Asymptotic Redundancy of the MTF Scheme for Stationary Ergodic Sources

Mitsuharu Arimura, *Member, IEEE,* and Hirosuke Yamamoto, *Senior Member, IEEE*

***Abstract*—The Move-to-front (MTF) scheme is a data-compression method which converts each symbol of a source sequence to a positive integer sequentially, and encodes it to a binary codeword. The compression performance of this algorithm has been analyzed usually under the assumption of the so-called symbol extension. But, in this paper, upper and lower bounds are derived for the redundancy of the MTF scheme without the symbol extension for stationary ergodic sources and Markov sources. It is also proved that for the stationary ergodic first-order Markov sources, the MTF scheme can attain the entropy rate if and only if the transition matrix of the source is a kind of doubly stochastic matrix. Moreover, if the source is a Kth-order Markov source ($K \geq 2$), the MTF scheme cannot attain the entropy rate of the source generally.**

***Index Terms*—Asymptotic redundancy, doubly stochastic process, Markov sources, move-to-front (MTF) scheme, stationary ergodic sources.**

I. Introduction

THE move-to-front (MTF) scheme [1] is a data-compression algorithm, which is equivalent to the Recency–Rank scheme [2] and the Book–Stack scheme [3]. The MTF scheme is often used in several efficient data-compression methods. For instance, in the block-sorting (BS) method [6], the MTF scheme with the Burrows–Wheeler transform attains very high compression performance for many practical files [6].

In the MTF scheme, each symbol X_n of a data sequence $X = X_1X_2X_3\cdots$ is transformed to a positive integer Y_n, and then Y_n is encoded to a binary codeword on the assumption that the process $\{Y_n\}$ is memoryless. The former transformation is called as the *MTF transform*.

Concerning the MTF scheme, some upper bounds of the average redundancy are derived by the information-theoretic analyses [1]–[3]. The average redundancy for stationary sources and the almost-sure redudancy for stationary ergodic sources and asymptotically mean stationary sources are also obtained in [4]. These analyses show that the redundancy approaches to zero as the size of the so-called symbol extension becomes large. Furthermore, the strict symbol-wise probability distribution of Y_n is studied for any independent and identically distributed (i.i.d.) source X in [5].

In these analyses, the symbol extension is introduced to show the asymptotic optimality, or source X is assumed to be i.i.d. In the analyses of the BS method [4], [7]–[9], the symbol extension is also used, because it is difficult to evaluate the performance of the MTF transform used in the BS method without the symbol extension. However, the symbol extension is not usually used in practical applications of the MTF scheme nor the BS method. Moreover, sources applied to the MTF transform are generally not memoryless. Hence, from practical and theoretical points of view, it is worth evaluating the compression performance of the MTF scheme without symbol extension for sources with memory. In this paper, the redundancy of the MTF scheme is derived for such practical cases.

This paper is organized as follows. In Section II, the MTF transform algorithm is reviewed, and the redundancy of the MTF scheme is defined in Section III. The redundancy is evaluated theoretically for stationary ergodic sources, finite-order Markov sources, and binary sources in Sections IV, V, and VI, respectively.

II. The MTF Transform Algorithm

For nonnegative integers m and n ($n \leq m$), let $X_n^m = X_nX_{n+1}\cdots X_m$ represent a sequence of random variables such that each X_n takes values in a finite set $\mathcal{A} = \{a_1, a_2, \ldots, a_A\}$ with cardinality $A = |\mathcal{A}| < \infty$. We consider a stationary ergodic stochastic process $X = X_1^\infty$ as an information source. For simplicity, it is assumed that the probability of X_n satisfies $P_{X_n}(a) > 0$ for all $a \in \mathcal{A}$. A stochastic process $Y = Y_1^\infty$ is obtained from X by the MTF transform, which has a sequence $S = S_0^\infty$ of *state* S_n. Each Y_n takes values in integer alphabet $\mathcal{Y} = \{1, 2, \ldots, A\}$ and each S_n takes values in $\mathcal{S}$, which consists of all permutations of $(a_1, a_2, \ldots, a_A)$. S_0 is the initial state of the MTF transform.

The algorithm of the MTF transform is defined as a map from X_1^N to (Y_1^N, S_k, k).

Algorithm—MTF Transform:

1) Generate S_0 randomly with a probability distribution P_{S_0}, which is independent of X.
2) $n := 1$.
3) Suppose that $S_{n-1} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \in \mathcal{S}$. Then, find the symbol $\hat{a}_i$ such that $X_n = \hat{a}_i$.
4) $Y_n := i$.
5) $S_n := (\hat{a}_i, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{i-1}, \hat{a}_{i+1}, \ldots, \hat{a}_A)$, which is obtained by moving $\hat{a}_i$ to the front of S_{n-1}.

Manuscript received September 8, 2002; revised February 15, 2005. This work was supported in part by the JSPS Grants-in-Aid for Scientific Research 17360174. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Lausanne, Switzerland, June/July 2002.

M. Arimura is with the Department of System and Communication Engineering, Faculty of Engineering, Shonan Institute of Technology, 1-1-25, Tsujido-nishi-kaigan, Fujisawa-shi, Kanagawa 251-8511, Japan (e-mail: arimura@ieee.org).

H. Yamamoto is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashinoha, Kashiwa-shi, Chiba 277-8561, Japan (e-mail: Hirosuke@ieee.org).

Communicated by S. A. Savari, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2005.856941

6) If $n = N$, exit with output (Y_1^N, S_k, k), where $k \in [0, N]$ can be chosen arbitrarily. If $n < N$, then let $n := n + 1$ and go to Step 3.

We can easily show that X_1^N can be reproduced from (Y_1^N, S_k, k) for any $k \in [0, N]$.

When we fix the parameter k as $k = 0$, then we need not encode k. Moreover, if the initial state S_0 is fixed or generated by a deterministic algorithm, S_0 is also unnecessary to be encoded. In this case, the output becomes Y_1^N only. Each Y_n is usually encoded to a binary codeword by a universal code for the positive integers [1], [2] or the arithmetic code under the assumption that Y is memoryless. In the case of arithmetic coding, we can know exactly the probability of Y_n for each n if the probabilities of X and S_0 are known. But, practically, adaptive arithmetic coding based on memoryless empirical distribution of Y_1^{n-1} may be used.

In the following sections, we assume that the probability distributions of X, S_0, and hence, Y_n are known. Such analyses give the best performance in all possible coding of Y_n under the memoryless assumption for the MTF scheme.

III. Redundancy of the MTF Scheme

In the MTF transform algorithm, source symbol X_n is independent of the initial state S_0 for any $n \geq 1$. But Y_n depends on S_0 generally, and Y may not be stationary even if X is stationary. Hence, at a glance, it seems to be difficult to derive some useful properties of Y. However, if X is a stationary, especially finite-order Markov, source, we can derive useful bounds for the redundancy of the MTF scheme.

First we define the following two kinds of symbolwise redundancies for a given stationary source X.

Definition 1:

$$\rho_n^{(1)}(X) \stackrel{\text{def}}{=} H(Y_n) - H\left(X_n \,|\, X_1^{n-1}\right) \tag{1}$$

$$\rho_n^{(2)}(X) \stackrel{\text{def}}{=} H(Y_n \,|\, S_0) - H\left(X_n \,|\, X_1^{n-1}\right) \tag{2}$$

where $H(X_n \,|\, X_1^{n-1}) = H(X_1)$ for $n = 1$.

In the case that the initial state S_0 is fixed, $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ coincide with each other. On the other hand, if S_0 is not fixed, the following relations hold for any $n \geq 1$:

$$0 \leq \rho_n^{(2)}(X) \leq \rho_n^{(1)}(X) \leq \log_2 A \tag{3}$$

$$\rho_n^{(1)}(X) - \rho_n^{(2)}(X) = I(Y_n; S_0). \tag{4}$$

$\rho_n^{(2)}(X)$ corresponds to the case that each Y_n is encoded using the information of initial state S_0 while $\rho_n^{(1)}(X)$ corresponds to the case that any information of S_0 is not used in the encoding of Y_n.

The entropy rate of a stationary source X is given by $\lim_{N\to\infty} \frac{H(X_1^N)}{N}$. On the other hand, since Y is encoded as memoryless data, if the probability distribution of Y_1^N is known to the encoder and the decoder, the best average code length per one symbol for Y_1^N is given by

$$\frac{1}{N}\sum_{n=1}^{N} H(Y_n) \quad \text{or} \quad \frac{1}{N}\sum_{n=1}^{N} H(Y_n \,|\, S_0)$$

depending on whether the knowledge of S_0 is used in the encoding of Y_n. Hence, the asymptotic upper and lower bounds of redundancy of the MTF scheme can be given by

$$\left(\limsup_{N\to\infty} \tilde{\rho}_N^{(1)}(X), \liminf_{N\to\infty} \tilde{\rho}_N^{(1)}(X)\right)$$

or

$$\left(\limsup_{N\to\infty} \tilde{\rho}_N^{(2)}(X), \liminf_{N\to\infty} \tilde{\rho}_N^{(2)}(X)\right)$$

where $\tilde{\rho}_N^{(1)}(X)$ and $\tilde{\rho}_N^{(2)}(X)$ are defined by

$$\tilde{\rho}_N^{(1)}(X) \stackrel{\text{def}}{=} \frac{1}{N}\left\{\sum_{n=1}^{N} H(Y_n) - H(X_1^N)\right\} = \frac{1}{N}\sum_{n=1}^{N} \rho_n^{(1)}(X) \tag{5}$$

$$\tilde{\rho}_N^{(2)}(X) \stackrel{\text{def}}{=} \frac{1}{N}\left\{\sum_{n=1}^{N} H(Y_n \,|\, S_0) - H(X_1^N)\right\} = \frac{1}{N}\sum_{n=1}^{N} \rho_n^{(2)}(X). \tag{6}$$

Then, since $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ are bounded as (3) for any $n \geq 1$, we have the following inequalities from Lemma 7 in the Appendix, which is an extension of the Cesáro mean (cf., [13, Theorem 4.2.3]).

$$\begin{aligned}\liminf_{n\to\infty} \rho_n^{(1)}(X) &\leq \liminf_{N\to\infty} \tilde{\rho}_N^{(1)}(X) \\ &\leq \limsup_{N\to\infty} \tilde{\rho}_N^{(1)}(X) \leq \limsup_{n\to\infty} \rho_n^{(1)}(X)\end{aligned} \tag{7}$$

$$\begin{aligned}\liminf_{n\to\infty} \rho_n^{(2)}(X) &\leq \liminf_{N\to\infty} \tilde{\rho}_N^{(2)}(X) \\ &\leq \limsup_{N\to\infty} \tilde{\rho}_N^{(2)}(X) \leq \limsup_{n\to\infty} \rho_n^{(2)}(X).\end{aligned} \tag{8}$$

Hence, if $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ converge, then $\tilde{\rho}_N^{(1)}(X)$ and $\tilde{\rho}_N^{(2)}(X)$ also converge to the same values as $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$, respectively.

From the above reason, we evaluate $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ as the redundancies of the MTF scheme.

In the following sections, we treat the cases of stationary ergodic sources, finite-order Markov sources, and binary sources.

IV. Redundancy for Stationary Ergodic Sources

Assume that X is stationary ergodic. We define a function f, which represents a relation of (X_n, Y_n, S_{n-1}) in the MTF transform.

Definition 2: Functions $f : \mathcal{S}\times\mathcal{Y} \to \mathcal{A}$ and $f^{(k)} : \mathcal{S}^k \to \mathcal{A}^k$ are defined as follows:

$$\begin{aligned} f(s, y) &\stackrel{\text{def}}{=} \hat{a}_y \quad \text{if } s = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \\ f^{(k)}\left(s_{t+1}^{t+k}, 1\right) &\stackrel{\text{def}}{=} f(s_{t+1}, 1) f(s_{t+2}, 1) \cdots f(s_{t+k}, 1). \end{aligned} \tag{9}$$

We note that (X_n, Y_n, S_{n-1}) and (X_{n-1}, S_{n-1}) satisfy

$$X_n = f(S_{n-1}, Y_n) \tag{10}$$

$$X_{n-1} = f(S_{n-1}, 1) \tag{11}$$

respectively.

Next, we show that if all symbols of the alphabet occur in x_1^ℓ, then S_j for $j \geq \ell$ does not depend on the initial state S_0. First we define the set of such sequences.

Definition 3: For $x_1^n \in \mathcal{A}^n, n \geq 1, \mathcal{A}(x_1^n)$ denote the set of all distinct letters included in x_1^n. Moreover, for each n and $k, 1 \leq k \leq n$, define a set

$$\mathcal{G}_n^{(k)} \stackrel{\text{def}}{=} \left\{x_1^n : \mathcal{A}\left(x_1^{n-k+1}\right) = \mathcal{A}\right\}$$

which is the set of x_1^n such that all symbols of the alphabet occur in the first $n-k+1$ symbols of x_1^n.

We show the properties of the set $\mathcal{G}_n^{(k)}$ in the following lemma.

Lemma 1: There exists a deterministic function g such that for any prefix x_1^j of $x_1^n \in \mathcal{G}_n^{(k)}, n-k+1 \leq j \leq n$, $g(x_1^j)$ gives the state of time j independently of the initial state s_0. If X is ergodic, it is satisfied that for any fixed k

$$\lim_{n\to\infty} \mathsf{Pr}\left\{X_1^n \in \mathcal{G}_n^{(k)}\right\} = 1. \tag{12}$$

Proof: Assume that $x_1^n \in \mathcal{G}_n^{(k)}$. Then because of the definition of $\mathcal{G}_n^{(k)}$, $x_1^j, j \geq n-k+1$ includes all symbols of the alphabet $\mathcal{A}$. Assume that such a sequence x_1^j has state $s_j = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A)$ after x_1^j is converted by the MTF transform. Letting n_i denote the last occurrence time of $\hat{a}_i$ in x_1^j, each n_i must satisfy that $j = n_1 > n_2 > \cdots > n_A \geq 1$. The set of x_1^j with $s_j = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A)$, say, $\mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A)$, can be represented as

$$\begin{aligned}
&\mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \\
&\quad \stackrel{\text{def}}{=} \left(\bigcup_{i=1}^{A}\{\hat{a}_i\}\right)^{n_A-1} \times \{\hat{a}_A\} \\
&\qquad \times \left(\bigcup_{i=1}^{A-1}\{\hat{a}_i\}\right)^{n_{A-1}-n_A-1} \times \{\hat{a}_{A-1}\} \times \cdots \\
&\qquad \times \{\hat{a}_2, \hat{a}_1\}^{n_2-n_3-1} \times \{\hat{a}_2\} \times \{\hat{a}_1\}^{n_1-n_2} \subseteq \mathcal{G}_j^{(1)}
\end{aligned}$$

where $\mathcal{A} \times \mathcal{B}$ means the Cartesian product set of the sets $\mathcal{A}$ and $\mathcal{B}$, and

$$\left(\bigcup_{i=1}^{k}\{\hat{a}_i\}\right)^{\ell} \stackrel{\text{def}}{=} \underbrace{\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k\} \times \cdots \times \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k\}}_{\ell \text{ times}}$$

stands for the ℓth Cartesian product set of the set $\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k\}$. If $\ell = 0, \mathcal{B}^\ell$ denote the string of length zero. Clearly, if $x_1^j \in \mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A), s_j$ satisfies $s_j = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A)$ independently of s_0. Note that the next relations hold for any $j \geq n-k+1$.

$$\begin{aligned}
&\mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \cap \mathcal{G}_j^{(1)}(\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_A) = \emptyset, \\
&\qquad \text{if } (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \neq (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_A) \\
&\bigcup_{(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) \in \mathcal{S}} \mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A) = \mathcal{G}_j^{(1)}.
\end{aligned}$$

Therefore, the function $g(x_1^j)$ can be defined as follows:

$$g(x_1^j) \stackrel{\text{def}}{=} (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A), \qquad \text{if } x_1^j \in \mathcal{G}_j^{(1)}(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_A).$$

Thus, the former part of Lemma 1 holds.

The latter part of Lemma 1, i.e., (12) holds from the ergodicity of X. □

Next we define the concatenation of the function g.

Definition 4: For each n and k, $1 \leq k \leq n$, function $g^{(k)} : \mathcal{G}_n^{(k)} \to \mathcal{S}^k$ is defined as

$$g^{(k)}(x_1^n) \stackrel{\text{def}}{=} g\left(x_1^{n-k+1}\right) g\left(x_1^{n-k+2}\right) \cdots g(x_1^n) = s_{n-k+1}^n, \quad \text{for } x_1^n \in \mathcal{G}_n^{(k)} \tag{13}$$

where $g(x_1^j)$ is the function given in Lemma 1.

Lemma 2: For any $n \geq 2$, x_n and y_n are one-to-one if s_{n-1} is given.

Proof: Fix $s_{n-1} \in \mathcal{S}$ arbitrarily. Then from Definition 2 and Algorithm 1, a unique $i \in \mathcal{Y}$ for each $x_n \in \mathcal{A}$ satisfies that

$$f(s_{n-1}, i) = x_n.$$

On the other hand, from Definition 2 and the reverse MTF transform, a unique $a \in \mathcal{A}$ for each $y_n \in \mathcal{Y}$ satisfies that

$$f(s_{n-1}, y_n) = a. \qquad \square$$

Using these lemmas, we now derive some bounds of redundancies $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$, which are used in the following sections.

Theorem 1: For any stationary ergodic source X with finite alphabet $\mathcal{A}$, the next relations hold.

$$\lim_{n\to\infty}\left(\rho_n^{(1)}(X) - \rho_n^{(2)}(X)\right) = 0 \tag{14}$$

$$\liminf_{n\to\infty}\left(\rho_n^{(1)}(X) - I\left(Y_n; S_{n-k}^{n-1}\right)\right) \geq 0, \qquad \text{for any } k \geq 1. \tag{15}$$

Furthermore, it is satisfied for any integers $k \geq 1$ and $n > k$ that

$$\rho_n^{(1)}(X) \leq I\left(Y_n; S_{n-k}^{n-1}\right) + I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}\right) \tag{16}$$

$$\rho_n^{(2)}(X) = I\left(Y_n; Y_1^{n-1} \,|\, S_0\right). \tag{17}$$

From (14), $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ are equal asymptotically. This means that the information of the initial state S_0 cannot decrease the redundancy asymptotically. Expressions (15) and (16) give lower and upper bounds of $\rho_n^{(1)}(X)$, respectively.

Proof: First we derive (14). Probability $P_{Y_n|S_0}(y|s)$ can be described for any $y \in \mathcal{Y}$ and any $s \in \mathcal{S}$ with $\mathsf{Pr}\{S_0 = s\} > 0$ as follows:

$$\begin{aligned}
&P_{Y_n|S_0}(y|s) \\
&\quad = \sum_{x_1^{n-1} \in \mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y, X_1^{n-1} = x_1^{n-1} | S_0 = s\right\} \\
&\qquad + \sum_{x_1^{n-1} \notin \mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y, X_1^{n-1} = x_1^{n-1} | S_0 = s\right\}.
\end{aligned} \tag{18}$$

In case of $x_1^{n-1} \notin \mathcal{G}_{n-1}^{(1)}$, the summation is bounded above as

$$\begin{aligned}
&\sum_{x_1^{n-1}\notin\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y, X_1^{n-1} = x_1^{n-1} \,|\, S_0 = s\right\} \\
&\quad = \sum_{x_1^{n-1}\notin\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{X_1^{n-1} = x_1^{n-1}\right\} \\
&\qquad \cdot \frac{\mathsf{Pr}\left\{Y_n = y, S_0 = s \,|\, X_1^{n-1} = x_1^{n-1}\right\}}{\mathsf{Pr}\{S_0 = s\}} \\
&\quad \le \sum_{x_1^{n-1}\notin\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{X_1^{n-1} = x_1^{n-1}\right\} \frac{1}{\mathsf{Pr}\{S_0 = s\}} \\
&\quad = \frac{1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(1)}\right\}}{\mathsf{Pr}\{S_0 = s\}}. \qquad (19)
\end{aligned}$$

On the other hand, if $x_1^{n-1} \in \mathcal{G}_{n-1}^{(1)}$, S_{n-1} is uniquelly determined by $S_{n-1} = g(X_1^{n-1})$. This means that Y_n, X_1^{n-1}, and S_0 make a Markov chain $Y_n - X_1^{n-1} - S_0$. Hence, in this case, the summation becomes

$$\begin{aligned}
&\sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y, X_1^{n-1} = x_1^{n-1} \,|\, S_0 = s\right\} \\
&\quad = \sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y \,|\, X_1^{n-1} = x_1^{n-1}, S_0 = s\right\} \\
&\qquad \cdot \mathsf{Pr}\left\{X_1^{n-1} = x_1^{n-1} \,|\, S_0 = s\right\} \\
&\quad = \sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y \,|\, X_1^{n-1} = x_1^{n-1}\right\} \\
&\qquad \cdot \mathsf{Pr}\left\{X_1^{n-1} = x_1^{n-1}\right\} \\
&\quad = \sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(1)}} \mathsf{Pr}\left\{Y_n = y, X_1^{n-1} = x_1^{n-1}\right\} \\
&\quad \le \mathsf{Pr}\{Y_n = y\}. \qquad (20)
\end{aligned}$$

From (18)–(20), it holds that

$$P_{Y_n\,|\,S_0}(y\,|\,s) \le P_{Y_n}(y) + \frac{1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(1)}\right\}}{\mathsf{Pr}\{S_0 = s\}}.$$

Therefore, we have from (12) that

$$\begin{aligned}
&\limsup_{n\to\infty} \left\{P_{Y_n\,|\,S_0}(y\,|\,s) - P_{Y_n}(y)\right\} \\
&\quad \le \limsup_{n\to\infty} \frac{1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(1)}\right\}}{\mathsf{Pr}\{S_0 = s\}} \\
&\quad = 0. \qquad (21)
\end{aligned}$$

Since it holds that

$$\sum_y \{P_{Y_n\,|\,S_0}(y\,|\,s) - P_{Y_n}(y)\} = 0$$

for any s, (21) means that

$$\lim_{n\to\infty} \left\{P_{Y_n\,|\,S_0}(y\,|\,s) - P_{Y_n}(y)\right\} = 0$$

and therefore,

$$\lim_{n\to\infty} I(Y_n; S_0) = 0. \qquad (22)$$

From (4) and (22), (14) is established.

Next we derive (15). $H(X_n\,|\,X_1^{n-1})$ can be bounded above for any $k \ge 1$ as follows:

$$\begin{aligned}
&H(X_n\,|\,X_1^{n-1}) \\
&\quad = \sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(k)}} P_{X_1^{n-1}}\left(x_1^{n-1}\right) H\left(X_n\,|\,X_1^{n-1} = x_1^{n-1}\right) \\
&\qquad + \sum_{x_1^{n-1}\notin\mathcal{G}_{n-1}^{(k)}} P_{X_1^{n-1}}\left(x_1^{n-1}\right) H\left(X_n\,|\,X_1^{n-1} = x_1^{n-1}\right) \\
&\quad \le \sum_{x_1^{n-1}\in\mathcal{G}_{n-1}^{(k)}} P_{X_1^{n-1}}\left(x_1^{n-1}\right) H\left(X_n\,|\,X_1^{n-1} = x_1^{n-1}\right) \\
&\qquad + \sum_{x_1^{n-1}\notin\mathcal{G}_{n-1}^{(k)}} P_{X_1^{n-1}}\left(x_1^{n-1}\right) \log_2 A \\
&\quad \overset{(a)}{=} \sum_{s_{n-k}^{n-1}\in\mathcal{S}^k} \sum_{\substack{x_1^{n-1}\in\mathcal{G}_{n-1}^{(k)}: \\ g^{(k)}(x_1^{n-1})=s_{n-k}^{n-1}}} P_{X_1^{n-1}}\left(x_1^{n-1}\right) \\
&\qquad \cdot H\left(X_n\,|\,S_{n-k}^{n-1} = s_{n-k}^{n-1}, X_1^{n-k-1} = x_1^{n-k-1}\right) \\
&\qquad + \left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A \\
&\quad \le \sum_{s_{n-k}^{n-1}\in\mathcal{S}^k} \sum_{x_1^{n-k-1}\in\mathcal{X}^{n-k-1}} P_{X_1^{n-k-1}}\left(x_1^{n-k-1}\right) \\
&\qquad \cdot P_{S_{n-k}^{n-1}|X_1^{n-k-1}}(s_{n-k}^{n-1}|x_1^{n-k-1}) \\
&\qquad \cdot H\left(X_n\,|\,S_{n-k}^{n-1} = s_{n-k}^{n-1}, X_1^{n-k-1} = x_1^{n-k-1}\right) \\
&\qquad + \left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A \\
&\quad = H\left(X_n\,|\,S_{n-k}^{n-1}, X_1^{n-k-1}\right) \\
&\qquad + \left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A \\
&\quad \le H(X_n|S_{n-k}^{n-1}) + \left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A \\
&\quad \overset{(b)}{=} H\left(Y_n\,|\,S_{n-k}^{n-1}\right) + \left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A
\end{aligned}$$

where equality (a) holds because S_{n-k}^{n-1} is uniquely determined from X_1^{n-1} when $X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}$, and X_{n-k}^{n-1} is uniquely determined from S_{n-k}^{n-1} by (11). Furthermore, equality (b) is induced from Lemma 2.

Therefore, for any n, the next inequality is satisfied.

$$\begin{aligned}
&\rho_n^{(1)}(X) - I\left(Y_n; S_{n-k}^{n-1}\right) \\
&\quad = H(Y_n) - H\left(X_n\,|\,X_1^{n-1}\right) - I\left(Y_n; S_{n-k}^{n-1}\right) \\
&\quad = H\left(Y_n\,|\,S_{n-k}^{n-1}\right) - H\left(X_n\,|\,X_1^{n-1}\right) \\
&\quad \ge -\left(1 - \mathsf{Pr}\left\{X_1^{n-1} \in \mathcal{G}_{n-1}^{(k)}\right\}\right) \log_2 A. \qquad (23)
\end{aligned}$$

Thus, (15) is obtained from (12) and (23).

Equation (16) is derived from Lemma 2 as follows:

$$\begin{aligned}
\rho_n^{(1)}(X) &= H(Y_n) - H\left(X_n\,|\,X_{n-k}^{n-1}\right) \\
&\quad + H\left(X_n\,|\,X_{n-k}^{n-1}\right) - H\left(X_n\,|\,X_1^{n-1}\right) \\
&\overset{(c)}{\le} H(Y_n) - H\left(X_n\,|\,S_{n-k}^{n-1}\right) \\
&\quad + H\left(X_n\,|\,X_{n-k}^{n-1}\right) - H\left(X_n\,|\,X_1^{n-1}\right) \\
&\overset{(b)}{=} H(Y_n) - H\left(Y_n\,|\,S_{n-k}^{n-1}\right)
\end{aligned}$$

$$+H\left(X_n \,|\, X_{n-k}^{n-1}\right) - H\left(X_n \,|\, X_1^{n-1}\right)$$
$$= I\left(Y_n; S_{n-k}^{n-1}\right) + I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}\right)$$

where inequality (c) holds because X_{n-k}^{n-1} is uniquely determined from S_{n-k}^{n-1} by (11) and hence we have that

$$H\left(X_n \,|\, S_{n-k}^{n-1}\right) \le H\left(X_n \,|\, X_{n-k}^{n-1}\right).$$

Finally we derive (17). The initial state S_0 is independent of X, S_{n-1} is uniquely determined from (S_0, X_1^{n-1}) or (S_0, Y_1^{n-1}), and X_n and Y_n are one-to-one when (S_0, X_1^{n-1}) or (S_0, Y_1^{n-1}) are given. Therefore, we have that

$$\begin{aligned} H\left(X_n \mid X_1^{n-1}\right) &= H\left(X_n \,|\, X_1^{n-1}, S_0\right) \\ &= H\left(Y_n \,|\, X_1^{n-1}, S_0\right) \\ &= H\left(Y_n \,|\, Y_1^{n-1}, S_0\right) \end{aligned}$$

which means that

$$\begin{aligned} \rho_n^{(2)}(X) &= H(Y_n \,|\, S_0) - H\left(X_n \mid X_1^{n-1}\right) \\ &= H(Y_n \,|\, S_0) - H\left(Y_n \,|\, Y_1^{n-1}, S_0\right) \\ &= I\left(Y_n; Y_1^{n-1} \,|\, S_0\right). \end{aligned}$$

□

V. Redundancy for Markov Sources

In this section, we consider finite-order Markov sources on the finite alphabet $\mathcal{A}$.

Theorem 2: If X is a stationary ergodic Kth-order Markov source, it holds that for any $k \ge K \ge 1$

$$\lim_{n\to\infty}\left(\rho_n^{(1)}(X) - I\left(Y_n; S_{n-k}^{n-1}\right)\right) = 0 \tag{24}$$
$$\rho_n^{(2)}(X) = I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right), \qquad \text{for any } n > k. \tag{25}$$

Proof: If X is stationary and ergodic, (15) holds. Moreover, if X is the Kth-order Markov source

$$I(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}) = 0$$

holds in (16) for any $k \ge K$ and $n > k$. These relations induce (24).

On the other hand, (25) can be derived from (17) as follows:

$$\begin{aligned} \rho_n^{(2)}(X) &= I\left(Y_n; Y_1^{n-1} \,|\, S_0\right) \\ &\overset{(d)}{=} I\left(Y_n; Y_1^{n-1}, S_{n-k}^{n-1} \,|\, S_0\right) \\ &= I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(Y_n; Y_1^{n-1} \,|\, S_{n-k}^{n-1}, S_0\right) \\ &\overset{(e)}{=} I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-1} \,|\, S_{n-k}^{n-1}, S_0\right) \\ &\overset{(f)}{=} I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-1} \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) \end{aligned} \tag{26}$$

where equality (d) holds because S_{n-k}^{n-1} is uniquely determined from S_0 and Y_1^{n-1}, equality (e) holds from Lemma 2 and the fact that X_1^{n-1} and Y_1^{n-1} are one-to-one when S_0 is given, and equality (f) comes from the fact that X_{n-k}^{n-1} is uniquely determined from S_{n-k}^{n-1} by (11). Furthermore, it holds that for any $k \ge K$

$$\begin{aligned} &I\left(X_n; X_1^{n-1} \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) \\ &\quad = H\left(X_n \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) - H\left(X_n \,|\, S_{n-k}^{n-1}, X_1^{n-1}, S_0\right) \\ &\quad \overset{(g)}{=} H\left(X_n \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) - H\left(X_n \,|\, X_1^{n-1}, S_0\right) \\ &\quad \overset{(h)}{=} H\left(X_n \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) - H\left(X_n \,|\, X_1^{n-1}\right) \\ &\quad \le H\left(X_n \,|\, X_{n-k}^{n-1}\right) - H\left(X_n \,|\, X_1^{n-1}\right) \\ &\quad = I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}\right) \\ &\quad \overset{(i)}{=} 0 \end{aligned}$$

where (g) holds because S_{n-k}^{n-1} is determined from (S_0, X_1^{n-1}), (h) holds since X is independent of S_0, and (i) is from the assumption that X is the Kth-order Markov source. □

Next we give the necessary and sufficient condition such that the MTF-scheme attains the entropy rate of the source asymptotically. First we give three lemmas to evaluate

$$P_{Y_n \,|\, S_{n-k}^{n-1}}(y \,|\, s_{n-k}^{n-1}).$$

Lemma 3: If X is the Kth-order Markov source, it holds that for any $n > K$, $y \in \mathcal{Y}$, and $s_{n-K}^{n-1} \in \mathcal{S}^K$

$$\begin{aligned} &\mathsf{Pr}\left\{Y_n = y \,|\, S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\} \\ &\qquad = \mathsf{Pr}\left\{X_n = f(s_{n-1}, y) \,|\, X_{n-K}^{n-1} = f^{(K)}\left(s_{n-K}^{n-1}, 1\right)\right\}. \end{aligned}$$

Proof: Lemma 2 implies that for any $y \in \mathcal{Y}$ and $s_{n-K}^{n-1} \in \mathcal{S}^K$, there exists $\tilde{x} = f(s_{n-1}, y) \in \mathcal{A}$ that satisfies

$$\mathsf{Pr}\left\{Y_n = y \,|\, S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\} = \mathsf{Pr}\left\{X_n = \tilde{x} \,|\, S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\}.$$

The right-hand side of the preceding equation can be expressed, from (9) and (11), as follows:

$$\begin{aligned} &\mathsf{Pr}\left\{X_n = \tilde{x} \,|\, S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\} \\ &\quad = \mathsf{Pr}\left\{X_n = \tilde{x} \,|\, X_{n-K}^{n-1} = f^{(K)}(s_{n-K}^{n-1}, 1), \right. \\ &\qquad\qquad \left. S_{n-K} = s_{n-k}\right\} \\ &\quad = \mathsf{Pr}\left\{X_n = \tilde{x} \,|\, X_{n-K}^{n-1} = f^{(K)}\left(s_{n-K}^{n-1}, 1\right)\right\} \end{aligned}$$

where the last equality holds because S_{n-K} is determined from X_1^{n-K} and S_0, X is the Kth-order Markov source, and therefore, $X_n - X_{n-K}^{n-1} - X_1^{n-K} S_0 - S_{n-K}$ make a Markov chain. □

Lemma 4: Assume that the stationary source X with finite alphabet $\mathcal{A}\,(A = |\mathcal{A}|)$ is converted to Y by the MTF transform with state sequence S. Let $P^{(K)}$ and π denote the transition probability and a stationary distibution of X. If

$$I\left(Y_n; S_{n-K}^{n-1}\right) < \varepsilon \tag{27}$$

holds for any $\varepsilon > 0$ and $n \ge A + K$, then it is satisfied for all $s_{n-K}^{n-1} \in \mathcal{S}^K$ with $P_{S_{n-K}^{n-1}}(s_{n-K}^{n-1}) > 0$ that

$$D\left(P_{Y_n \,|\, S_{n-K}^{n-1}}\left(\cdot \,|\, s_{n-K}^{n-1}\right) \| P_{Y_n}(\cdot)\right) < \frac{\varepsilon}{\alpha\beta^{A-1}} \tag{28}$$

where α and β are defined by

$$\alpha \overset{\text{def}}{=} \min_{\substack{a_1^K \in \mathcal{A}^K: \\ \pi(a_1^K) > 0}} \pi\left(a_1^K\right)$$
$$\beta \overset{\text{def}}{=} \min_{\substack{a_1^K \in \mathcal{A}^K, a_{K+1} \in \mathcal{A}: \\ P^{(K)}(a_{K+1} \,|\, a_1^K) > 0}} P^{(K)}\left(a_{K+1} \,|\, a_1^K\right).$$

α and β are the minimum nonzero probability of the stationary distribution and the minimum nonzero transition probability, respectively.

Proof: We note that, letting $a_{n-K+k} = f(s_{n-K+k}, 1)$ for $1 \le k \le K-1$, the next relation holds.

$$\begin{aligned}&\text{Event}\left\{S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\}\\ &\quad = \text{Event}\{S_{n-K} = s_{n-K}, X_{n-K+1} = a_{n-K+1},\\ &\qquad\qquad \ldots, X_{n-1} = a_{n-1}\}. \quad (29)\end{aligned}$$

Moreover, for any

$$s_{n-K} = (a_{n-K}, a_{n-K-1}, \ldots, a_{n-K-A+1})$$

S_{n-K} satisfies $S_{n-K} = s_{n-K}$ if $X = X_1^\infty$ satisfies

$$X_{n-K-A+1} = a_{n-K-A+1}, \ldots, X_{n-K} = a_{n-K}$$

for $n > A$. This means that

$$\begin{aligned}&\text{Event}\{S_{n-K} = s_{n-K}\}\\ &\quad \supseteq \text{Event}\{X_{n-K-A+1} = a_{n-K-A+1}, \ldots X_{n-K} = a_{n-K}\}.\end{aligned} \quad (30)$$

Expressions (29) and (30) imply that

$$\begin{aligned}&\mathsf{Pr}\left\{S_{n-K}^{n-1} = s_{n-K}^{n-1}\right\}\\ &\quad = \mathsf{Pr}\{S_{n-K} = s_{n-K},\\ &\qquad X_{n-K+1} = a_{n-K+1}, \ldots, X_{n-1} = a_{n-1}\}\\ &\quad \ge \mathsf{Pr}\{X_{n-K-A+1} = a_{n-K-A+1}, \ldots, X_{n-K} = a_{n-K},\\ &\qquad X_{n-K+1} = a_{n-K+1}, \ldots, X_{n-1} = a_{n-1}\}\\ &\quad = \pi\left(a_{n-K-A+1}^{n-K}\right) \prod_{k=1}^{A-1} P^{(K)}\left(a_{n-k} \,|\, a_{n-k-K}^{n-k-1}\right)\\ &\quad \ge \alpha\beta^{A-1}. \end{aligned} \quad (31)$$

Hence, we have that for any $n > A$ and any s_{n-K}^{n-1} such that $P_{S_{n-K}^{n-1}}(s_{n-K}^{n-1}) > 0$

$$\begin{aligned}&I\left(Y_n; S_{n-K}^{n-1}\right)\\ &\quad = \sum_{\hat{s}_{n-K}^{n-1} \in \mathcal{S}^K} P_{S_{n-K}^{n-1}}\left(\hat{s}_{n-K}^{n-1}\right)\\ &\qquad \cdot D\left(P_{Y_n \,|\, S_{n-K}^{n-1}}\left(\cdot \,|\, \hat{s}_{n-K}^{n-1}\right) \,\|\, P_{Y_n}(\cdot)\right)\\ &\quad \ge \sum_{\hat{s}_{n-K}^{n-1} \in \mathcal{S}^K} \alpha\beta^{A-1} D\left(P_{Y_n \,|\, S_{n-K}^{n-1}}\left(\cdot \,|\, \hat{s}_{n-K}^{n-1}\right) \,\|\, P_{Y_n}(\cdot)\right)\\ &\quad \ge \alpha\beta^{A-1} D\left(P_{Y_n \,|\, S_{n-K}^{n-1}}\left(\cdot \,|\, s_{n-K}^{n-1}\right) \,\|\, P_{Y_n}(\cdot)\right).\end{aligned}$$

Finally, the inequality $I(Y_n; S_{n-K}^{n-1}) < \varepsilon$ implies for all $s_{n-K}^{n-1} \in \mathcal{S}^K$ with $P_{S_{n-K}^{n-1}}(s_{n-K}^{n-1}) > 0$ that

$$D\left(P_{Y_n \,|\, S_{n-K}^{n-1}}\left(\cdot \,|\, s_{n-K}^{n-1}\right) \| P_{Y_n}(\cdot)\right) < \frac{\varepsilon}{\alpha\beta^{A-1}}. \qquad \square$$

Lemma 5: Assume that X is the Kth-order Markov source with positive transition probabilities and S is the process obtained by applying the MTF transform to X. Then, there exists only one stationary distribution $\pi_S^{(K)}(s_1^K)$ that satisfies

$$\lim_{n\to\infty} \max_{s_1^K \in \mathcal{S}^K} \left| P_{S_{n-K}^{n-1}}\left(s_1^K\right) - \pi_S^{(K)}\left(s_1^K\right)\right| = 0. \quad (32)$$

Proof: S is clearly a finite-state Markov process which takes values in the set $\mathcal{S}$ with $|\mathcal{S}| = A!$. First, we show that S is irreducible. For any $s \in \mathcal{S}$ and $s' = (a_A, a_{A-1}, \ldots, a_1) \in \mathcal{S}$, assume that $S_i = s$ at the time i. Then $S_{i+A} = s'$ can be attained by $X_{i+1} = a_1, X_{i+2} = a_2, \ldots, X_{i+A} = a_A$. Since all the transition probabilities of X are assumed to be positive, there is a transition path of length A from s to s' with positive probability. This means that S is irreducible. Next, we show that S is aperiodic. For any $s, s' \in \mathcal{S}$, if there is a transition path from the state s to s' of length m, there is also a transition path of length $m+1$ because for any state $s' = (a_A, a_{A-1}, \ldots, a_1) \in \mathcal{S}$, it is possible to transit s' to s' with length one by $X_i = a_A$. Therefore, S is aperiodic.

Because S is a finite-state, irreducible, aperiodic Markov chain, its marginal distribution converges to the unique stationary distribution, i.e.,

$$\forall s_1^K \in \mathcal{S}^K, \quad \lim_{n\to\infty} \left| P_{S_{n-K}^{n-1}}\left(s_1^K\right) - \pi_S^{(K)}\left(s_1^K\right)\right| = 0.$$

Since s_1^K can take only finite states, (32) holds. $\square$

Using Lemmas 3–5, we next give the necessary and sufficient condition to attain the entropy rate of the K-th order Markov source asymptotically by the MTF scheme.

Theorem 3: Let X be a stationary ergodic Kth-order Markov source with finite alphabet $\mathcal{A}(A = |\mathcal{A}|)$, and let $P^{(K)} = \{P^{(K)}(x \,|\, x_1^K)\}$ denote the set of transition probabilities, which are positive for all $x \in \mathcal{A}$ and $x_1^K \in \mathcal{A}^K$. Then $\lim_{n\to\infty} \rho_n^{(1)}(X) = 0$ if and only if $K = 1$ and $P^{(K)}$ is given by

$$\begin{aligned}\tilde{P}_p^{(K)}\left(\tilde{x} \,|\, x_1^K\right) &= \tilde{P}_p^{(1)}(\tilde{x} \,|\, x_K)\\ &= \begin{cases} 1 - (A-1)p, & \text{if } \tilde{x} = x_K\\ p, & \text{otherwise}\end{cases}\end{aligned} \quad (33)$$

where p is a parameter satisfying $0 < p < 1/(1-A)$.

In the case of i.i.d. sources, the above condition can be simplified as follows.

Corollary 1: Let X be an i.i.d. source with finite alphabet $\mathcal{A}$. Then $\lim_{n\to\infty} \rho_n^{(1)}(X) = 0$ if and only if X_n is uniformly distributed over $\mathcal{A}$.

Proof of Theorem 3: First we derive (33) under the assumption

$$\lim_{n\to\infty} \rho_n^{(1)}(X) = 0. \quad (34)$$

Equations (24) and (34) means that

$$\lim_{n\to\infty} I(Y_n; S_{n-K}^{n-1}) = 0 \quad (35)$$

i.e., that for any $\varepsilon_1 > 0$ and sufficiently large n

$$I\left(Y_n; S_{n-K}^{n-1}\right) < \varepsilon_1. \quad (36)$$

From Lemma 4, (36) implies that for any $s_{n-K}^{n-1} \in \mathcal{S}^K$ with $P_{S_{n-K}^{n-1}}(s_{n-K}^{n-1}) > 0$

$$D\left(P_{Y_n \mid S_{n-K}^{n-1}}\left(\cdot \mid s_{n-K}^{n-1}\right) \| P_{Y_n}(\cdot)\right) < \frac{\varepsilon_1}{\alpha\beta^{A-1}}. \tag{37}$$

Then, because of Lemma 8 in the Appendix, it holds that for any $y \in \mathcal{Y}$ and any $s_{n-K}^{n-1} \in \mathcal{S}^K$

$$\begin{aligned}&\left| P_{Y_n \mid S_{n-K}^{n-1}}\left(y \mid s_{n-K}^{n-1}\right) \right.\\ &\quad \left. - \sum_{\tilde{s}^K \in \mathcal{S}^K} P_{S_{n-K}^{n-1}}(\tilde{s}^K) P_{Y_n \mid S_{n-K}^{n-1}}\left(y \mid \tilde{s}_1^K\right) \right|\\ &\quad = \left| P_{Y_n \mid S_{n-K}^{n-1}}\left(y \mid s_{n-K}^{n-1}\right) - P_{Y_n}(y) \right| < \sqrt{\frac{2\varepsilon_1}{\alpha\beta^{A-1}}}. \end{aligned} \tag{38}$$

From Lemma 3, $P_{Y_n \mid S_{n-K}^{n-1}}$ in (38) can be replaced with $P_{X_n \mid X_{n-K}^{n-1}}$, and $P_{X_n \mid X_{n-K}^{n-1}}$ can be represented by $P^{(K)}$ since X is stationary. Hence, we obtain that

$$\begin{aligned}&\left| P^{(K)}\left(f(s_{n-1}, y) | f^{(K)}\left(s_{n-K}^{n-1}, 1\right)\right) \right.\\ &\quad \left. - \sum_{\tilde{s}^K \in \mathcal{S}^K} P_{S_{n-K}^{n-1}}(\tilde{s}^K) \cdot P^{(K)}(f(\tilde{s}_K, y) | f^{(K)}(\tilde{s}^K, 1)) \right|\\ &\quad < \sqrt{\frac{2\varepsilon_1}{\alpha\beta^{A-1}}}. \end{aligned} \tag{39}$$

On the other hand, from Lemma 5, there exists the unique stationary distribution $\pi_S^{(K)}(\cdot)$ for S. Hence, from (32), it holds for any $\tilde{s}^K \in \mathcal{S}^K$, any $\varepsilon_2 > 0$, and sufficiently large n that

$$\left| P_{S_{n-K}^{n-1}}(\tilde{s}^K) - \pi_S^{(K)}(\tilde{s}^K) \right| < \varepsilon_2.$$

Then if we define $C(y)$ as

$$C(y) \stackrel{\text{def}}{=} \sum_{\tilde{s}^K \in \mathcal{S}^K} \pi_S^{(K)}(\tilde{s}^K) \cdot P^{(K)}(f(\tilde{s}_K, y) | f^{(K)}(\tilde{s}^K, 1))$$

the following holds;

$$\begin{aligned}&\left| \sum_{\tilde{s}^K \in \mathcal{S}^K} P_{S_{n-K}^{n-1}}(\tilde{s}^K) \right.\\ &\quad \left. \cdot P^{(K)}(f(\tilde{s}_K, y) | f^{(K)}(\tilde{s}^K, 1)) - C(y) \right|\\ &\quad \le \sum_{\tilde{s}^K \in \mathcal{S}^K} \left| P_{S_{n-K}^{n-1}}(\tilde{s}^K) - \pi_S^{(K)}(\tilde{s}^K) \right|\\ &\qquad \cdot P^{(K)}(f(\tilde{s}_K, y) | f^{(K)}(\tilde{s}^K, 1))\\ &\quad < \sum_{\tilde{s}^K \in \mathcal{S}^K} \varepsilon_2 P^{(K)}\left(f(\tilde{s}_K, y) | f^{(K)}\left(\tilde{s}_1^K, 1\right)\right)\\ &\quad \le \sum_{\tilde{s}^K \in \mathcal{S}^K} \varepsilon_2 \le \varepsilon_2 (A!)^K. \end{aligned} \tag{40}$$

Combining (39) and (40), we obtain that

$$\begin{aligned}&\left| P^{(K)}\left(f(s_{n-1}, y) \mid f^{(K)}\left(s_{n-K}^{n-1}, 1\right)\right) - C(y) \right|\\ &\qquad < \sqrt{\frac{2\varepsilon_1}{\alpha\beta^{A-1}}} + \varepsilon_2 (A!)^K. \end{aligned} \tag{41}$$

Letting $x_n = f(s_{n-1}, y)$ and $x_{n-K}^{n-1} = f^{(K)}(s_{n-K}^{n-1}, 1)$, since (41) holds for any $s_{n-K}^{n-1} \in \mathcal{S}^K$ and $y \in \mathcal{Y}$, it is satisfied for any $\varepsilon_1 > 0, \varepsilon_2 > 0$, and sufficiently large n that

$$\left| P^{(K)}\left(x_n \mid x_{n-K}^{n-1}\right) - C(y) \right| < \sqrt{\frac{2\varepsilon_1}{\alpha\beta^{A-1}}} + \varepsilon_2 (A!)^K. \tag{42}$$

This means that for $x_n = f(s_{n-1}, y)$

$$\lim_{n\to\infty} P^{(K)}\left(x_n \mid x_{n-K}^{n-1}\right) = C(y). \tag{43}$$

Futhermore, since X is stationary, the left-hand side of (43) does not depend on time n. Therefore, $P^{(K)}$ must satisfy that for any $x_1^K \in \mathcal{A}^K, s_K \in \mathcal{S}, y \in \mathcal{Y}$ with $\tilde{x} = f(s_K, y) \in \mathcal{A}$

$$P^{(K)}\left(\tilde{x} \mid x_1^K\right) = C(y). \tag{44}$$

With the preceding formula, we consider two cases of $y = 1$ and $y \ge 2$. We note from the assumption that $P(\tilde{x} \mid x_1^K) > 0$ for all $\tilde{x} \in \mathcal{A}$ and all $x_1^K \in \mathcal{A}^K$.

1) In the case of $y = 1$, we have $\tilde{x} = x_K$ from (10) and (11). Hence, (44) means that $P^{(K)}(\tilde{x} \mid x_1^K)$ is constant for all x_1^K and all $\tilde{x}$ such that $\tilde{x} = x_K$.
2) In the case of $y \ge 2$, $\tilde{x}$ and x_K must be different from each other. Consider a case such that $x_1^K = aa\cdots a$ and $\tilde{x} \ne a$. Then, if s_0 has a and $\tilde{x}$ as the first and yth elements, respectively, s_K also has a and $\tilde{x}$ as the first and yth elements, respectively. Hence, by selecting s_0 adequately, every y except 1 can satisfy $\tilde{x} = f(s_K, y)$. This means from (44) that $P^{(K)}(\tilde{x} \mid aa\cdots a) = C(y)$ for any fixed $\tilde{x}$, a, and all $y = 2, 3, \cdots, A$, that is, $C(y)$ is constant for $y = 2, 3, \cdots, A$. Therefore, we conclude from (44) that $P^{(K)}(\tilde{x} \mid x_1^K)$ must be constant for all x_1^K and all $\tilde{x}$ such that $\tilde{x} \ne x_K$.

Combining the results of these two cases, $P^{(K)}(\tilde{x} \mid x_1^K)$ depends only on $\tilde{x}$ and x_K but not x_1^{K-1}. Therefore, X becomes the first-order Markov source with $P^{(K)}(\tilde{x} \mid x_1^K) = P^{(1)}(\tilde{x} \mid x_K)$ satisfying that

1) $P^{(1)}(x \mid x)$ is constant for all $x \in \mathcal{A}$,
2) $P^{(1)}(\tilde{x} \mid x)$ is constant for all $x, \tilde{x} (\ne x) \in \mathcal{A}$.

Such distribution $P^{(1)}(\tilde{x} \mid x)$ is given by (33).

Conversely, if $P^{(K)}$ is given by (33), it can be easily checked that $\lim_{n\to\infty} I(Y_n; S_{n-K}^{n-1}) = 0$ for any K, and therefore, $\lim_{n\to\infty} \rho_n^{(1)}(X) = 0$. □

In Theorem 3, we assume that transition probabilities $P^{(K)}(\tilde{x} \mid x_1^K)$ for all $\tilde{x} \in \mathcal{A}$ and all $x_1^K \in \mathcal{A}^K$ are positive. If there exists a pair $(\tilde{x}, x_1^K)$ such that $P^{(K)}(\tilde{x} \mid x_1^K) = 0$, we cannot conclude that (33) is necessary for $\lim_{n\to\infty} \rho_n^{(1)}(X) = 0$ because probability $\mathsf{Pr}\{S_{n-K}^{n-1} = s_{n-K}^{n-1}\}$ cannot be bounded below for all s_{n-K}^{n-1} as (31). In this case, we must exclude state s_{n-K}^{n-1} with

$$\lim_{n\to\infty} \mathsf{Pr}\{g^{(K)}(X_1^n) = s_{n-K}^{n-1}\} = 0$$

in the MTF transform. But, such treatment is burdensome.

In the case that the transition probability is different from (33), the redundancy can be bounded as follows.

Theorem 4: Let X be a stationary ergodic Kth-order Markov source with finite alphabet $\mathcal{A}$ and transition probability $P^{(K)}$.

Then the redundancy $\rho_n^{(1)}(X)$ is bounded above asymptotically by

$$\begin{aligned}&\limsup_{n\to\infty}\rho_n^{(1)}(X)\\&\quad\le \inf_{0<p<\frac{1}{A-1}} D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|\,X_1^K\right) \quad (45)\\&\quad\le \inf_{0<p<\frac{1}{A-1}}\max_{x_1^K\in\mathcal{A}^K} D\left(P^{(K)}\left(\cdot\,|\,x_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,x_K)\right) \quad (46)\end{aligned}$$

where the right-hand side in (45) is the conditional divergence defined by

$$\begin{aligned}&D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|X_1^K\right)\\&\qquad=\sum_{x_1^K\in\mathcal{A}^K}\pi\left(x_1^K\right)D\left(P^{(K)}\left(\cdot\,|\,x_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,x_K)\right).\end{aligned}$$

Theorem 4 means that if the transition probability of the Kth-order Markov source is close to the one given by (33) in the sense of divergence, the redundancy is small.

Proof: It holds for any $P_{Y_n}(\cdot),\tilde{P}_{Y_n}(\cdot)$, and $P_{Y_n\,|\,S_{n-K}^{n-1}}(\cdot\,|\,\cdot)$ that

$$\begin{aligned}&\log_2\frac{1}{P_{Y_n}(Y_n)}+\log_2\frac{P_{Y_n}(Y_n)}{\tilde{P}_{Y_n}(Y_n)}\\&\quad=\log_2\frac{1}{P_{Y_n\,|\,S_{n-K}^{n-1}}\left(Y_n\,|\,S_{n-K}^{n-1}\right)}\\&\qquad+\log_2\frac{P_{Y_n\,|\,S_{n-K}^{n-1}}\left(Y_n\,|\,S_{n-K}^{n-1}\right)}{\tilde{P}_{Y_n}(Y_n)}.\end{aligned}$$

Taking the average of both sides with respect to (Y_n,S_{n-K}^{n-1}), we have

$$\begin{aligned}H(Y_n)+D\left(P_{Y_n}(\cdot)\middle\|\tilde{P}_{Y_n}(\cdot)\right)&=H\left(Y_n\,|\,S_{n-K}^{n-1}\right)\\&\quad+D\left(P_{Y_n\,|\,S_{n-K}^{n-1}}\left(\cdot\,|\,S_{n-K}^{n-1}\right)\middle\|\tilde{P}_{Y_n}(\cdot)\middle|S_{n-K}^{n-1}\right).\end{aligned}$$

Using this equation and the nonnegativity of divergence, $I(Y_n;S_{n-K}^{n-1})$ is upper-bounded as follows:

$$\begin{aligned}&I\left(Y_n;S_{n-K}^{n-1}\right)\\&\quad=H(Y_n)-H\left(Y_n\,|\,S_{n-K}^{n-1}\right)\\&\quad=D\left(P_{Y_n\,|\,S_{n-K}^{n-1}}\left(\cdot\,|\,S_{n-K}^{n-1}\right)\middle\|\tilde{P}_{Y_n}(\cdot)\middle|\,S_{n-K}^{n-1}\right)\\&\qquad-D\left(P_{Y_n}(\cdot)\middle\|\tilde{P}_{Y_n}(\cdot)\right)\\&\quad\le D\left(P_{Y_n\,|\,S_{n-K}^{n-1}}\left(\cdot\,|\,S_{n-K}^{n-1}\right)\middle\|\tilde{P}_{Y_n}(\cdot)\middle|\,S_{n-K}^{n-1}\right). \quad (47)\end{aligned}$$

Now, set the probability distribution $\tilde{P}_{Y_n\,|\,S_{n-K}^{n-1}}(\cdot\,|\,\cdot)$ as

$$\begin{aligned}\tilde{P}_{Y_n\,|\,S_{n-K}^{n-1}}\left(y\,|\,s_{n-K}^{n-1}\right)&=\tilde{P}_{Y_n}(y)\\&\stackrel{\text{def}}{=}\begin{cases}1-(A-1)p, & \text{if } y=1\\ p, & \text{otherwise}\end{cases}\end{aligned}$$

which is independent of the state $s_{n-K}^{n-1}\in\mathcal{S}^K$ and parameterized by p. Then if we rewrite this probability $\tilde{P}_{Y_n}=\tilde{P}_{Y_n\,|\,S_{n-K}^{n-1}}$ to a conditional probability of X_n given X_{n-1} by Lemma 3, it becomes $\tilde{P}_p^{(K)}$ defined by (33). On the other hand, $P_{Y_n\,|\,S_{n-K}^{n-1}}$ can be rewritten to $P^{(K)}$ by Lemma 3.

By these replacements, (47) can be represented as follows:

$$\begin{aligned}&I\left(Y_n;S_{n-K}^{n-1}\right)\\&\quad\le D\left(P^{(K)}\left(\cdot\,|\,X_{n-K}^{n-1}\right)\middle\|\tilde{P}_p^{(K)}\left(\cdot\,|\,X_{n-K}^{n-1}\right)\middle|\,X_{n-K}^{n-1}\right)\\&\quad=D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|\,X_1^K\right).\end{aligned}$$

Hence, we obtain from Theorem 2 that

$$\begin{aligned}&\limsup_{n\to\infty}\left(\rho_n^{(1)}(X)-D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|\,X_1^K\right)\right)\\&\qquad\le 0\end{aligned}$$

which means

$$\begin{aligned}&\limsup_{n\to\infty}\rho_n^{(1)}(X)\\&\quad\le D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|\,X_1^K\right) \quad (48)\\&\quad=\sum_{x_1^K\in\mathcal{A}^K}\pi\left(x_1^K\right)D\left(P^{(K)}\left(\cdot\,|\,x_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,x_K)\right)\\&\quad\le\max_{x_1^K\in\mathcal{A}^K}D\left(P^{(K)}\left(\cdot\,|\,x_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,x_K)\right). \quad (49)\end{aligned}$$

Finally, since (48) and (49) hold for any p, the bounds can be taken the infimum by p as follows:

$$\begin{aligned}&\limsup_{n\to\infty}\rho_n^{(1)}(X)\\&\quad\le\inf_{0<p<\frac{1}{A-1}}D\left(P^{(K)}\left(\cdot\,|\,X_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,X_K)\middle|\,X_1^K\right)\\&\quad\le\inf_{0<p<\frac{1}{A-1}}\max_{x_1^K\in\mathcal{A}^K}D\left(P^{(K)}\left(\cdot\,|\,x_1^K\right)\middle\|\tilde{P}_p^{(1)}(\cdot\,|\,x_K)\right).\end{aligned}$$

□

VI. Redundancy for Binary Sources

In this section, we consider the case that the source X is binary.

Lemma 6: If the source X is binary, then s_n and x_n are one-to-one for any $n\ge 1$.

Proof: Letting the alphabet be $\mathcal{A}=\{a,b\}$, the next relations hold.

$$\begin{aligned}x_n=a&\iff s_n=(a,b)\\x_n=b&\iff s_n=(b,a).\end{aligned}$$

□

In the binary case, $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ can be evaluated exactly for any n as shown by the next theorem.

Theorem 5: If X is a binary source, the next equations hold for any $k\ge 1$ and $n>k$

$$\rho_n^{(1)}(X)=I\left(Y_n;S_{n-k}^{n-1}\right)+I\left(X_n;X_1^{n-k-1}\,|\,X_{n-k}^{n-1}\right) \quad (50)$$

$$\rho_n^{(2)}(X)=I\left(Y_n;S_{n-k}^{n-1}\,|\,S_0\right)+I\left(X_n;X_1^{n-k-1}\,|\,X_{n-k}^{n-1}\right). \quad (51)$$

Proof: For a given binary source, $\rho_n^{(1)}(X)$ is evaluated as follows:

$$\begin{aligned}\rho_n^{(1)}(X)&=H(Y_n)-H\left(X_n\,|\,X_{n-k}^{n-1}\right)\\&\quad+H\left(X_n\,|\,X_{n-k}^{n-1}\right)-H\left(X_n\,|\,X_1^{n-1}\right)\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{=} H(Y_n) - H\left(X_n \,|\, S_{n-k}^{n-1}\right) \\
&\quad + H\left(X_n \,|\, X_{n-k}^{n-1}\right) - H\left(X_n \,|\, X_1^{n-1}\right) \\
&\overset{(b)}{=} H(Y_n) - H\left(Y_n \,|\, S_{n-k}^{n-1}\right) \\
&\quad + H\left(X_n \,|\, X_{n-k}^{n-1}\right) - H\left(X_n \,|\, X_1^{n-1}\right) \\
&= I\left(Y_n; S_{n-k}^{n-1}\right) + I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}\right)
\end{aligned}$$

where equalities (a) and (b) hold from Lemmas 6 and 2, respectively.

Similarly, $\rho_n^{(2)}(X)$ is evaluated from (26) as follows:

$$\begin{aligned}
\rho_n^{(2)}(X) &= I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-1} \,|\, S_{n-k}^{n-1}, X_{n-k}^{n-1}, S_0\right) \\
&\overset{(a)}{=} I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-1} \,|\, X_{n-k}^{n-1}, S_0\right) \\
&= I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}, S_0\right) \\
&\overset{(c)}{=} I\left(Y_n; S_{n-k}^{n-1} \,|\, S_0\right) + I\left(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}\right)
\end{aligned}$$

where equality (c) comes from the fact that S_0 is independent of X. □

If X is the Kth-order binary Markov sources, the next theorem can easily be obtained from Theorem 5.

Theorem 6: Let X be the Kth-order binary Markov source. If $\rho_n^{(1)}(X) = 0$ or $\rho_n^{(2)}(X) = 0$ for some $n \geq 2$, then the Markov order must be $K = 1$. In other words, in the case of $K \geq 2$, $\rho_n^{(1)}(X)$, and $\rho_n^{(2)}(X)$ are strictly positive for any $n \geq 2$.

Proof: Since any binary source satisfies (50), $\rho_n^{(1)}(X) = 0$ if and only if both of

$$I(Y_n; S_{n-k}^{n-1}) = 0 \quad \text{and} \quad I(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}) = 0$$

are satisfied. We note that the condition

$$I(X_n; X_1^{n-k-1} \,|\, X_{n-k}^{n-1}) = 0$$

for any $k \geq 1$ means that $K = 1$.

We have the similar argument for $\rho_n^{(2)}(X)$ from (51). □

This theorem means that for any n, the MTF scheme can encode only some of the first order binary Markov sources at the entropy rate. Moreover, it cannot attain the entropy rate if the order of the binary Markov source is strictly greater than one.

In the case of the first-order binary Markov sources, we can obtain simple upper and lower bounds of the redundancy from Theorem 4.

Theorem 7: Let X be a stationary ergodic first-order Markov source with binary alphabet $\mathcal{A} = \{a, b\}$. Then $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ satisfy that for $n \geq 2$

$$\rho_n^{(1)}(X) = \rho_n^{(2)}(X) = I(Y_n; X_{n-1}). \tag{52}$$

Letting $p_a = P_{X_n \,|\, X_{n-1}}(b \,|\, a), p_b = P_{X_n \,|\, X_{n-1}}(a \,|\, b)$, and $\lambda = \pi(a)$, where π is a stationary distribution of X, then $\rho_n^{(1)}(X)$ and $\rho_n^{(2)}(X)$ are bounded for any $n \geq 2$ as follows:

$$\begin{aligned}
\rho_n^{(1)}(X) &= \rho_n^{(2)}(X) \\
&\leq \lambda(1-\lambda)\{d(p_a \,\|\, p_b) + d(p_b \,\|\, p_a)\} \\
&\leq \frac{1}{2}\max\{d(p_a \,\|\, p_b), d(p_b \,\|\, p_a)\} \qquad (53) \\
\rho_n^{(1)}(X) &= \rho_n^{(2)}(X) \\
&\geq \min\{d(p_a \| \lambda p_a + (1-\lambda)p_b), \\
&\qquad d(p_b \,\|\, \lambda p_a + (1-\lambda)p_b)\} \qquad (54)
\end{aligned}$$

where $d(p \,\|\, q)$ is the binary divergence defined by

$$d(p \,\|\, q) \overset{\text{def}}{=} p \log_2 \frac{p}{q} + (1-p)\log_2 \frac{1-p}{1-q}.$$

Proof: When X is the first-order Markov source, $I(X_n; X_1^{n-2} \,|\, X_{n-1})$ is equal to zero. Hence, (50) and (51) with $k = 1$ imply

$$\begin{aligned}
\rho_n^{(1)}(X) &= I(Y_n; S_{n-1}) \\
\rho_n^{(2)}(X) &= I(Y_n; S_{n-1} \,|\, S_0).
\end{aligned}$$

Moreover, from Lemma 6, $\rho_n^{(1)}(X)$ can be represented as

$$\rho_n^{(1)}(X) = I(Y_n; X_{n-1}). \tag{55}$$

On the other hand, $\rho_n^{(2)}(X)$ can be evaluated as follows:

$$\begin{aligned}
\rho_n^{(2)}(X) &= I(Y_n; S_{n-1} \,|\, S_0) \\
&\overset{(a)}{=} I(Y_n; X_{n-1} \,|\, S_0) \\
&= H(X_{n-1} \,|\, S_0) - H(X_{n-1} \,|\, Y_n, S_0) \\
&\overset{(d)}{\geq} H(X_{n-1}) - H(X_{n-1} \,|\, Y_n) \\
&= I(Y_n; X_{n-1}) \qquad (56)
\end{aligned}$$

where inequality (d) comes from the fact that $H(X_{n-1} \,|\, S_0) = H(X_{n-1})$ because S_0 is independent of X and that $H(X_{n-1} \,|\, Y_n, S_0) \leq H(X_{n-1} \,|\, Y_n)$.

From (3), (55), and (56), we conclude that $\rho_n^{(1)}(X) = \rho_n^{(2)}(X)$ for any $n \geq 2$.

Next we derive (53). In the binary case, X_n and Y_n are one-to-one when X_{n-1} is given. Hence, we have that

$$\begin{aligned}
P_{Y_n \,|\, X_{n-1}}(1 \,|\, a) &= P_{X_n \,|\, X_{n-1}}(a \,|\, a) = 1 - p_a \\
P_{Y_n \,|\, X_{n-1}}(2 \,|\, a) &= P_{X_n \,|\, X_{n-1}}(b \,|\, a) = p_a \\
P_{Y_n \,|\, X_{n-1}}(1 \,|\, b) &= P_{X_n \,|\, X_{n-1}}(b \,|\, b) = 1 - p_b \\
P_{Y_n \,|\, X_{n-1}}(2 \,|\, b) &= P_{X_n \,|\, X_{n-1}}(a \,|\, b) = p_b.
\end{aligned}$$

Furthermore, it holds that

$$\begin{aligned}
P_{Y_n}(1) &= \lambda P_{Y_n \,|\, X_{n-1}}(1 \,|\, a) + (1-\lambda) P_{Y_n \,|\, X_{n-1}}(1 \,|\, b) \\
&= \lambda(1-p_a) + (1-\lambda)(1-p_b) \\
P_{Y_n}(2) &= \lambda P_{Y_n \,|\, X_{n-1}}(2 \,|\, a) + (1-\lambda) P_{Y_n \,|\, X_{n-1}}(2 \,|\, b) \\
&= \lambda p_a + (1-\lambda) p_b.
\end{aligned}$$

Hence, we can bound $\rho_n^{(1)}(X)$ above as follows:

$$\begin{aligned}
\rho_n^{(1)}(X) &= I(Y_n; X_{n-1}) \\
&= P_{X_{n-1}}(a) D\left(P_{Y_n \,|\, X_{n-1}}(\cdot \,|\, a) \,\|\, P_{Y_n}(\cdot)\right) \\
&\quad + P_{X_{n-1}}(b) D\left(P_{Y_n \,|\, X_{n-1}}(\cdot \,|\, b) \,\|\, P_{Y_n}(\cdot)\right) \\
&= \lambda d(p_a \,\|\, \lambda p_a + (1-\lambda)p_b) \\
&\quad + (1-\lambda) d(p_b \,\|\, \lambda p_a + (1-\lambda)p_b) \\
&\leq \lambda(\lambda d(p_a \,\|\, p_a) + (1-\lambda) d(p_a \,\|\, p_b)) \\
&\quad + (1-\lambda)(\lambda d(p_b \,\|\, p_a) + (1-\lambda) d(p_b \,\|\, p_b)) \\
&= \lambda(1-\lambda)(d(p_a \,\|\, p_b) + d(p_b \,\|\, p_a)) \\
&\leq \frac{1}{4}(d(p_a \,\|\, p_b) + d(p_b \,\|\, p_a)) \\
&\leq \frac{1}{2}\max\{d(p_a \,\|\, p_b), d(p_b \,\|\, p_a)\} \qquad (57)
\end{aligned}$$

where the first inequality holds because of the convexity of divergence.

On the other hand, $\rho_n^{(1)}(X)$ is bounded below from the third equality in (57) as the following:

$$\begin{aligned}\rho_n^{(1)}(X) &= \lambda d(p_a \,\|\, \lambda p_a + (1-\lambda)p_b) \\ &\quad + (1-\lambda) d(p_b \,\|\, \lambda p_a + (1-\lambda)p_b) \\ &\geq \min\{d(p_a \,\|\, \lambda p_a + (1-\lambda)p_b), \\ &\qquad d(p_b \,\|\, \lambda p_a + (1-\lambda)p_b)\}\end{aligned}$$

where the inequality holds because $\rho_n^{(1)}(X)$ is a weighted average of the two values

$$d(p_a \,\|\, \lambda p_a + (1-\lambda)p_b) \quad \text{and} \quad d(p_b \| \lambda p_a + (1-\lambda)p_b)$$

for $0 < \lambda < 1$. □

We note from (53) that for the first-order binary Markov source, the redundancy of the MTF scheme becomes small as the transition probability becomes close to the symmetric distribution with $p_a = p_b$.

From Theorem 4, we can derive another upper bound of the redundancy.

Theorem 8: Let X be the first-order binary Markov source with $p_a = P_{X_n \,|\, X_{n-1}}(b \,|\, a)$ and $p_b = P_{X_n \,|\, X_{n-1}}(a \,|\, b)$, $0 < p_a, p_b < 1$. Then the redundancy is upper-bounded as

$$\limsup_{n\to\infty} \rho_n^{(1)}(X) \leq d(p_a \,\|\, \hat{p}) = d(p_b \,\|\, \hat{p}) \tag{58}$$

where $\hat{p}$ is given by

$$\hat{p} = \left(\exp_2 \frac{h(p_a) - h(p_b)}{p_a - p_b} + 1 \right)^{-1} \tag{59}$$

and $h(z)$ is the binary entropy function defined by

$$h(z) \stackrel{\text{def}}{=} z \log_2 \frac{1}{z} + (1-z) \log_2 \frac{1}{1-z}.$$

Proof: In the binary case, $\tilde{P}_p^{(1)}(\cdot \,|\, a)$ defined in (33) can represent any binary symmetric probability distribution $\tilde{P}_p^{(1)}(b \,|\, a) = \tilde{P}_p^{(1)}(a \,|\, b) = p$. Hence, the following upper bound is derived from (46) and Lemma 9 in the Appendix:

$$\begin{aligned}\limsup_{n\to\infty} \rho_n^{(1)}(X) &\leq \min_p \max_{x\in\mathcal{A}} D\left(P^{(1)}(\cdot \,|\, x) \,\middle\|\, \tilde{P}_p^{(1)}(\cdot \,|\, x) \right) \\ &= \min_p \max\{d(p_a \,\|\, p), d(p_b \,\|\, p)\} \\ &= \min_p \max\{g_p(p_a) - h(p_a), g_p(p_b) - h(p_b)\}\end{aligned} \tag{60}$$

where $g_p(q)$ is defined by

$$g_p(q) = h(p) + (-\log_2 p + \log_2(1-p))(q-p). \tag{61}$$

Since $h(q)$ is concave and $g_p(q)$ is the tangent line of $h(q)$ at $q = p$, $g_p(q) - h(q)$ monotonically increases as $|p - q|$ becomes large. Hence, moving p from p_a to p_b, $g_p(p_a) - h(p_a)$ monotonically increases from zero and $g_p(p_b) - h(p_b)$ monotonically decreases to zero. Therefore, (60) can be minimized for given p_a and p_b by selecting p that satisfies $g_p(p_a) - h(p_a) = g_p(p_b) - h(p_b)$. Such optimal $\hat{p}$ is given by (59), and the upper bound of (60) becomes

$$\min_p \max_{x\in\mathcal{A}} D\left(P^{(1)}(\cdot \,|\, x) \,\middle\|\, \tilde{P}_p^{(1)}(\cdot \,|\, x) \right) = d(p_a \,\|\, \hat{p}) = d(p_b \,\|\, \hat{p}).$$

□

VII. Concluding Remarks

In this paper, we have clarified what class of sources can be compressed efficiently by the MTF scheme. We showed that the MTF scheme cannot attain the entropy rate of any Kth-order Markov source for $K \geq 2$, but it can attain the entropy rate if the source is the first-order Markov source given by (33). Furthermore, the closer the source distribution becomes to (33), the smaller the redundancy is.

We note that the first-order Markov source with transition probability given by (33) is a doubly stochastic process. Since X is assumed to be ergodic, its symbol-wise stationary distribution is the uniform distribution (cf., [13, Problem 1 in Ch. 4]). Therefore, if we encode the source X as an i.i.d. source, X cannot be compressed at all. However, we note from Theorem 3 that, if we convert X to Y by the MTF transform and encode Y as an i.i.d. source to a binary codeword, X can be encoded at the entropy rate asymptotically. Especially, in the case of small p in (33), each symbol tends to occur continuously, and the entropy rate of X is much smaller than $\log_2 A$ that is the entropy of symbol-wise stationary distribution of X. This case corresponds to an empirically known result such that *"the data which consist of many runs of symbols can be compressed efficiently by the MTF scheme."*

We think that our theoretical analyses of the MTF scheme are also useful to analyze and/or improve the performance of the BS method without the symbol extension.

Appendix

Lemma 7: If a sequence $\{a_n\}_{n=1}^{\infty}$ is bounded, then it holds that

$$\limsup_{n\to\infty} a_n \geq \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_n \tag{62}$$

$$\liminf_{n\to\infty} a_n \leq \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_n. \tag{63}$$

Proof: We prove (62) only since (63) can be proved in the same way.

Suppose that

$$\limsup_{n\to\infty} a_n = \alpha \tag{64}$$

which implies that, for any $\varepsilon > 0$, there exists $n_0(\varepsilon)$ such that for any $n > n_0(\varepsilon)$, $a_n < \alpha + \varepsilon$. Moreover, from the assumption that $\{a_n\}$ is bounded, there exists $A < \infty$ such that for any

n, $a_n \le A$. Using these bounds, $\frac{1}{N}\sum_{n=1}^{N} a_n$ is bounded above for $N > n_0(\varepsilon)$ as follows;

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} a_n &= \frac{1}{N}\sum_{n=1}^{n_0(\varepsilon)} a_n + \frac{1}{N}\sum_{n=n_0(\varepsilon)+1}^{N} a_n \\ &< \frac{1}{N}\sum_{n=1}^{n_0(\varepsilon)} A + \frac{1}{N}\sum_{n=n_0(\varepsilon)+1}^{N} (\alpha+\varepsilon) \\ &= \frac{n_0(\varepsilon)}{N}A + \frac{N-n_0(\varepsilon)}{N}(\alpha+\varepsilon) \\ &< \alpha + \frac{n_0(\varepsilon)(A-\alpha)}{N} + \varepsilon. \end{aligned} \tag{65}$$

Since (65) is satisfied for any $\varepsilon > 0$ and $N > n_0(\varepsilon)$, we can conclude that

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} a_n \le \alpha$$

which means that (62) holds. □

Lemma 8: For any probability measures P and Q on alphabet $\mathcal{A}$, and any $\varepsilon > 0$, if

$$D(P\|Q) < \varepsilon \tag{66}$$

is satisfied, then the difference of the probabilities P and Q is bounded for any $a \in \mathcal{A}$ by

$$|P(a) - Q(a)| < \sqrt{2\varepsilon}. \tag{67}$$

Proof: Note that the variational distance $d_V(P,Q)$ between two probability distributions P and Q defined by

$$d_V(P,Q) = \sum_{a\in\mathcal{A}} |P(a) - Q(a)|$$

satisfies [12]

$$\frac{1}{2} d_V(P,Q)^2 \le D(P\|Q).$$

Hence, if $D(P\|Q) < \varepsilon$ for $\varepsilon > 0$, then we have for any $a \in \mathcal{A}$ that

$$|P(a) - Q(a)|^2 < \left\{ \sum_{\tilde{a}\in\mathcal{A}} |P(\tilde{a}) - Q(\tilde{a})| \right\}^2 < 2\varepsilon. \qquad \square$$

Lemma 9—([14, Example 2.4]): Let $z = g_p(q)$ be the tangent line of the binary entropy function $z = h(q)$ at $q = p \in (0,1)$. Then $g_p(q)$ is given by (61). Furthermore, for any $p \in (0,1)$, the difference of $g_p(q)$ and $h(q)$ is given by

$$g_p(q) - h(q) = d(q\|p).$$

Proof: The proof is omitted since it is simple. □

REFERENCES

[1] J. L. Bentley, D. D. Sleator, R. E. Tarjan, and V. K. Wei, "A locally adaptive compression scheme," *Commun. Assoc. Comp. Mach.*, vol. 29, pp. 320–330, 1986.

[2] P. Elias, "Interval and recency rank source coding: Two on-line adaptive variable-length schemes," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 1, pp. 3–10, Jan. 1987.

[3] B. Y. Ryabko, "Data compression by means of a "book stack"," *Probl. Inf. Transm.*, vol. 16, pp. 265–269, 1980.

[4] J. Muramatsu, "On the performance of recency-rank and block-sorting universal lossless data compression algorithms," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2621–2625, Sep. 2002.

[5] J. A. Fill, "An exact formula for the move-to-front rule for self-organizing lists," *J. Theor. Probab.*, vol. 9, pp. 113–160, 1996.

[6] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Digital Systems Res. Ctr., SRC Res. Rep. 124, Palo Alto, CA, 1994.

[7] M. Arimura and H. Yamamoto, "Asymptotic optimality of the block sorting data compression algorithm," *IEICE Trans. Fundamentals*, vol. E81-A, pp. 2117–2122, 1998.

[8] ——, "Almost sure convergence coding theorem for block sorting data compression algorithm," in *Proc. 1998 Int. Symp. Information Theory and Its Applications (ISITA98)*, Mexico City, Mexico, Oct. 1998, pp. 286–289.

[9] M. Effros, K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Universal lossless source coding with the burrows wheeler transform," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1061–1081, May 2002.

[10] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 3, pp. 337–343, May 1977.

[11] ——, "Compression of indivisual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 5, pp. 530–536, Sep. 1978.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Akadémiai Kiadó, 1981.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[14] T. S. Han and K. Kobayashi, "Mathematics of information and coding," in *Translations of Mathematical Monographs*. Providence, RI: Amer. Math. Soc., 2002, vol. 203.