Average Redundancy of the Shannon Code for Markov Sources

Neri Merhav
Dep. of Electrical Engineering
Technion – Israel Institute of Technolgy
Technion City, Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

Wojciech Szpankowski
Dep. of Computer Science
Purdue University
West Lafayette, IN 47909-2066, U.S.A
Email: spa@cs.purdue.edu

***Abstract*—It is known that for memoryless sources, the average and maximal redundancy of fixed–to–variable length codes, such as the Shannon and Huffman codes, exhibit two modes of behavior for long blocks. It either converges to a limit or it has an oscillatory pattern, depending on the irrationality or rationality, respectively, of certain parameters that depend on the source. Here, we extend these findings for the Shannon code to the case of a Markov source, which is considerably more involved. We provide a precise characterization of the redundancy of the Shannon code redundancy for a class of irreducible, periodic and aperiodic Markov sources.**

I. Introduction

Recent years have witnessed a resurgence of interest in redundancy rates of lossless coding, see, e.g., [1], [3], [4], [7], [9], [10], [12], [13], [14]. In particular, in [14] Szpankowski derived asymptotic expressions of the (unnormalized) average redundancy R_n, as a function of the block length n, for the Shannon code, the Huffman code, and other codes, focusing primarily on the binary memoryless source parameterized by p – the probability of '1'. A rather interesting behavior of R_n was revealed in [14], especially in the cases of the Shannon code and the Huffman code: When $\alpha \stackrel{\triangle}{=} \log_2[(1-p)/p]$ is irrational, then R_n converges to a constant (which is $1/2$ for the Shannon code), as $n \to \infty$. On the other hand, when α is rational, R_n has a non–vanishing oscillatory term whose fundamental frequency and amplitude depend on the source statistics in an explicit manner.

More precisely, confining the discussion to the Shannon code, in [14] the average unnormalized redundancy

$$R_n = \boldsymbol{E}\left\{\lceil -\log_2 P(X_1,\ldots,X_n)\rceil + \log_2 P(X_1,\ldots,X_n)\right\}, \tag{1}$$

was analyzed for large n, assuming that the source P governing the data to be compressed, $X_1, X_2, \ldots$, is a binary memoryless source. A straightforward extension (see also [10]) of the Shannon–code redundancy result of [14], to a general r–ary alphabet memoryless source, with letter probabilities $p_1, \ldots, p_r$, yields the following expression:

$$R_n = \begin{cases} \frac{1}{2} + \frac{1}{M}\left(\frac{1}{2} - \langle \beta M n\rangle\right) + o(1) & \text{all } \{\alpha_j\} \text{ are rational} \\ \frac{1}{2} + o(1) & \text{otherwise} \end{cases} \tag{2}$$

where $\beta \stackrel{\triangle}{=} -\log p_1$, $\alpha_j = \log p_j/p_1$, $j = 2, 3, \ldots, r$, $\langle u\rangle$ is the fractional part of a real number u (i.e., $\langle u\rangle = u - \lfloor u\rfloor$), and M is the smallest common multiple of all denominators of the rational numbers $\{\alpha_j\}$ when presented as ratios between two relatively prime integers. This erratic behavior, where R_n is either convergent (and then the limit is always $1/2$) or oscillatory, depending on the rationality of $\{\alpha_j\}$, was related in [10] to wave diffraction patterns of scattering from partially disordered media, where the existence/non–existence of Bragg peaks depends on the rationality/irrationality of certain optical distance ratios.

Our goal is to extend the scope of this analysis to irreducible Markov sources and to evaluate precisely (for large n) the average redundancy of the Shannon code for a finite alphabet, first order Markov source with given transition probabilities. In doing so, we also provide a more complete analysis than in [10] and [14]. As will be seen, this extension to the Markov case appears rather non–trivial, both from the viewpoint of the conditions for oscillatory behavior and from the aspect of the asymptotic expression of R_n in the oscillatory mode. These depend strongly on the dominant eigenvalues and on the detailed structure of the matrix of transition probabilities. For example, in contrast to the memoryless case, where there is only one oscillatory term, when it comes to the Markov case, in the oscillatory mode there are, in general, contributions from multiple oscillatory terms, and in the convergent mode, R_n may converge to a constant other than $1/2$ (see Example 2). Moreover, it turns out that the behavior of the redundancy depends quite strongly on important dynamical properties of the Markov chain, such as reducibility/irreducibility and periodicity/aperiodicity.

We begin our study from the relatively simple case where all single–step state transitions have positive probability. Our main result in Section 2, Theorem 1, is then an extension of formula (2) to the Markov case with strictly positive state transition probabilities. To give the reader a general idea of this theorem, an informal description of it can be stated as follows: Rather than the parameters $\{\alpha_j\}$ of the memoryless case, we now define a matrix $\{\alpha_{jk}\}_{j,k=1}^r$ of log–ratios of certain transition probabilities (the exact definition will be provided in the sequel). If at least one of these parameters is irrational, then similarly as in the memoryless case, $R_n = \frac{1}{2} + o(1)$. If, on the other hand, all these parameters are rational, then as in the memoryless case, let M be their smallest common denominator. In this case, $R_n = \Omega_n + o(1)$, for "most large values" of n, where Ω_n is a linear combination of certain functions of n, for which we have an explicit formula in terms of the source parameters. These functions oscillate as n varies, with amplitude $1/M$ and a fundamental frequency that depends on the source parameters.

Then we relax the strict positivity assumption, but still assume the Markov chain to be irreducible. Under this assumption, we first assume that the chain is also aperiodic, and then further extend the scope to allow periodicity. It is also demonstrated (in Example 2) that the irreducibility assumption is essential, since the above described two–mode behavior ceases to exist when this assumption is dropped.

We should point out that minimax redundancy and regret for the *class* of Markov sources were studied in the past – see, e.g., [7], [12]. Interestingly enough, the minimax regret for memoryless and Markov sources does not exhibit the two–mode behavior of either convergent or oscillatory mode [3]. This dichotomy, of convergent vs. oscillatory behavior, with dependence on rationality/irrationality of certain parameters, is a well recognized phenomenon in mathematics and physics, ranging across a large variety of areas, including renewal theory, ergodic theory [5], local limit theorems and large deviations for discrete distributions [2], This phenomenon, however, was observed in information theory only relatively recently [5], [14]. On the other hand, the oscillatory phenomenon for discrete random structures is a well known fact in analysis of algorithms [15], and also in information theory [3], [9], [15].

In this conference version we omit most proofs which can be found in the full version of this paper [11].

II. Main Results for Positive Transition Matrices

Throughout, we adopt the customary notation conventions: Random variables will be denoted by capital letters (e.g., X), specific values they may take will be denoted by the corresponding lower–case letters (e.g., x), and their alphabets will be denoted by the corresponding calligraphic letters (e.g., $\mathcal{X}$). Random vectors of length n (e.g., $(X_1, X_2, \ldots, X_n)$) will be denoted by capital letters superscripted by n (e.g., X^n), and specific values of these vectors (e.g., $(x_1, x_2, \ldots, x_n)$) will be denoted by lower–case letters superscripted by n (e.g., x^n). Finally, the set of vectors of length n, with components taking on values in $\mathcal{X}$, will be denoted by $\mathcal{X}^n$. Logarithms will always be understood to be taken w.r.t. the base 2. The function $\mathcal{I}(\cdot)$ will denote the indicator function, that is, for a given statement E, $\mathcal{I}(E) = 1$ if E is true, and $\mathcal{I}(E) = 0$ if E is false.

Consider a source sequence $X_1, X_2, \ldots, X_t \in \mathcal{X} = \{1, 2, \ldots, r\}$ (r – positive integer), $t = 1, 2, \ldots$, governed by a first–order Markov chain with a given matrix P of state–transition probabilities $\{p(j|k)\}_{j,k=1}^r$. The initial state probabilities will be denoted by p_k, $k = 1, 2, \ldots, r$. The stationary state probabilities will be denoted by π_k, $k = 1, 2, \ldots, r$. Thus, the probability of a given source string $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, under the given Markov source, is

$$\mu(x^n) = p_{x_1} \prod_{t=2}^n p(x_t|x_{t-1}). \tag{3}$$

The average unnormalized redundancy of the Shannon code is defined as

$$R_n \stackrel{\triangle}{=} \boldsymbol{E}\{\lceil -\log \mu(X^n) \rceil + \log \mu(X^n)\}, \tag{4}$$

where here and throughout the sequel, $\boldsymbol{E}\{\cdot\}$ denotes the expectation operator w.r.t. the underlying Markov source μ.

As mentioned in the Introduction, we assume that P is irreducible, that is, there is positive probability to pass from every state $j \in \mathcal{X}$ to every state $k \in \mathcal{X}$ within a finite number of steps. Furthermore, the *period* d_j of a state j is the greatest common divisor of all integers n for which $\Pr\{X_n = j | X_0 = j\} > 0$. A state is called periodic if $d_j > 1$ and aperiodic if $d_j = 1$. The case where all entries of P are positive, henceforth referred to as the case of a *positive matrix* P, is obviously a case of an irreducible, aperiodic Markov chain.

Theorem 1: Consider the Shannon code of block length n for a Markov source μ with a given vector $\boldsymbol{p} = (p_1, \ldots, p_r)$ of initial state probabilities and a positive state transition matrix P. Define

$$\alpha_{jk} = \log \left[\frac{p(j|1)p(j|j)}{p(k|1)p(j|k)} \right], \qquad j, k \in \{1, 2, \ldots, r\}. \tag{5}$$

Then, the redundancy R_n is characterized as follows:
(a) If not all $\{\alpha_{jk}\}$ are rational, then

$$R_n = \frac{1}{2} + o(1). \tag{6}$$

(b) If all $\{\alpha_{jk}\}$ are rational, then for every $j, k \in \{1, \ldots, r\}$, let

$$\begin{aligned}\zeta_{jk}(n) \quad &= \quad M[-(n-1)\log p(1|1) + \log p(j|1) - \\ &\qquad \log p(k|1) - \log p_j],\end{aligned} \tag{7}$$

and

$$\Omega_n = \frac{1}{2}\left(1 - \frac{1}{M}\right) + \frac{1}{M} \sum_{j=1}^r \sum_{k=1}^r p_j \pi_k \varrho[\zeta_{jk}(n)], \tag{8}$$

where $\varrho(u) \stackrel{\triangle}{=} \lceil u \rceil - u$ and M is the smallest common integer multiple of the denominators of $\{\alpha_{jk}\}$, when each one of these numbers is represented as a ratio between two relatively prime integers. Then, there exists a positive sequence $\xi_n \to 0$, which depends only the source parameters, such that R_n is upper bounded and lower bounded as follows:

$$R_n \le \Omega_n + \frac{1}{M} \sum_{j=1}^r \sum_{k=1}^r p_j \pi_k \mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\} + o(1). \tag{9}$$

$$R_n \ge \Omega_n - \frac{1}{M} \sum_{j=1}^r \sum_{k=1}^r p_j \pi_k \mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\} - o(1). \tag{10}$$

As a technical comment, it should be pointed out that the choice of the index 1 in the conditioning of $p(j|1)$ and

$p(k|1)$, that appear in the definition of α_{jk} and in (7), is completely arbitrary. One may choose any other index in $\{1, 2, \ldots, r\}$, as long as it is the same index in both places in the expression of α_{jk}, as well as in the second and third terms in the square brackets of (7). Also, $p(1|1)$ in (7) can be replaced independently by $p(l|l)$ for any $l \in \{1, 2 \ldots, r\}$.

Discussion. Theorem 1 tells us that, similarly as in the memoryless case, in the positive matrix case, R_n has two modes of behavior. In the convergent mode, which happens when at least one α_{jk} is irrational, $R_n \to 1/2$. In the oscillatory mode, which happens when all $\{\alpha_{jk}\}$ are rational, R_n oscillates and it asymptotically coincides with Ω_n for *most large values*[*] of n, provided that $\log p(1|1)$ is irrational. This follows from the following consideration: If $\log p(1|1)$ is irrational, then by Weyl's equidistribution theorem [8], the sequences $\{\zeta_{jk}(n)\}_{n\geq 1}$ are uniformly distributed modulo 1, i.e., they fill the unit interval mod 1 with a uniform density as n exhausts the positive integers. Thus, for every fixed ξ, $\varrho[\zeta_{jk}(n)] \notin (\xi, 1-\xi)$ for a fraction 2ξ of the values of n. This means that for $\xi_n \to 0$, the terms $\mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\}$ vanish for most large values of n, and then the lower bound and the upper bound on R_n asymptotically coincide with Ω_n. If, on the other hand, $\log p(1|1)$ is rational, then $\varrho[\zeta_{jk}(n)]$ are periodic sequences. If for none of the values n in a period, $\varrho[\zeta_{jk}(n)] = 0$, then beyond a certain value of n, ξ_n is smaller than the minimum value of $\varrho[\zeta_{jk}(n)]$ along the period and $1 - \xi_n$ is larger than the maximum, and so, $\mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\}$ all vanish for *all* large n. The expression

$$\frac{1}{M}\sum_{j=1}^{r}\sum_{k=1}^{r} p_j \pi_k \mathcal{I}\{\varrho[\zeta_{jk}(n)] \notin (\xi_n, 1-\xi_n)\},$$

which generates the gap between the upper bound and the lower bound on R_n, can be interpreted as an asymptotic approximation of the probability that $-\log \mu(X^n)$ falls in the vicinity (within distance $O(\xi_n)$) of an integer. For example, when the source is purely dyadic ($M = 1$), then $-\log \mu(X^n)$ is integer with probability 1, and indeed, the expression in the last display is equal to 1. In this case, Theorem 1 is useless, but it is also redundant, because in this case, we clearly know that R_n vanishes. The reason for this "uncertainty" around integer values of $-\log \mu(X^n)$ is that these are the discontinuity points of the function $\varrho[-\log \mu(X^n)]$, and in the proof of Theorem 1, the function ϱ is expanded as a series of trigonometric polynomials whose convergence is problematic in the neighborhood of discontinuities. Thus, we believe that the uncertainty in the characterization of R_n around these points should be attributed more to the limitations of the analysis methods than to the real behavior of R_n. In other words, we conjecture that, in fact, $R_n = \Omega_n + o(1)$ for *all* large n, and not just for most large values of n.

[*]The statement "R_n asymptotically coincides with Ω_n for most large values of n" means that for every $\epsilon > 0$, the fraction of values of n, within the range $\{1, \ldots, N\}$, for which $|R_n - \Omega_n| > \epsilon$, tends to zero as $N \to \infty$.

The expression of the oscillatory case, Ω_n, is not quite intuitive at first glance, therefore, in this paragraph, we make an attempt to give some quick insight, which captures the essence of the main points. The arguments here are informal and non-rigorous (the rigorous proof can be found in [11]). The Fourier series expansion of the periodic function ϱ is given by

$$\varrho(u) = \frac{1}{2} + \sum_{m\neq 0} a_m e^{2\pi i m u} \tag{11}$$

and the important fact about the coefficients is that they are inversely proportional to m, so that for every two integers k and m, $a_{m\cdot k} = a_m/k$. Now, when computing $R_n = \boldsymbol{E}\{\varrho[-\log \mu(X^n)]\}$, let us take the liberty of exchanging the order between the expectation and the summation, i.e.,

$$R_n = \frac{1}{2} + \sum_{m\neq 0} a_m \boldsymbol{E}\{e^{-2\pi i m \log \mu(X^n)}\}. \tag{12}$$

It turns out that under the conditions of the oscillatory mode, $\boldsymbol{E}\{e^{-2\pi i m \log \mu(X^n)}\}$ tends to zero as $n \to \infty$ for all m, except for multiples[†] of M, namely, $m = \ell M$, $l = \pm 1, \pm 2, \ldots$. Thus, for large n, we have

$$\begin{aligned} R_n &\approx \frac{1}{2} + \sum_{\ell\neq 0} a_{\ell M} \boldsymbol{E}\{e^{-2\pi i \ell M \log \mu(X^n)}\} \\ &= \frac{1}{2} + \frac{1}{M}\left\{\boldsymbol{E}\varrho[-M\log\mu(X^n)] - \frac{1}{2}\right\} \\ &= \frac{1}{2}\left(1 - \frac{1}{M}\right) + \frac{1}{M}\boldsymbol{E}\varrho[-M\log\mu(X^n)]. \end{aligned} \tag{13}$$

Now, consider the set of all $\{x^n\}$ that begin from state $x_1 = j$ and end at state $x_n = k$. Their total probability is about $p_j \pi_k$ for large n since X_n is almost independent of X_1. It turns out that all these sequences have exactly the same value of $\varrho[-M\log\mu(x^n)]$, which is exactly $\varrho[\zeta_{jk}(n)]$ (or, in other words, $\varrho[-M\log\mu(x^n)] = \varrho[\zeta_{x_1 x_n}(n)]$ independently of $x_2, \ldots, x_{n-1}$) and this explains the expression of Ω_n. The reason for this property of $\varrho[-M\log\mu(x^n)]$ is the rationality conditions $\langle M \cdot \alpha_{uv}\rangle = 0$, $u, v \in \{1, 2, \ldots, r\}$, which imply that $\langle M \log p(x_t|x_{t-1})\rangle = \langle M \log[p(x_t|1)p(1|1)/p(x_{t-1}|1)]\rangle$, and so, that mod 1 we have

$$\langle -M\log\mu(x^n)\rangle = \langle -M\log p_j\rangle + \sum_{t=2}^{n}\langle -M\log p(x_t|x_{t-1})\rangle$$

$$= \langle -M\log p_j\rangle + \sum_{t=2}^{n}\langle -M\log[p(x_t|1)p(1|1)/p(x_{t-1}|1)]\rangle$$

which, thanks to the telescopic summation, is easily seen to coincide with the fractional part of $\zeta_{jk}(n)$, and of course, $\varrho[\zeta_{jk}(n)]$ depends on $\zeta_{jk}(n)$ only via its fractional part.

Main Ideas of the Proof. We sketch here informally the main ingredients of the proof (see [11]). Let us pick up at equation

[†]The convergent mode can be treated as a special case of this statement with $M = \infty$.

(12). We can re-write R_n as

$$\frac{1}{2}+\sum_{m\neq 0} a_m \sum_{\boldsymbol{x}\in\mathcal{X}^n}\prod_{t=1}^{n} p(x_t|x_{t-1})\exp\left\{-2\pi i m\log p(x_t|x_{t-1})\right\}.$$

Define the $r\times r$ complex matrix A_m whose entries are

$$a_{jk}(m)=p(k|j)\exp\left[-2\pi i m\log p(k|j)\right], \qquad j,k=1,\ldots,r. \tag{14}$$

Also define the r–dimensional column vectors

$$\boldsymbol{c}_m=(p_1\exp[-2\pi i m\log p_1)],\ldots,p_r\exp[-2\pi i m\log p_r])^T, \tag{15}$$

and $\mathbf{1}=(1,1,\ldots,1)^T$, where the superscript T denotes vector/matrix transposition. Then

$$R_n=\frac{1}{2}+\sum_{m\neq 0} a_m\cdot \boldsymbol{c}_m^T A_m^{n-1}\mathbf{1}. \tag{16}$$

Let $\boldsymbol{l}_{i,m}$ and $\boldsymbol{r}_{i,m}$ be, respectively, the left eigenvector and the right eigenvector (such that the scalar product $\boldsymbol{l}_{i,m}^T\boldsymbol{r}_{i,m}=1$) pertaining to the eigenvalue $\lambda_{i,m}$ ($i=1,2,\ldots,r$) of the matrix A_m, listed in non-increasing order of their modulus, that is,

$$\rho(A_m):=|\lambda_{1,m}|\geq|\lambda_{2,m}|\geq\cdots|\lambda_{r,m}|. \tag{17}$$

Then by the spectral representation of matrices [6], we have

$$A_m^{n-1}\mathbf{1}=\sum_{i=1}^{r}\lambda_{i,m}^{n-1}\cdot\boldsymbol{l}_{i,m}^T\mathbf{1}\cdot\boldsymbol{r}_{i,m}, \tag{18}$$

leading to

$$R_n=\frac{1}{2}+\sum_{m\neq 0}a_m\cdot\sum_{i=1}^{r}\lambda_{i,m}^{n-1}\cdot\boldsymbol{l}_{i,m}^T\mathbf{1}\cdot\boldsymbol{c}_m^T\boldsymbol{r}_{i,m}. \tag{19}$$

Now, in principle, if all eigenvalues $\lambda_{i,m}<1$, then by Fejer's theorem we can conclude that $R_n\to\frac{1}{2}$. For the oscillatory case we need to argue that the largest modulus of the eigenvalues (spectral radius) of A_m and P are equal to 1. This can be accomplished by appealing to the following lemma that is at the heart of our analysis.

Lemma 1: [6, Theorem 8.4.5, p. 509] Let $F=\{f_{kj}\}$ and $G=\{g_{kj}\}$ be two $r\times r$ matrices. Assume that F is a real, non–negative and irreducible matrix, G is a complex matrix, and $f_{kj}\geq|g_{kj}|$ for all $k,j\in\{1,2,\ldots,r\}$. Then, the spectral radii $\rho(G)\geq\rho(F)$ with equality if and only if there exist real numbers s, and $w_1,\ldots,w_r$ such that $G=e^{2\pi i s}DFD^{-1}$, where $D=\mathrm{diag}\{e^{2\pi i w_1},\ldots,e^{2\pi i w_r}\}$.

In our case, P is represented by F of Lemma 1 and the matrix A_m takes the role of G. By non–negativity and irreducibility of P, we conclude that its spectral radius is $\rho(P)=1$. It is also obvious that the elements of P are the absolute values of the corresponding elements of A_m, and so, all the conditions of Lemma 1 clearly apply. The lemma then tells us that $\rho(A_m)=\rho(P)=1$ if and only if there exist real numbers s and $w_1,\ldots w_r$ such that:

$$-m\log p(j|k)=(s+w_k-w_j)\bmod 1, \quad j,k=1,\ldots,r, \tag{20}$$

where $x=y$ mod 1 means that the fractional parts of x and y are equal, that is, $\langle x\rangle=\langle y\rangle$.

Example 1. Consider a Markov source for which the rows of P are all permutations of the first row, which is $\boldsymbol{p}=(p_1,\ldots,p_r)$. Now, assuming that $\alpha_j\stackrel{\triangle}{=}\log(p_1/p_j)$ are all rational, let M be the least common multiple of their denominators (i.e., the common denominator) when each one of them is expressed as a ratio between two relatively prime integers. Then,

$$\begin{aligned}\varrho[\zeta_{jk}(n)] &= \varrho[-M(n-1)\log p(1|1)+M\log p(j|1)\\ &- M\log p(k|1)-M\log p_j]=\\ &=\varrho[-M(n-1)\log p_1+M\log p_j-M\log p_k-M\log p_j]\\ &=\varrho(-Mn\log p_1),\end{aligned}$$

where in the last step, we have used the fact that $(M\log p_1-M\log p_k)$ is integer and that ϱ is a periodic function with period 1. Thus,

$$\begin{aligned}R_n &= \frac{1}{2}\left(1-\frac{1}{M}\right)+\frac{1}{M}\sum_{j=1}^{r}\sum_{k=1}^{r}p_j\pi_k\varrho[\zeta_{jk}(n)]+o(1)\\ &= \frac{1}{2}\left(1-\frac{1}{M}\right)+\frac{1}{M}\varrho(-nM\log p_1)+o(1). \end{aligned} \tag{21}$$

If not all α_j are rational, then $R_n\to 1/2$, as predicted by Theorem 1. Note that the memoryless source is a special case of this example, where the rows of P are all identical to the first row, $(p_1,\ldots,p_r)$. Indeed, eq. (21) coincides with the expression of the memoryless case (see [10], [14] and the Introduction of this paper).

III. Extensions

We now discuss some extensions of Theorem 1, namely we first assume that P corresponds to an irreducible aperiodic Markov source, and then even drop the aperiodicity constraint.

Irreducible Aperiodic Markov Sources. When some of the entries of the matrix P vanish, then obviously, Theorem 1 cannot be used as is since the corresponding parameters α_{jk} are no longer well defined. Lemma 1 can still be used as long as P is irreducible, but more caution should be exercised.

For example, if one or more diagonal element of P is positive, and at least one row of P is strictly positive, say, row number l, then the rationality condition of Theorem 1 (see also (20)) is replaced by the condition that

$$\alpha'_{jk}=\log\left[\frac{p(j|l)p(l|l)}{p(k|l)p(j|k)}\right] \tag{23}$$

must be rational for all (j,k) with $p(j|k)>0$. The bounds on R_n in the oscillatory mode would be exactly as in Theorem 1. Theorem 2 below describes the behavior of R_n for general irreducible, aperiodic Markov sources.

Theorem 2: Consider the Shannon code of block length n for an irreducible aperiodic Markov source. Let M be defined as the smallest positive integer m such that

$$\rho(A_m)\equiv|\lambda_{1,m}|=1 \tag{24}$$

and as $M = \infty$ if (24) does not hold for any positive integer m. Then, R_n is characterized as follows:
(a) If $M = \infty$, then

$$R_n = \frac{1}{2} + o(1). \tag{25}$$

(b) If $M < \infty$, then the bounds of Theorem 1, part (b), hold with $\zeta_{jk}(n)$ being redefined according to

$$\zeta_{jk}(n) = M[(n-1)s + w_j - w_k - \log p_j], \tag{26}$$

where $s = \arg\{\lambda_{1,M}\}/2\pi$ and $w_j = \arg\{x_j\}/2\pi$, $j = 1, 2, \ldots, r$, x_j being the j–th component of the right eigenvector $\boldsymbol{x}$ of A_M, which is associated with the dominant eigenvalue $\lambda_{1,M}$.

Finally, it is instructive to demonstrate an example of a *reducible* Markov source, for which Theorems 1 and 2 do not hold, and see that even in a simple situation ($r = 2$), once the irreducibility assumption is dropped, the two–mode behavior, predicted by Theorems 1 and 2, disappears.

Example 2. *Reducible Markov source*. Consider the case $r = 2$, where $p(1|2) = 0$ and $\alpha \stackrel{\triangle}{=} p(2|1) \in (0, 1)$. Assume also that $p_1 = 1$ and $p_2 = 0$. Since this is a reducible Markov source (once in state 2, there is no way back to state 1), we cannot use Theorems 1 and 2, but we can still find an asymptotic expression of the redundancy in a direct manner: Note that the chain starts at state '1' and remains there for a random duration, which is a geometrically distributed random variable with parameter $(1-\alpha)$. Thus, the probability of k 1's (followed by $n - k$ 2's) is about $(1-\alpha)^k \cdot \alpha$ (for large n) and so the argument of the function $\varrho(\cdot)$ should be the negative logarithm of this probability. Taking the expectation w.r.t. the randomness of k, we readily have

$$R_n = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k \varrho[-\log\alpha - k\log(1-\alpha)] + o(1). \tag{27}$$

We see then that there is *no oscillatory mode* in this case, as R_n always tends to a constant that depends on α, in contrast to the convergent mode of Theorems 1 and 2, where the limit is always $1/2$, independently of the source statistics. To summarize, it is observed that the behavior here is very different from that of the irreducible case, characterized by Theorems 1 and 2.

Irreducible Periodic Markov Sources. Consider now an irreducible *periodic* Markov source. The Perron-Frobenius theorem and Lemma 1 still hold [6]. However, the matrix P now has d eigenvalues on the unit circle, namely, all the d–th roots of unity [6], where d is the period, i.e.,

$$\lambda'_t = e^{2\pi i t/d}, \qquad t = 0, 1, \ldots, d-1. \tag{28}$$

The analysis is similar as in the aperiodic case, except that we now have d oscillatory terms, one for each eigenvalue on the unit circle. Indeed, suppose that for some m, the matrix A_m has a modulus–1 eigenvalue $\lambda = e^{2\pi i s}$. Therefore, A_m has the following eigenvalues on the unit circle:

$$\lambda_{t,m} = e^{2\pi i(s+t/d)}, \quad t = 0, 1, \ldots, d-1. \tag{30}$$

Let us relabel, if necessary, the eigenvalues of A_m such that $s \in [0, 1/d)$. This means that the definition of s in Theorem 2 should be restricted to the half open interval $[0, 1/d)$. Thus, Theorem 2 holds except that $\zeta_{jk}(n)$ are replaced by

$$\zeta_{jkt}(n) \stackrel{\triangle}{=} M\left[(n-1)\left(s + \frac{t}{d}\right) + w_j - w_k - \log p_j\right], \tag{31}$$

for $j, k \in \{1, 2, \ldots, r\}$, $t = 0, 1, \ldots, d-1$, and the double summations over (j, k) with weights $p_j\pi_k$, are replaced by corresponding triple summations over (j, k, t) with weights $p_j r_{t,j} l_{t,k}$, where $l_{t,k}$ is the k–th component of $\boldsymbol{l}_t$ and $r_{t,j}$ is the j–th component of $\boldsymbol{r}_t$. Here $\boldsymbol{r}_t$ and $\boldsymbol{l}_t$ are the right- and the left eigenvectors of P that are associated with λ'_t. Note that $r_{0,j} = 1$ and $l_{0,k} = \pi_k$, so for $d = 1$ we indeed obtain the expression (26) of the aperiodic case as a special case.

Acknowledgment

N. Merhav's work was supported by the Israeli Science Foundation (ISF), grant no. 208/08. W. Szpankowski is also a Visiting Professor at ETI, Gdańsk University of Technology, Poland. His work was supported in part by the NSF Science and Technology Center Grant CCF-0939370, NSF Grant CCF-0830140, AFOSR Grant FA8655-11-1-3076, and NSA Grant H98230-11-1-0141.

References

[1] J. Abrahams, "Code and parse trees for lossless source encoding," *Communications in Information and Systems*, vol. 1, no. 2, pp. 113–146, April 2001.
[2] D. Blackwell and J. L. Hodges, "The probability in the extreme tail of a convolution," *Ann. Math. Stat.*, vol. 30, pp. 1113–1120, 1959.
[3] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2686–2707, November 2004.
[4] R. G. Gallager, "Variations on the theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT–24, no. 6, pp. 668–674, November 1978.
[5] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1220-1244, November 1990.
[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[7] P. Jacquet and W. Szpankowski, "Markov types and minimax redundancy for Markov sources, *IEEE Trans. Inform. Theory*, vol. 50, no. 7, pp. 1393–1402, July 2004.
[8] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Wiley, New York, 1974.
[9] G. Louchard and W. Szpankowski, "Average redundancy of the Lempel–Ziv code," *IEEE Trans. Inform. Theory*, vol. 43, no. 1, pp. 2–8, January 1997.
[10] N. Merhav, "Relations between redundancy patterns of the Shannon code and wave diffraction patterns of partially disordered media," *IEEE Trans. Inform. Theory*, vol. 58, no. 6, pp. 3402–3406, June 2012.
[11] N. Merhav and W. Szpankowski, Average Redundancy of the Shannon Code for Markov Sources, arXiv:1208.1225 [cs.IT], August, 2012.
[12] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT–32, no. 4, pp. 526–532, July 1986.
[13] S. A. Savari, "Variable–to–fixed length codes for predictable sources," *Proc. Data Compression Conference (DCC)*, Snowbird, UT, pp. 481–490, 1998.
[14] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block Codes," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2434–2443, November 2000.
[15] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.