}$. Akcoglu, del Junco and Rahe, too, claim without proof that the code length should have a finite expectation.

There has been renewed interest in finitary coding in connection with non-commutative Bernoulli schemes. Actually, the fact that the Keane–Smorodinsky coding has finite expected length is essential in a recent work of Hamachi and Keane [4].

The aim of this paper is to present a proof that for mixing Markov shifts the code length has finite expectation. We prove even more by showing that it is an L^p function for some $p > 1$.

First we deal with a Markov-to-Bernoulli coding, in which case we simplify and correct the Bernoulli-to-Bernoulli argument in [6]. In particular, it follows that the length of the Keane–Smorodinsky code between two Bernoulli processes of unequal entropies has finite L^p-norm for all $p < 2$. It remains an open question whether the variance is finite. Our method relies on an application of Hölder inequality for appropriately chosen exponents that assure convergence of certain series. In contrast to [6] we consider individual fillers globally without classifying them according to length. This will be effective thanks to an application of a simplified version of the Bernstein inequality. The formulas for some distributions such as u_{-1}, b_r and l_0 will be derived in a detailed fashion in a preliminary section (we note that in [6] they are only asymptotically correct). A separate section is devoted to Bernoulli-to-Markov coding. Finally, as in [1], the Markov-to-Markov coding in obtained as a composition.

2. Definitions and notation

By a **process** with alphabet A we mean a dynamical system (X, μ, T), where $X = A^{\mathbf{Z}}$, T is the left shift, and μ is a shift-invariant Borel probability measure on X. All processes studied in this paper are with finite alphabet. Generally, the probability measure for any process will also be denoted by P.

Our notation will be based on [1] and [6]. For a Markov process given by a transition matrix (p_{ij}) we set $p_{\min} = \min\{p_{ij} : p_{ij} \neq 0\}$. The process will be assumed mixing, which means that the matrix is irredudicible and aperiodic. In particular, there exists a unique strictly positive vector (p_i) such that the powers of the transition matrix converge with exponential speed to the matrix with all its rows identical with (p_i). In the associated Markov process (X, μ, T) the p_i's represent the stationary marginal probabilities, $p_i = P(x_n = a_i)$ for every $n \in \mathbf{Z}$. The entropy of the Markov process X is given by the well-known formula $h(X) = -\sum_{i,j} p_i p_{ij} \log p_{ij}$. Recall that for Markov processes mixing

implies $h(X) > 0$ (except for the trivial case $|A| = 1$) and moreover the process has a **marker** M, i.e. a collection of blocks of the same length k such that

(i) each word in M begins with the same symbol a_1,

(ii) no word in M overlaps a word in M,

(iii) arbitrarily long concatenations of words from M occur with positive probabilities.

According to the construction in [1], the length k can be chosen arbitrarily large and the probability that a marker occurs at a given position decays exponentially with k. By ergodicity, almost every source sequence x in X splits into *runs* of markers labeled in a natural manner by $\pm 1, \pm 2, \ldots$ and *separating blocks* labeled $0, \pm 1, \pm 2, \ldots$ We assume that the 0-th coordinate of x is covered by either the run of markers labeled -1 or the subsequent 0-th separating block. By u_j we denote the number of markers in the j-th run while l_j stands for the length of the j-th separating block.

For every $r = 1, 2, \ldots$ we denote by $s_r = s_r(x)$ the **skeleton of rank** r. This is defined as the truncation of x to a finite segment around 0 such that the separating blocks in x are replaced by gaps of the same length, and with the property that the extreme left and right runs of markers each contain at least r markers while the internal runs, if any, each contain less than r markers. Moreover, neither the immediately preceding nor the immediately following k-block of x is a marker block. We denote by $-m_r < 0$ and $n_r > 0$ the label of the first and the last run of markers in s_r, respectively. Whenever convenient we will also consider a skeleton as an appropriate non-indexed finite sequence s consisting of runs of a k-block labeled M separated by gaps. For a skeleton s we denote by $l(s)$ the length of s minus the length of the last run of markers,

$$l(s_r) = ku_{-m_r} + l_{-m_r+1} + \cdots + ku_{-1} + l_0 + ku_1 + \cdots + l_{n_r-1}.$$

As in [1], the final block of markers is only needed to determine the occurrence of $s_r(x)$ but is not considered to be part of that occurrence. We define b_r to be the number of separating blocks in s_r so $b_r = m_r + n_r - 1$. A block in x occurring along a single run of markers followed by a separating block will be called an **order one filler**. The concatenation of all the order one fillers in s_r will be referred to as the **filler of** s_r. Clearly the length of the filler is equal to $l(s_r)$.

For a fixed non-indexed skeleton s the **filler measure** μ_s is defined on the $l(s)$-blocks as the projection of the conditional measure $\mu(\cdot|\mathcal{S})$ where $\mathcal{S}$ is the event

that s occurs at $[0, l(s)-1]$ in x. According to [1], Lemma 5.1, the filler measure μ_s is the product of the filler measures corresponding to order one subskeletons of s. Regardless of the skeleton rank, any filler measure will be denoted by μ_s.

Recall that the **marker process** is a stationary 0–1 process $\hat{X}$ defined by $\hat{x}_i = 0$ iff $x_i \cdots x_{i+k-1}$ is a marker block. If the marker length k is sufficiently large, then the entropy of the marker process can be made as small as needed. We define the **filler entropy** $f = h(X) - h(\hat{X})$.

Let ϕ: $X \to \bar{X}$ be a homomorphism, referred to as **coding**, between two processes (X, μ, T), $(\bar{X}, \bar{\mu}, \bar{T})$. This means that ϕ is measure preserving and $\phi(Tx) = \bar{T}\phi(x)$ a.e. The **code length** is defined as the least positive integer $C = C(x)$ such that there exists an integer interval J of length C containing 0 with the property that for a.e. y the condition $y_j = x_j$ for $j \in J$ implies that the encoded sequences $\phi(x), \phi(y)$ agree at the 0-th coordinate, $\phi(y)_0 = \phi(x)_0$. We let $C(x) = \infty$ if such a finite J does not exist. In this paper we will study finitary codings, i.e. such that $C(x)$ is finite with probability one. The code length $C(x)$ will be treated as a random variable.

We will only consider mixing Markov processes. As in [1], the coding between two such processes will be achieved in two steps.

In the first step, referred to as Markov-to-Bernoulli coding, we study a mixing Markov process (X, μ, T) and a Bernoulli process $(\bar{X}, \bar{\mu}, \bar{T})$ with $h(\bar{X}) = \bar{h} < h = h(X)$. A marker in X can be selected as a single word $a_1 \cdots a_k$ in such a way that the filler entropy f still exceeds the entropy $\bar{h}$. No marker will be needed in the Bernoulli process $\bar{X}$. We fix $\epsilon < (f - \bar{h})/3$. A filler F in the skeleton $s_r(x)$ of the source sequence $x \in X$, is called **bad** if $\mu_s(F) > e^{-l(s_r)(f-\epsilon)}$. On the other hand, a corresponding $l(s_r)$-block $\bar{F}$ in $\bar{x} \in \bar{X}$ will be called a bad filler if $\bar{\mu}(\bar{F}) < e^{-l(s_r)(\bar{h}+\epsilon)}$. According to [1], only good fillers will be encoded to good fillers. If a filler is bad, it will be encoded as a part of a longer good filler at a later stage. The coding is carried out for a given source sequence x by looking at the ascending skeletons $s_r(x)$, $r = 1, 2, \ldots$. By means of an "assignment" defined in [1], the filled skeleton $s_r(x)$ will be encoded in a consistent way if the filler F is good, except for a small set of exceptional cases. The conditional probability, given a marker structure of x, that $C(x) \geq c_{r+1}$ is bounded by (see [5], Lemma 14)

$$2^{b_r - cl(s_r)} + \mu_s(F \text{ is bad}) + \bar{\mu}(\bar{F} \text{ is bad}),$$

where $c = (f - \bar{h} - 2\epsilon)/\log 2 > 0$ and $F, \bar{F}$ denote the s_r-fillers in $X, \bar{X}$, respectively. Our aim is to prove that if the parameter k is chosen sufficiently large, then EC^p is finite for every $p < 2$. Clearly, EC^p is finite if

$\sum Ec_{r+1}{}^p P(C \geq c_{r+1}) < \infty$, so it suffices to show that the three following series converge:

$$S_1(p) = \sum_{r=1}^{\infty} E(c_{r+1}{}^p 2^{b_r - cl(s_r)}),$$

$$S_2(p) = \sum_{r=1}^{\infty} E(c_{r+1}{}^p \mu_s(F \text{ is bad})),$$

$$S_3(p) = \sum_{r=1}^{\infty} E(c_{r+1}{}^p \bar{\mu}(\bar{F} \text{ is bad})).$$

The second step is a Bernoulli-to-Markov coding. Now X is Bernoulli and $\bar{X}$ is mixing Markov with $h(\bar{X}) = \bar{h} < h = h(X)$. Moreover, as in [1], by extending $\bar{X}$ to another mixing Markov process with a slightly larger entropy (the extension taking place by a length-one coding) we may assume that there exist a marker $\bar{M}$ in $\bar{X}$ and a single-word marker M in X such that the corresponding marker processes have the same distribution. Therefore the two marker processes can be identified as a common factor $\hat{X}$ of X and $\bar{X}$; now the marker process in $\bar{X}$ will be referred to as "independent". Here the filler entropies are $f = h - h(\hat{X})$ and $\bar{f} = \bar{h} - h(\hat{X})$, respectively. The bad fillers in X and $\bar{X}$ are defined as in the Markov-to-Bernoulli case with $\bar{f}$ in place of $\bar{h}$ and $\bar{\mu}_s$ in place of $\bar{\mu}$. The finiteness of EC^p will be concluded similarly by studying the three series.

In a preliminary section, distributions of the parameters of the construction are calculated. Here the Markov property will be exploited on several occasions to assure a sufficient degree of independence of random variables under consideration. The distribution of b_r will play a decisive role in future calculations. Next, in the Markov-to-Bernoulli coding we will obtain a bound for the L^N-norm of c_{r+1} and will use it (for a large N) along with Hölder inequality to prove that each of the three sums is finite. For $S_2(p)$ and $S_3(p)$ we will need a form of the Bernstein inequality (or a large deviation theorem). The proof in the Bernoulli-to-Markov case is similar, but $S_3(p)$ is now more difficult to handle. In the last section we prove a lemma which enables us to compose codes without losing some features of the code length. In particular, this yields a coding of finite expected length between any two mixing Markov processes of unequal entropies.

3. Distributions

In this section we assume that (X, μ, T) is a mixing Markov process and calculate the distributions of the random variables u_1, u_{-1}, m_r, n_r, b_r, l_1, l_0. We also

assume that the marker M is a single word $a_1 \dots a_k$ and write

$$\eta = P(x_1 \cdots x_k = M \mid x_0 = a_k).$$

In other words, $\eta = p_{k1}p_{12} \cdots p_{k-1,k}$, where we write p_{ij} for $p_{a_i a_j}$. Finally, we set $\gamma = p_1/p_{k1}$ so $\gamma\eta$ is the probability that M occurs at a given position.

Recall that η decays exponentially with k. It will turn out that, as in [6], the probability $P(b_r = t)$ decays in r rapidly enough to ensure the convergence of the pertinent series.

By the Markov property and stationarity it follows that the u_j's are independent and, for $j \neq -1$, identically distributed. Clearly $P(u_1 \geq 1) = 1$ and

$$\begin{aligned} P(u_1 \geq t) =& P(x_1 \dots x_{k(t-1)} = M^{t-1} | x_{-k+1} \dots x_0 = M) \\ =& P(x_1 \dots x_{k(t-1)} = M^{t-1} | x_0 = a_k) = \eta^{t-1} \end{aligned}$$

for $t > 1$.

The distribution of u_{-1} is quite different. First we define auxiliary events: B_0 will mean "x_0 is contained in a marker", $B_{0,i}$ will denote "x_0 is contained in the i-th marker of the run of markers containing x_0", and B_0^c is the negation of B_0. For any $t \geq 1$ we have by the Markov property

$$P(u_{-1} \geq t | B_0^c) = \eta^{t-1} \quad \text{and} \quad P(u_{-1} \geq t | B_{0,i}) = \eta^{t-i}.$$

Therefore

$$\begin{aligned} P(u_{-1} \geq t) =& P(u_{-1} \geq t | B_0^c) P(B_0^c) + \sum_{i=1}^{\infty} P(u_{-1} \geq t | B_{0,i}) P(B_{0,i}) \\ =& \eta^{t-1} P(B_0^c) + \sum_{i=1}^{t} \eta^{t-i} P(B_{0,i}) + \sum_{i=t+1}^{\infty} P(B_{0,i}). \end{aligned}$$

Since the probability that M occurs at a given position equals $\gamma\eta$ and the reverse time process is Markov, we get for every $j = 1, \dots, k$

$$P(B_{0,i},\ x_0 \text{ is the } j\text{-th element of } M) = \gamma\eta\eta^{i-1}(1-\eta),$$

so $P(B_{0,i}) = k\gamma\eta^i(1-\eta)$. It is clear that $P(B_0) = k\gamma\eta$. Consequently

$$\begin{aligned} P(u_{-1} \geq t) &= \eta^{t-1}(1 - k\gamma\eta) + \sum_{i=1}^{t} k\gamma(1-\eta)\eta^t + k\gamma\eta^{t+1} \\ &= \eta^{t-1}\left(1 + k\gamma\eta(t-1)(1-\eta)\right). \end{aligned}$$

Our next aim is to calculate the distribution of b_r. We observe that

$$P(m_r = 1) = P(u_{-1} \geq r) = \eta^{r-1}\left(1 + k\gamma\eta(r-1)(1-\eta)\right),$$

while for $i > 1$

$$\begin{aligned} P(m_r = i) =& P(u_{-1} < r, u_{-2} < r, \ldots, u_{-i+1} < r, u_{-i} \geq r) \\ =& \left(1 - \eta^{r-1} - k\gamma\eta^r(r-1)(1-\eta)\right)(1-\eta^{r-1})^{i-2}\eta^{r-1} \end{aligned}$$

and clearly

$$P(n_r = i) = (1-\eta^{r-1})^{i-1}\eta^{r-1}$$

for all i. The Markov property implies that m_r and n_r are independent, so

$$\begin{aligned} P(b_r = t) =& P(m_r + n_r = t+1) = \sum_{i=1}^{t} P(m_r = i)P(n_r = t-i+1) \\ =& \left(1 + k\gamma(r-1)\eta(1-\eta)\right)\eta^{r-1}\,(1-\eta^{r-1})^{t-1}\eta^{r-1} \\ &+ \sum_{i=2}^{t}\left(1 - \eta^{r-1} - k\gamma(r-1)\eta^r(1-\eta)\right) \\ &\quad \times (1-\eta^{r-1})^{i-2}\eta^{r-1}\,(1-\eta^{r-1})^{t-i}\eta^{r-1} \\ =& (1-\eta^{r-1})^{t-1}\eta^{2r-2} \\ &\times \left(1 + k\gamma(r-1)\eta(1-\eta) + (t-1)(1 - \frac{k\gamma(r-1)\eta^r(1-\eta)}{1-\eta^{r-1}})\right). \end{aligned}$$

Therefore, if k is chosen sufficiently large so that $k\eta < 1/\gamma$, we get

$$P(b_r = t) \leq (1-\eta^{r-1})^{t-1}\eta^{2r-2}(1 + r - 1 + t - 1) \leq rt(1-\eta^{r-1})^{t-1}\eta^{2r-2}.$$

Now we proceed to the calculation of the distributions of the l_j's. Like for u_j's we note that the l_j's are independent and, for $j \neq 0$, identically distributed.

We denote by l' the separation between two consecutive markers. More formally $P(l' = t-1)$ is the conditional probability that t is the least positive integer such that $x_t \cdots x_{t+k-1} = M$ given that $x_{-k+1} \cdots x_0 = M$. Clearly the renewal sequence for $l' + k$ is

$$q_0 = 1, \quad q_1 = \cdots = q_{k-1} = 0, \quad q_k = \eta, \quad q_{k+j} = \frac{p_{k1}^{(j+1)}}{p_{k1}}\eta,$$

for $j \geq 0$, where $p_{k1}^{(j)}$ denotes the j-step transition probability from a_k to a_1. The renewal function for $l' + k$ is

$$Q(s) = 1 + \frac{\eta}{p_{k1}}\sum_{j=0}^{\infty} p_{k1}^{(j+1)} s^{k+j}$$

and hence the generating funcion is given by the formula

$$1 - \frac{1}{Q(s)} = \frac{\frac{\eta s^k}{p_{k1}} \sum_{j=0}^{\infty} p_{k1}^{(j+1)} s^j}{1 + \frac{\eta s^k}{p_{k1}} \sum_{j=0}^{\infty} p_{k1}^{(j+1)} s^j}.$$

On dividing by s^k we obtain the generating function of l':

$$G_{l'}(s) = \sum_{j=0}^{\infty} P(l' = j) s^j = \eta \frac{\sum_{j=0}^{\infty} \pi^{(j)} s^j}{1 + \eta s^k \sum_{j=0}^{\infty} \pi^{(j)} s^j},$$

where $\pi^{(j)} = p_{k1}^{(j+1)} / p_{k1}$.

As it is clear that for $t > 0$

$$P(l_1 = t) = \frac{P(l' = t)}{P(l' > 0)},$$

we have

$$\begin{aligned} G_{l_1}(s) &= \frac{G_{l'}(s) - G_{l'}(0)}{1 - G_{l'}(0)} \\ &= \frac{\eta}{1 - \eta} \frac{\sum_{j=0}^{\infty} \pi^{(j)} s^j}{1 + \eta s^k \sum_{j=0}^{\infty} \pi^{(j)} s^j} - \frac{\eta}{1 - \eta} \\ &= \frac{\eta}{1 - \eta} \frac{(1 - \eta s^k) \sum_{j=0}^{\infty} \pi^{(j)} s^j - 1}{1 + \eta s^k \sum_{j=0}^{\infty} \pi^{(j)} s^j}. \end{aligned}$$

For a mixing Markov chain we have $\pi^{(j)} \to \gamma$ exponentially so $\pi^{(j)} = \gamma + \alpha_j$, where $\alpha_j \to 0$ exponentially. Consequently, the function

$$h(s) = \sum_{j=0}^{\infty} \alpha_j s^j$$

is analytic on a disk of radius strictly greater than 1. We can write

$$G_{l_1}(s) = \frac{\eta}{1 - \eta} \frac{\gamma(1 - \eta s^k) + (1 - \eta s^k)(1 - s)h(s) - 1 + s}{1 - s + \gamma \eta s^k + \eta(1 - s) s^k h(s)}.$$

The value of the last denominator at $s = 1$ is equal to $\gamma\eta$ while, if k is sufficiently large, its derivative is close to -1. This implies that the denominator does not vanish in a certain neighborhood of $s = 1$. Therefore the power series

$$G_{l_1}(s) = \sum_{j=1}^{\infty} P(l_1 = j) s^j$$

has an analytic continuation through the point $s = 1$. Since the coefficients are nonnegative, we conclude that the power series converges in a disk $|s| < \rho$, where

$\rho > 1$. Equivalently, the probabilities $P(l_1 = j)$ decay exponentially and l_1 has exponential moments. In particular,

$$E({l_1}^N) = C_1(k, N) < \infty$$

for every $N \geq 1$. Actually, by estimating the derivatives of the generating function it is not hard to see that $E({l_1}^N) \leq C_2(N)k^N\eta^{-N}$, but we will not need this formula.

Finally we will examine the distribution of l_0. Denote by τ the least non-negative integer t such that a marker starts at t, i.e. $x_t \cdots x_{t+k-1} = M$. It is clear that the conditional distribution of τ given that x_0 is in a marker coincides with that of l_1.

Now, for every i in the alphabet let q_i be the minimal positive integer q such that $p^{(q)}_{a_k,i} > 0$. We have

$$\begin{aligned}
P(\tau \geq t,\ B_0^c) &= \sum_i P(\tau \geq t,\ B_0^c \mid x_0 = i)P(x_0 = i) \\
&= \sum_i P(\tau \geq t,\ B_0^c \mid x_0 = i,\ x_{-q_i-k+1} \cdots x_{-q_i} = M)P(x_0 = i) \\
&= \sum_i P(\tau \geq t,\ B_0^c,\ x_0 = i,\ x_{-q_i-k+1} \cdots x_{-q_i} = M) \\
&\quad \times P(x_0 = i)/P(x_0 = i,\ x_{-q_i-k+1} \cdots x_{-q_i} = M) \\
&\leq \max_i P(\tau \geq t,\ B_0^c,\ x_0 = i,\ x_{-q_i-k+1} \cdots x_{-q_i} = M) \\
&\quad \times \sum_i (\gamma\eta p^{(q_i)}_{a_k,i})^{-1} \\
&\leq C_3\eta^{-1} \max_i P(\tau \geq t,\ B_0^c,\ x_0 = i,\ x_{-q_i-k+1} \cdots x_{-q_i} = M) \\
&= C_3\gamma\eta^{-1} \max_i P(\tau \geq t,\ B_0^c,\ x_0 = i \mid x_{-q_i-k+1} \cdots x_{-q_i} = M)\eta.
\end{aligned}$$

But by the minimality of q_i we get

$$\begin{aligned}
&P(\tau \geq t,\ B_0^c,\ x_0 = i \mid x_{-q_i-k+1} \cdots x_{-q_i} = M) \\
&\leq P(l' \geq q_i + t) \leq P(l' > 0)P(l_1 \geq t),
\end{aligned}$$

so

$$P(\tau \geq t,\ B_0^c) \leq C_4 P(l_1 \geq t).$$

The latter decays exponentially and we have

$$E(1_{B_0^c}{\tau}^N) \leq C_4 E({l_1}^N) \leq C_5(k, N) < \infty.$$

In a symmetric fashion we define τ_- for the reverse time process. It is clear that if x_0 is not in a marker, then $l_0 = \tau + \tau_- - 1$. On the other hand, if x_0 is in a marker, then the (conditional) distribution of l_0 is the same as that of l_1. It follows that l_0 has exponential moments and

$$E({l_0}^N) \leq C_6(k, N) < \infty.$$

4. Norm of skeleton length

For a positive integer $N \geq 2$ we are going to estimate the L^N-norm of c_{r+1}. Clearly

$$c_{r+1} = k + l(s_{r+1}) + k(r+1),$$

and if $b_{r+1} = t$ then

$$l(s_{r+1}) = ku_{-m_{r+1}} + l_0 + \xi_1 + \xi_2,$$

where ξ_1 is a sum of $t-1$ independent random variables l_j $(j \neq 0)$, and $\xi_2 \leq k(r+1)(t-1)$ corresponds to the joint length of markers inside the skeleton.

For ξ_1 we write

$$E\xi_1^N = \sum_{t=1}^{\infty} E(\xi_1^N | b_{r+1} = t) P(b_{r+1} = t).$$

Since

$$(v_1 + \cdots + v_{t-1})^N \leq (t-1)^{N-1}(v_1^N + \cdots + v_{t-1}^N)$$

for any $v_j \geq 0$, we get, using the upper bound for $P(b_{r+1} = t)$, that

$$\begin{aligned} E\xi_1^N \leq & \sum_{t=1}^{\infty} (t-1)^N E{l_1}^N (r+1) t (1-\eta^r)^{t-1} \eta^{2r} \\ \leq & C_1(k, N)(r+1)\eta^{2r} \sum_{t=1}^{\infty} t^{N+1}(1-\eta^r)^{t-1}. \end{aligned}$$

On the other hand, by approximating the integral $\Gamma(N+2) = \int_0^\infty x^{N+1} e^{-x}\, dx$ by its Riemann sums it is not hard to see that

$$\frac{\sum_{t=1}^{\infty} t^{N+1}(1-\delta)^t}{\Gamma(N+2)\delta^{-(N+2)}} \to 1$$

as $0 < \delta \to 1$. Consequently, for $\delta = \eta^r$,

$$E\xi_1^N \leq C_1(k, N)(r+1)\eta^{-Nr}.$$

As to u_{-m_r}, we have $P(u_{-m_r} \geq r + j) = \eta^j$ for $j \geq 0$. Therefore for $\eta < 1/2$ we can write

$$\begin{aligned} E(u_{-m_r}^N) &= \sum_{t=r}^{\infty} t^N \eta^{t-r}(1-\eta) \leq \sum_{j=0}^{\infty} (r+j)^N 2^{-j} \\ &= r^N \sum_{j=0}^{\infty} \left(1 + \frac{j}{r}\right)^N 2^{-j} \leq r^N \sum_{j=0}^{\infty} (1+j)^N 2^{-j} \\ &= C_7(N) r^N. \end{aligned}$$

We also estimate

$$\begin{aligned} E\xi_2^N &\leq k^N (r+1)^N E(b_{r+1} - 1)^N \\ &\leq k^N (r+1)^N \sum_{t=1}^{\infty} (t-1)^N (r+1) t (1-\eta^r)^{t-1} \eta^{2r} \\ &\leq C_8(k, N)(r+1)^{N+1} \eta^{-rN}, \end{aligned}$$

as for ξ_1.

Altogether the three upper bounds yield

$$(E c_{r+1}{}^N)^{1/N} \leq C_9(k, N) r^2 \eta^{-r}.$$

5. Convergence of the first series

Given $1 < p < 2$ we let $p' = p'(N) = N/p$ and q' be such that

$$\frac{1}{p'} + \frac{1}{q'} = 1.$$

Note that $1 < q' \leq 2$ if $N \geq 4$.

We will now show that $S_1(p)$ is finite.

By the Hölder inequality we obtain

$$E(c_{r+1}{}^p 2^{b_r - cl(s_r)}) \leq (E c_{r+1}{}^{pp'})^{1/p'} (E 2^{q'(b_r - cl(s_r))})^{1/q'}.$$

But, given that $b_r = t$, the random variable $l(s_r)$ is bounded below by the sum of $t-1$ independent copies of l_1. Therefore

$$\begin{aligned} E 2^{q'(b_r - cl(s_r))} &\leq \sum_{t=1}^{\infty} 2^{q't} (G_{l_1}(2^{-c}))^{t-1} P(b_r = t) \\ &\leq \sum_{t=1}^{\infty} ((G_{l_1}(2^{-c}))^{-1} (4 G_{l_1}(2^{-c}))^t r t \eta^{2r-2} \\ &= C_{10}(k) r \eta^{2r}, \end{aligned}$$

where

$$C_{10}(k) = \eta^{-2} \sum_{t=1}^{\infty} ((G_{l_1}(2^{-c}))^{-1} t (4G_{l_1}(2^{-c}))^t$$

is finite if k is sufficiently large (so that η is small enough for $4G_{l_1}(2^{-c}) < 1$). Therefore

$$\begin{aligned} E(c_{r+1}{}^p 2^{b_r - cls_r}) &\leq (Ec_{r+1}{}^N)^{p/N} C_{10}(k)^{1/q'} r\eta^{2r(N-p)/N} \\ &\leq \left(C_9(k,N) r^2 \eta^{-r}\right)^p C_{10}(k) r \eta^{2r(N-p)/N} \\ &\leq C_{11}(k,N,p) r^5 \eta^{(2-p-2p/N)r}, \end{aligned}$$

so $\sum_r E(c_{r+1}{}^p 2^{b_r - cl(s_r)}) < \infty$ because we can choose $N > 2p/(2-p)$.

6. Bad fillers in the Markov process

In this section we show that $S_2(p)$ is finite. We will apply the Bernstein inequality to the random variables $-\log \mu_s(F)$, where F runs over fillers of the order one subskeletons of the skeleton s_r. As a result we get a bound for the probability that the empirical mean value of $-\log \mu_s(F)$ differs essentially from $gE(ku_1 + l_1)$. This will ensure an exponential bound for the probability that the filler is bad given $b_r = t$.

The following useful lemma is an immediate consequence of the Bernstein inequality, which actually gives an exponential bound for the probability ([2], Ch. 2). It can also be deduced from Cramér's large deviation theorem (see e.g. [3], XVI, §6, Theorem 1). For the reader's convenience we give a direct elementary proof of the lemma.

LEMMA 1: *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables such that $Ee^{\alpha|\xi_1|} < \infty$ for some $\alpha > 0$ and $E\xi_1 = 0$. Then for every $\delta > 0$ there exists t_0 such that*

$$P(|\xi_1 + \cdots + \xi_t|/t > \delta) < 1/t^4$$

for all $t \geq t_0$.

Proof: Instead of $Ee^{\alpha|\xi_1|} < \infty$ it suffices to assume $E\xi^N < \infty$ for an even integer $N \geq 10$. Let $p(i, N)$ be the number of ways N can be represented as a sum of i integers ≥ 2, including the order of summation; we have, e.g.,

$$p(1, N) = 1, \quad p(2, N) = (N-2)/2, \quad p(N/2, N) = 1.$$

We also define $p(N) = \max(p(1, N), p(2, N), \ldots, p(N/2, N))$ and let

$$m(N) = \max\{|E\xi_1{}^j| : j = 0, 1, \ldots, N\},$$

where we set $E\xi_j^0 = 1$. Since $E\xi_j = 0$, we clearly have

$$E\left((\xi_1 + \cdots + \xi_t)^N\right) = \sum \binom{N}{k_1, \ldots, k_t} E\xi_1^{k_1} \cdots E\xi_t^{k_t},$$

where the summation is over all nonnegative integer vectors $(k_1, \ldots, k_t)$ such that $k_1 + \cdots + k_t = N$ and $k_j \geq 2$ whenever $k_j \neq 0$. If $t > N$ then the number of such vectors with exactly i nonzero coordinates is

$$\binom{t}{i} p(i, N) \leq \binom{t}{N/2} p(N),$$

so the sum is bounded by

$$N!(N/2 - 1)\binom{t}{N/2} p(N) m(N) t^{N/2} \leq C'(N) t^{N/2}.$$

Now the Chebychev inequality applied to ξ^N yields

$$P(|\xi_1 + \cdots + \xi_t| > \delta t) \leq \frac{C'(N) t^{N/2}}{\delta^N t^N} \leq C'(N, \delta) t^{-N/2},$$

which ends the proof of the lemma. ∎

Given that $b_r = t$, the filler F of s_r is a concatenation of t fillers of order one subskeletons of s_r. By the Markov property these order one fillers are independent and equally distributed except for the initial filler of length $ku_{-m_r} + l_{-m_r+1}$ and the central filler of length $ku_{-1} + l_0$. We denote them by F_{-1} and F_0, respectively. The other order one fillers in F will be denoted $F_1, F_2, \ldots$. Clearly it may happen that $-m_r = -1$, so $F_{-1} = F_0$, in which case there are $t - 1$ other fillers; otherwise there are only $t - 2$. We will write $\zeta_j = -\log \mu_s(F_j)$. The conditional measure $\mu_s(F_j)$ depends only on the manner in which the block of length l_j between two consecutive runs of markers is filled in. It follows that the variables ζ_j are independent and, for $j \neq 0$, identically distributed. Notice that $\mu_s(F_j) \geq p_{\min}^{l_j}$, where $p_{\min}$ is the least positive entry of the transition matrix. Therefore there exists $\alpha > 0$ such that, for $j \neq 0$,

$$E(e^{\alpha\zeta_j}) = \sum_{l=1}^{\infty} E(e^{-\alpha \log \mu_s(F_j)} \mid l_j = l) P(l_j = l) \leq G_{l_1}(1/p_{\min}^{\alpha}) < \infty,$$

since $G_{l_1}(s)$ has no poles in a disk of radius > 1 around the origin. It follows that the random variables $\zeta_j - E\zeta_j$, $j \neq 0$ satisfy the assumption of Lemma 1. By the same token

$$E(e^{\alpha\zeta_0}) < \infty$$

for some $\alpha > 0$.

By Lemma 3.1 in [1] we have

$$-\lim_r \log \mu_s(F)/l(s_r) = f$$

with probability 1. On the other hand, by the law of large numbers,

$$l(s_r)/b_r \to E(ku_1 + l_1) = \lambda \quad \text{a. s.}$$

It should be noted that the exceptional quantities u_{-m_r}, u_{-1}, and l_0 do not affect the value of the limit. Indeed, it is easy to see that $l_0/b_r \to 0$ while for u_{-1} and u_{-m_r} we observe that $u_{-1} \le u_{-m_r} = k(r-1) + u$, where u is independent of b_r and distributed as u_1. Therefore it remains to show $r/b_r \to 0$. This, however, is a simple consequence of the Borel–Cantelli lemma, for given $\alpha > 0$ we can write

$$\begin{aligned}\sum_r P(r/b_r > \alpha) &= \sum_r P(b_r < r/\alpha) \\ &\le \sum_r r\eta^{2r-2} \sum_{t=1}^{[r/\alpha]} t \le \sum_r r^2\eta^{2r-2}/\alpha^2 < \infty.\end{aligned}$$

Consequently, $-\log \mu_s(F)/b_r \to f\lambda$ and, thanks to Lemma 5.1 in [1], $E\zeta_j = f\lambda$ for $j \neq 0$.

Now we define an auxiliary event

$$B_r = \{|l(s_r)/b_r - \lambda| > \delta\},$$

where $\delta < \min(\epsilon/4f, \lambda/24, \epsilon\lambda/16)$. From now on, for the sake of simplicity, we only consider the case where $-m_r \neq -1$, the other case being handled similarly.

Since clearly $l(s_r) > kb_r$,

$$\begin{aligned}P(B_r^c,\ F \text{ is bad} \mid b_r = t) &= P(B_r^c,\ \mu_s(F) > e^{(\epsilon-f)l(s_r)} \mid b_r = t) \\ &= P(B_r^c,\ \zeta_{-1} + \zeta_0 + \cdots + \zeta_{t-2} < (f-\epsilon)l(s_r) \mid b_r = t) \\ &\le P\left(B_r^c,\ f\left|\lambda - \frac{l(s_r)}{t}\right| > \frac{\epsilon k}{2} \mid b_r = t\right) \\ &\quad + P\left(B_r^c,\ \left|\frac{\zeta_{-1} + \zeta_0 + \cdots + \zeta_{t-2}}{t} - f\lambda\right| > \frac{\epsilon l(s_r)}{2t} \mid b_r = t\right).\end{aligned}$$

The first term vanishes by the choice of δ. The second term is bounded by

$$\begin{aligned}&P\left(\frac{\zeta_0}{t} > \frac{\epsilon l(s_r)}{4t} \mid b_r = t\right) \\ &+P\left(B_r^c,\ \left|\frac{\zeta_{-1} + \zeta_1 + \cdots + \zeta_{t-2}}{t} - f\lambda\right| > \frac{\epsilon l(s_r)}{4t} \mid b_r = t\right)\end{aligned}$$

$$\leq P\left(\zeta_0 > \frac{\epsilon k t}{4}\right) + P\left(\left|\frac{\zeta_{-1} + \zeta_1 + \cdots + \zeta_{t-2}}{t-1} - f\lambda\right| > \frac{\epsilon\lambda}{8} - \frac{f\lambda}{t-1}\right),$$

which is less than $e^{-\beta_1 t} + 1/(t-1)^4$ for some positive constant β_1 and all sufficiently large t (independently of r). The last assertion is a consequence of the exponential decay of the distribution of ζ_0, and of Lemma 1 applied to $\zeta_j - E\zeta_j$.

Therefore we obtain

$$\begin{aligned} P(B_r^c,\ F \text{ is bad}) &= \sum_{t=1}^{\infty} P(B_r^c,\ F \text{ is bad} \mid b_r = t) P(b_r = t) \\ &\leq \sum_{t=1}^{t_0-1} tr\eta^{2r-2} + \sum_{t=t_0}^{\infty} (e^{-\beta_1 t} + 1/(t-1)^4) tr\eta^{2r-2} \\ &\leq C_{12}(k) r\eta^{2r}. \end{aligned}$$

Now we are going to obtain an exponential bound for the probability of B_r. Clearly the skeleton length $l(s_k)$ decomposes into the following five terms: ku_{-m_r}, ku_{-1}, l_0, $k(u_1' + \cdots + u_{t-2}')$, and $l_1' + \cdots + l_{t-1}'$, where the u_j' and l_j' are the random variables u_j, $j \neq -m_r, -1$, and $l_j, j \neq 0$ occurring within the skeleton s_r, relabeled accordingly. Therefore

$$\begin{aligned} P(B_r \mid b_r = t) \leq & P\left(u_{-m_r} > \frac{\delta t}{5k} \mid b_r = t\right) \\ & + P\left(u_{-1} > \frac{\delta t}{5k} \mid b_r = t\right) \\ & + P\left(l_0 > \frac{\delta t}{5} \mid b_r = t\right) \\ & + P\left(\left|\frac{u_1' + \cdots + u_{t-2}'}{t} - Eu_1\right| > \frac{\delta}{5k} \mid b_r = t\right) \\ & + P\left(\left|\frac{l_1' + \cdots + l_{t-1}'}{t} - El_1\right| > \frac{\delta}{5} \mid b_r = t\right). \end{aligned}$$

We first note that since the distribution of l_0 decays exponentially, the third summand is bounded by $e^{-\beta_2 t}$ for some $\beta_2 > 0$. Since $u_{-1} \leq u_{-m_r}$, the second summand is bounded by the first.

For the first term we write $u_{-m_r} = k(r-1) + u$ as above and note that for some $\beta_3 > 0$

$$\begin{aligned} P(u_{-m_r} > \delta t/5k \mid b_r = t) &\leq P(u_1 > \delta t/10k) + P((r-1) > \delta t/10k \mid b_r = t) \\ &< e^{-\beta_3 t} + \sigma_1(r,t), \end{aligned}$$

where $\sigma_1(r,t) = 1$ or 0 according as t is or is not less than $10k(r-1)/\delta$.

For the fourth summand we will apply Lemma 1. To avoid the error term caused by the condition of being in the skeleton s_r we introduce new random variables. Let $u_1'', u_2'', \ldots$ be independent and distributed as u_1. Now, for each j, whenever $u_j'' \geq r$ occurs, we will repeat independent trials until a value less than r appears. The first appearance of a "short" result will be called u_j'. The apparent abuse of notation will not affect the calculation because the new random vector $(u_1', \ldots, u_{t-2}')$ will have the same distribution as its counterpart that appears in the fourth term. We will write $w_j = u_j'' - u_j'$. It is clear that these are nonnegative i.i.d. variables with $w_j < u_j''$ and $w_j = 0$ iff $u_j'' < r$. Observe that $w_j \neq 0$ happens with probability $P(u_j'' \geq r) = P(u_1 \geq r) = \eta^{r-1}$, so

$$Ew_j \leq \sum_{i=r}^{\infty} i\eta^{i-1} \leq 4r\eta^{r-1}.$$

The fourth summand is thus majorized by

$$\begin{aligned} &P\left(\left|\frac{u_1'' + \cdots + u_{t-2}''}{t} - Eu_1\right| > \frac{\delta}{10k}\right) + P\left(\frac{w_1 + \cdots + w_{t-2}}{t} > \frac{\delta}{10k}\right) \\ \leq &P\left(\left|\frac{u_1'' + \cdots + u_{t-2}''}{t-2} - Eu_1\right| > \frac{\delta}{10k} - \frac{2Eu_1}{t-2}\right) \\ &+ P\left(\frac{w_1 + \cdots + w_{t-2}}{t-2} > \frac{\delta}{10k} - \frac{2}{t-2}\right). \end{aligned}$$

Since for t sufficiently large the right hand sides of both inequalities are greater than $\delta/20k$, we can use Lemma 1 for the first probability. The second is bounded by

$$P\left(\left|\frac{w_1 + \cdots + w_{t-2}}{t-2} - Ew_1\right| > \frac{\delta}{20k} - Ew_1\right).$$

Since $Ew_1 \leq 4r\eta^{r-1}$, we have $\delta/20k - Ew_1 > \delta/40k$ for $r \geq r_0$, say. This allows us to use Lemma 1 for the last probability, too. Consequently, the fourth term is bounded by $2/(t-2)^4$ for all $r \geq r_0$, and all sufficiently large t (independently of r).

Now it should be clear how to majorize the fifth summand. We can just apply Lemma 1 to the independent random variables l_j' and get a similar bound of the form $1/(t-1)^4$. As a result we obtain that, for $r \geq r_0$ and all sufficiently large t, $t \geq t_0$, say,

$$P(B_r \mid b_r = t) \leq e^{-\beta_3 t} + \sigma_1(r,t) + 2e^{-\beta_2 t} + \frac{2}{(t-2)^4} + \frac{1}{(t-1)^4}.$$

Consequently,

$$\begin{aligned} P(B_r) \leq & \sum_{t=1}^{t_0} tr\eta^{2r-2} + \sum_{t=t_0}^{[10k(r-1)/\delta]} tr\eta^{2r-2} \\ & + \sum_{t=t_0}^{\infty} \left(e^{-\beta_3 t} + 2e^{-\beta_2 t} + \frac{2}{(t-2)^4} + \frac{1}{(t-1)^4} \right) tr\eta^{2r-2} \\ \leq & C_{13}(k) r^3 \eta^{2r}. \end{aligned}$$

In order to prove that $S_2(p)$ converges we now argue as in the proof for $S_1(p)$. As above, we denote by F the filler of s_r. For N sufficiently large we have

$$E\left(c_{r+1}{}^p \mu_s(F \text{ is bad})\right) \leq (Ec_{r+1}{}^N)^{p/N} \left(E(\mu_s(F \text{ is bad})^{N/(N-p)}) \right)^{(N-p)/N}$$

Since $N/(N-p) > 1$, we may write

$$\begin{aligned} E\left(\mu_s(F \text{ is bad})^{N/(N-p)}\right) \leq & E(\mu_s(F \text{ is bad})) = P(F \text{ is bad}) \\ \leq & C_{12}(k) r\eta^{2r} + C_{13}(k) r^3 \eta^{2r} \leq C_{14}(k) r^3 \eta^{2r}. \end{aligned}$$

Consequently,

$$\begin{aligned} S_2(p) = & \sum_r E(c_{r+1}{}^p \mu_s(F \text{ is bad})) \\ \leq & \sum_r C_9(k,N)^p r^{2p} \eta^{-rp} (C_{14}(k) r^3 \eta^{2r})^{(N-p)/N} \\ \leq & C_{15}(k,N,p) \sum_r r^7 \eta^{(2-p-2p/N)r} \end{aligned}$$

and the series $S_2(p)$ converges if $N > 2p/(2-p)$.

7. Bad fillers in the Bernoulli process

By the nature of the coding there is no marker structure in the Bernoulli process $\bar{X}$. Therefore in order to calculate the probability that a filler $\bar{F}$ in $\bar{X}$ corresponding to the skeleton s_r on top is bad, we will condition on $l(s_r)$. The following crude estimate will be sufficient for our purpose:

$$P(l(s_r) = t) \leq P(l(s_r) \leq t) \leq P(b_r \leq t) \leq \sum_{j=1}^{t} jr\eta^{2r-2} \leq t^2 r\eta^{2r-2}.$$

Now we can write

$$P(\bar{F} \text{ is bad}) \leq \sum_t P(\bar{F} \text{ is bad} \mid l(s_r) = t) t^2 r\eta^{2r-2}.$$

For $l(s_r) = t$ we clearly have

$$-\log \bar{\mu}(\bar{F}) = \sum_{j=1}^{t} \bar{\zeta}_j,$$

where the random variables $\bar{\zeta}_j$, corresponding to single symbols in $\bar{F}$, are independent and distributed as $-\log \bar{\mu}(\bar{x}_0)$. It is clear that $E(\bar{\zeta}_1) = \bar{h}$. As the condition $\bar{F}$ is bad is equivalent to the inequality $-\log \bar{\mu}(\bar{F}) > l(s_r)(\bar{h} + \epsilon)$, we can apply Lemma 1 to the variables $\bar{\zeta}_j - \bar{h}$ and obtain as before, for all sufficiently large t,

$$P(\bar{F} \text{ is bad} \mid l(s_r) = t) \leq C_{16}(k)/t^4,$$

which readily implies $S_3(p) < \infty$.

We have proved the following theorem.

THEOREM 1: *Let the processes $X, \bar{X}$ be mixing Markov and Bernoulli, respectively. If $h(X) > h(\bar{X})$, then there exists a finitary coding from X to $\bar{X}$ such that for every $p < 2$ the code length is an L^p random variable.*

It should be observed that, if the marker length k is large enough, our proof applies to the code constructed in [1] and, in the Bernoulli-to-Bernoulli case, to the Keane–Smorodinsky code [5].

8. Bernoulli-to-Markov coding

In this section X will stand for a Bernoulli process with entropy $h(X) = h$ and $\bar{X}$ will denote a mixing Markov process with $h(\bar{X}) = \bar{h} < h$. According to [1] we may assume that there is an "independent" marker process, common to X and $\bar{X}$. The marker in $\bar{X}$ will be denoted by $\bar{M}$. Unlike in the Markov-to-Bernoulli case, where M was a single word, this marker is a collection of words of length k all starting from the same symbol $\bar{a}_1$, say. The marker in X is still a single word. The measure of $\bar{M}$ decays exponentially with k, because markers in X and $\bar{X}$ have the same measure. The filler entropies in X and $\bar{X}$ will be denoted by f and $\bar{f}$, respectively. By choosing k sufficiently large we may assure $f - \bar{f}$ positive and in fact arbitrarily close to $h - \bar{h}$. Other parameters of the coding, such as $l(s_r), b_r, l_0, m_r, u_{-1}$, etc., are determined by the marker process so they are functions of the Bernoulli process X.

The outline of the argument is the same as before. The convergence of $S_1(p)$ and $S_2(p)$ follows from the previous part as a special case. The proof for $S_3(p)$ is somewhat different, because the distribution of the filler measure now depends on

the parameter r which influences the filler length (the block filling in the ku_j part of the corresponding order one filler in $\bar{X}$ is now random, unlike in the previous case where M was a single word).

If $\bar{F}$ is a filler of the skeleton s_r in $\bar{X}$, then $\bar{F}$ has length $l(s_r)$ and is a concatenation of b_r blocks (order one fillers) each corresponding to a run of markers followed by a gap of length l_j. Since each marker starts from the same definite symbol $\bar{a}_1$, by the Markov property the order one fillers $\bar{F}_j$ are independent. Moreover, except for the two exceptional fillers $\bar{F}_{-m_r+1}$ and $\bar{F}_0$, they are identically distributed (both unconditionally and conditionally given that $b_r = t$). As in the Markov-to-Bernoulli case, it will be useful to relabel the fillers by giving the exceptional fillers the indices $-1, 0$, thus $\bar{F}_{-m_r+1} = \bar{F}'_{-1}$, $\bar{F}_0 = \bar{F}'_0$, and for $b_r = t$ we have

$$\bar{F} = \bar{F}_{-m_r+1} \cdots \bar{F}_{-1}\bar{F}_0 \cdots \bar{F}_{n_r-1} = \bar{F}'_{-1}\bar{F}'_1 \cdots \bar{F}'_{m_r-2}\bar{F}'_0\bar{F}'_{m_r-1} \cdots \bar{F}'_{t-2}.$$

The corresponding quantities l_j and u_j will be relabeled accordingly. Of course it may happen that $-m_r = -1$, in which case there is only one exceptional filler; the argument in this case will be quite similar and is left to the reader.

We set $\bar{\zeta}_j = -\log \bar{\mu}_s(\bar{F}'_j)$. By [1], Lemma 5.1, we have $-\log \bar{\mu}_s(\bar{F}) = \bar{\zeta}'_{-1} + \bar{\zeta}'_0 + \bar{\zeta}'_1 + \cdots + \bar{\zeta}'_{t-2}$. An important feature of $\bar{\zeta}'_j$ is that its distribution depends not only on $l_j{}'$ but also on $u_j{}'$, hence on r (because the condition $u_j{}' < r$ for $j > -1$ alters the distribution of $u_j{}'$). This is because the part of the filler corresponding to the run of markers of length $ku_j{}'$ is now filled by a random sequence. To overcome this difficulty we use again the random variables $u_j{}'$ and $u_j{}''$. We recall that the $u_j{}'$, $j = 1, 2, \ldots, t-2$, can be viewed as produced by a random experiment with $u_j{}' = u_j{}'' - w_j$, where the $u_j{}''$ are independent and distributed as u_1. It should be clear that to each $u_j{}''$, $j > 0$, there corresponds an extension $\bar{F}''_j$ of $\bar{F}'_j$ of length $ku_j{}'' + l_j{}'$. Here $\bar{F}'_j$ has been extended by adjoining a prefix consisting of a run of w_j markers chosen at random according to the distribution of the Markov process so that the resulting $\bar{F}''_j$ has the same distribution as the unconditional $\bar{F}_1$. The random variables $\bar{\zeta}''_j = -\log \bar{\mu}_s(\bar{F}''_j)$ are independent and, for $j > 0$, distributed as $-\log \bar{\mu}_s(\bar{F}_1)$, so their distribution is independent of r. We will apply Lemma 1 to the $\bar{\zeta}''_j - E\bar{\zeta}''_j$. This will be possible because $Ee^{\alpha\bar{\zeta}''_j} < \infty$ for some $\alpha > 0$. Indeed, the length of $\bar{F}''_j$ is distributed as $ku_1 + l_1$. The lengths ku_1 and l_1 are determined by a Bernoulli process, so are independent and have generating functions which are analytic on a disk of radius > 1. The generating function G of $ku_1 + l_1$ therefore has the same property, whence

$$E(e^{\alpha\bar{\zeta}''_j}) \leq \sum_t e^{\alpha \log(1/p_{\min})^t} P(ku_1 + l_1 = t) = G(1/{p_{\min}}^\alpha) < \infty$$

for α small enough. Similarly we get $Ee^{\alpha\bar{\zeta}_0''} < \infty$. The event B_r will be defined as before, so it is determined by the Bernoulli process X. As in the Markov-to-Bernoulli case we can write

$$\begin{aligned}
&P(B_r^c,\ \bar{F} \text{ is bad} \mid b_r = t) = P(B_r^c,\ \bar{\mu}_s(\bar{F}) < e^{-(\bar{f}+\epsilon)l(s_r)} \mid b_r = t)\\
=&P(B_r^c,\ \bar{\zeta}_{-1}' + \bar{\zeta}_0' + \bar{\zeta}_1' + \ldots + \bar{\zeta}_{t-2}' > (\bar{f}+\epsilon)l(s_r) \mid b_r = t)\\
\le&P(B_r^c,\ \bar{f}\left|\lambda - \frac{l(s_r)}{t}\right| > \frac{\epsilon k}{2} \mid b_r = t)\\
&+ P\left(B_r^c,\ \left|\frac{\bar{\zeta}_{-1}' + \bar{\zeta}_0' + \ldots + \bar{\zeta}_{t-2}'}{t} - \bar{f}\lambda\right| > \frac{\epsilon l(s_r)}{2t} \mid b_r = t\right).
\end{aligned}$$

The first term vanishes by the choice of δ and the second is bounded by

$$\begin{aligned}
&P\left(\frac{\bar{\zeta}_{-1}'}{t} > \frac{\epsilon l(s_r)}{6t} \mid b_r = t\right) + P\left(\frac{\bar{\zeta}_0'}{t} > \frac{\epsilon l(s_r)}{6t} \mid b_r = t\right)\\
&+P\left(B_r^c,\ \left|\frac{\bar{\zeta}_1' + \ldots + \bar{\zeta}_{t-2}'}{t} - \bar{f}\lambda\right| > \frac{\epsilon l(s_r)}{6t} \mid b_r = t\right)\\
\le\ &P\left(\bar{\zeta}_{-1}' > \frac{\epsilon kt}{6}\right) + P\left(\bar{\zeta}_0' > \frac{\epsilon kt}{6}\right) + P\left(\left|\frac{\bar{\zeta}_1' + \ldots + \bar{\zeta}_{t-2}'}{t-2} - \bar{f}\lambda\right| > \frac{\epsilon\lambda}{8} - \frac{2\bar{f}\lambda}{t-2}\right).
\end{aligned}$$

The first summand is bounded by $P(\bar{\zeta}_1 + kr\log(1/p_{\min}) > \epsilon kt/6) \le e^{-\beta_4 t} + \sigma_2(r,t)$, where $\beta_4 > 0$ and $\sigma_2(r,t)$ vanishes for $t > 12\epsilon^{-1}r\log(1/p_{\min})$. It is clear that the second summand is bounded by $e^{-\beta_5 t}$ for some $\beta_5 > 0$. For the third summand we will use the random variables $\bar{\zeta}_j''$. First note that the variables

$$\theta_j = \bar{\zeta}_j'' - \bar{\zeta}_j'$$

satisfy the inequality $\theta_j \le \bar{\zeta}_j''$. Now using Lemma 1 we can see that for large t the third summand is bounded by

$$P\left(\left|\frac{\bar{\zeta}_1'' + \cdots + \bar{\zeta}_{t-2}''}{t-2} - \bar{f}\lambda\right| > \frac{\epsilon\lambda}{20}\right) + P\left(\frac{\theta_1 + \cdots + \theta_{t-2}}{t-2} > \frac{\epsilon\lambda}{20}\right),$$

which is less than or equal to

$$\frac{1}{(t-2)^4} + P\left(\left|\frac{\theta_1 + \cdots + \theta_{t-2}}{t-2} - E\theta_1\right| > \frac{\epsilon\lambda}{20} - E\theta_1\right) \le \frac{2}{(t-2)^4},$$

because if k is chosen sufficiently large then

$$\begin{aligned} E\theta_1 \leq & E(1_{\{\theta_1 \neq 0\}} \bar{\zeta}_1'') \\ = & \sum_{i=r}^{\infty} E(\bar{\zeta}_1 \mid u_1 = i) P(u_1 = i) \\ \leq & \sum_{i=r}^{\infty} ik \log(1/p_{\min}) P(u_1 = i) \\ \leq & k \log(1/p_{\min})(1-\eta) \sum_{i=r}^{\infty} i\eta^{i-1} \\ \leq & 2k \log(1/p_{\min}) r\eta^{r-1} < \epsilon\lambda/40 \end{aligned}$$

for all $r > 1$. We have obtained, for $r > 1$ and all sufficiently large t,

$$P(B_r^c,\ \bar{F} \text{ is bad} \mid b_r = t) \leq e^{-\beta_4 t} + \sigma_2(r,t) + e^{-\beta_5 t} + \frac{2}{(t-2)^4}.$$

Therefore, by a calculation as in Section 6 we get

$$P(B_r^c,\ \bar{F} \text{ is bad}) \leq C_{17}(k) r^3 \eta^{2r}.$$

The proof of the convergence of $S_3(p)$ is now concluded along the same lines as the convergence of $S_2(p)$ in the Markov-to-Bernoulli case. Since the coding from $\bar{X}$ with an independent marker process to the given mixing Markov process of a lower entropy simply consists in lumping certain states (so it has code length equal to one), we obtain the following result.

THEOREM 2: *Let the processes X and $\bar{X}$ be Bernoulli and mixing Markov, respectively. If $h(X) > h(\bar{X})$, then there exists a finitary coding from X to $\bar{X}$ with the property that the code length is an L^p random variable for all $p < 2$.*

9. Markov-to-Markov coding

In this section we consider two mixing Markov processes X_1 and X_2 such that $h(X_1) > h(X_2)$. According to [1] there exist a Bernoulli process Y and a mixing Markov process Z with "independent" markers (common for Y and Z) such that $h(X_1) > h(Y) > h(Z) > h(X_2)$ and X_2 is obtained by lumping certain states in Z. By Theorem 1 there exists a finitary code ϕ from X_1 to Y such that its code length is an L^p function for all $p < 2$. Similarly, by Theorem 2, there is a $\psi : Y \to Z$ with the same property of the code length. We now wish to examine the code length of the composed coding $X_1 \to X_2$.

LEMMA 2: *Let X, Y, Z be arbitrary stationary processes and let $\phi : X \to Y$ and $\psi : Y \to Z$ be finitary codes. Assume that for some $p_1, p_2 > 1$ with $p_1 \leq p_2 + 1$ the code length of ϕ is in L^p for all $p < p_1$ and the code length of ψ is in L^p for all $p < p_2$. Then the composed code $\psi \circ \phi$ has code length in L^p for all $p < p_1 p_2/(p_2 + 1)$.*

Proof: Let C_1, C_2, C be the code length of $\phi, \psi, \psi \circ \phi$, respectively. By the definition of code length we have

$$C(x) \leq C_2(\phi(x)) + 2\max\{C_1(T^j x) : |j| \leq C_2(\phi(x))\}.$$

Since every $p < p_1 p_2/(p_2 + 1)$ is less than p_2 and by assumption $E({C_2}^p) < \infty$, it suffices to consider the second term. Recall that for every nonnegative random variable ξ and a fixed $q > 0$ the quantity $E(\xi^p)$ is finite for all $p < q$ iff for all $p < q$ we have $P(\xi > \lambda) = O(\lambda^{-p})$ as $\lambda \to \infty$.

Now we set $\xi = \max\{C_1(T^j x) : |j| \leq C_2(\phi(x))\}$ and let θ be a positive real number. For any $q_1 < p_1$ and $q_2 < p_2$ we have

$$\begin{aligned} P(\xi > \lambda) =& P(\xi > \lambda,\, C_2(\phi(x)) > \lambda^\theta) + P(\xi > \lambda,\, C_2(\phi(x)) \leq \lambda^\theta) \\ \leq& O(\lambda^{-\theta q_2}) + \sum_{|j| \leq \lambda^\theta} P(C_1 T^j x > \lambda) \\ =& O(\lambda^{-\theta q_2}) + \lambda^\theta O(\lambda^{-q_1}). \end{aligned}$$

Now setting $\theta = p_1/(p_2 + 1)$ yields $P(\xi > \lambda) = O(\lambda^{-p})$ for all $p < p_1 p_2/(p_2 + 1)$, which concludes the proof of the lemma. ∎

The next result follows readily.

THEOREM 3: *Let X_1 and X_2 be mixing Markov processes such that $h(X_1) > h(X_2)$. Then there exists a finitary coding from X_1 to X_2 such that the code length is in L^p for all $p < 4/3$.*

The following corollary applies in particular to the Keane–Smorodinsky coding of Bernoulli processes of unequal entropies.

COROLLARY: *For $n \geq 2$ let $X_1, \ldots, X_{n+1}$ be stationary processes and $\phi_i : X_i \to X_{i+1}$, $i = 1, \ldots, n$, be finitary codes with code length in L^p for every $p < 2$. Then the expected length of the composed code $\phi_n \circ \cdots \circ \phi_1$ is finite.*

Proof: We prove by induction that the length of the composed code is an L^p function for all $p < 2^n/(2^n - 1)$. To this end we let $\phi = \phi_1$, $p_1 = 2$, $\psi = \phi_n \circ \cdots \circ \phi_2$, $p_2 = 2^{n-1}/(2^{n-1} - 1)$. Now apply Lemma 2. ∎

References

[1] M. A. Akcoglu, A. del Junco and M. Rahe, *Finitary codes between Markov processes*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **47** (1979), 305–314.

[2] S. N. Bernstein, *Probability Theory* (in Russian), Gosizdat, Moscow, 1927.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York, 1966.

[4] T. Hamachi and M. S. Keane, *Homomorphisms of noncommutative Bernoulli schemes, II. Finitary codes*, preprint.

[5] M. Keane and M. Smorodinsky, *A class of finitary codes*, Israel Journal of Mathematics **26** (1977), 352–371.

[6] J. Serafin, *The finitary coding of two Bernoulli schemes with unequal entropies has finite expectation*, Indagationes Mathematicae. New Series **7** (1996), 503–519.