PAPER *Special Section on Information Theory and Its Applications*

Asymptotic Optimality of the Block Sorting Data Compression Algorithm*

Mitsuharu ARIMURA†, *Student Member and* Hirosuke YAMAMOTO†, *Member*

SUMMARY In this paper the performance of the Block Sorting algorithm proposed by Burrows and Wheeler is evaluated theoretically. It is proved that the Block Sorting algorithm is asymptotically optimal for stationary ergodic finite order Markov sources. Our proof is based on the facts that symbols with the same Markov state (or context) in an original data sequence are grouped together in the output sequence obtained by Burrows-Wheeler transform, and the codeword length of each group can be bounded by a function described with the frequencies of symbols included in the group.

key words: *lossless data compression, block sorting algorithm, stationary ergodic Markov source, asymptotic optimality*

1. Introduction

The Block Sorting (BS) method was proposed by Burrows and Wheeler [2] as a new lossless data compression algorithm. In the BS method, a given data block x^N is first converted to another sequence y^N with the same length by sorting N sequences generated by cyclic shifts of x^N, and y^N is finally encoded to a binary sequence by the Move-To-Front (MTF) coding scheme [10]**. The practical performance of the BS method is examined by compressing many kinds of files, and it is shown that the BS method attains a high performance compared with other data compression methods [3]–[6]. However, the performance has not been analyzed theoretically well, and even the asymptotic optimality has not been proved yet. The difficulty in theoretical analysis comes from the fact that the probability distribution of y^N cannot easily be obtained because the sorting operation destroys the probabilistic structure of x^N.

In this paper, we evaluate the average codeword length of the BS method for stationary ergodic finite order Markov sources by extending the proof of the MTF coding scheme devised in [10]. In Sect. 2, we review the encoding algorithm of the BS method. The average codeword length of the BS method is theoretically evaluated in Sect. 3. We show that the codeword length can be bounded above by a function described with the numbers of symbols and Markov states (or contexts) included in x^N. This bound means that the average codeword length converges to the entropy of the Markov source asymptotically.

2. Encoding Algorithm of Block Sorting Method

In this section, we review the BS method proposed by Burrows and Wheeler [2]. Let $\widehat{x}^N = x_N x_{N-1} \cdots x_2 x_1$ be a block to encode***, where x_i takes values in a finite discrete ordered set $\mathcal{A} = \{a_1, a_2, \cdots, a_A\}$. $A = |\mathcal{A}|$ is the cardinality of $\mathcal{A}$.

We first obtain N sequences by shifting $\widehat{x}^N$ cyclically. Letting $M(\widehat{x}^N)$ be a $N \times N$ matrix with the obtained N sequences as rows, it is given by

$$M(\widehat{x}^N) = \begin{bmatrix} x_N & x_{N-1} & \cdots & x_2 & x_1 \\ x_{N-1} & x_{N-2} & \cdots & x_1 & x_N \\ \vdots & \vdots & & \vdots & \vdots \\ x_1 & x_N & \cdots & x_3 & x_2 \end{bmatrix}. \qquad (1)$$

Next, we sort the rows of the matrix $M(\widehat{x}^N)$ in lexicographical order. Assume that $\widehat{x}^N$ is located at R$(\widehat{x}^N)$-th row in the sorted matrix $\widetilde{M}(\widehat{x}^N)$. Then we use the last column of $\widetilde{M}(\widehat{x}^N)$, say y^N, and the number R$(\widehat{x}^N)$ in the encoding.

The last column $y^N = y_1 y_2 \cdots y_N$ is encoded to a sequence of positive integers $z^N = n_1 n_2 \cdots n_N$ by the MTF coding scheme with an initial list $L = (a_1, a_2, \cdots, a_A)$. Each y_l is mapped to integer n_l if y_l is equal to the n_l-th element of L, and the n_l-th element is moved to the front of list L. Each obtained integer n_l is finally converted to a binary sequence by a universal code of the positive integers [12]–[14] or entropy coding like Huffman code, Arithmetic code, etc. On the other hand, R$(\widehat{x}^N)$ can be represented by a binary number with $\lceil \log N \rceil$ bits.

The encoding x^N to y^N is called B-W (Burrows-Wheeler) transform. Refer [2] for more details of the BS algorithm and how to decode $\widehat{x}^N$ from y^N and R$(\widehat{x}^N)$.

Example
In the case of $\widehat{x}^N$ =`arbadacarba` (which is the reverse of x^N =`abracadabra`),

Manuscript received January 16, 1998.
Manuscript revised April 13, 1998.

†The authors are with the School of Engineering, the University of Tokyo, Tokyo, 113–8656 Japan.

*This work was presented at the 1997 Symposium on Information Theory and Its Applications, Matsuyama, Japan.

**It is known that the MTF coding scheme is equivalent to the recency rank coding proposed in [11].

***In order to simplify the theoretical analysis, we use the reversed block $\widehat{x}^N = x_N x_{N-1} \cdots x_2 x_1$ instead of $x^N = x_1 x_2 \cdots x_{N-1} x_N$ in this paper.

1. the B-W transform:

$$M(\widehat{x}^N) = \begin{bmatrix} \texttt{arbadacarba} \\ \texttt{rbadacarbaa} \\ \texttt{badacarbaar} \\ \texttt{adacarbaarb} \\ \texttt{dacarbaarba} \\ \texttt{acarbaarbad} \\ \texttt{carbaarbada} \\ \texttt{arbaarbadac} \\ \texttt{rbaarbadaca} \\ \texttt{baarbadacar} \\ \texttt{aarbadacarb} \end{bmatrix}$$

$$\widetilde{M}(\widehat{x}^N) = \begin{bmatrix} \texttt{aarbadacarb} \\ \texttt{acarbaarbad} \\ \texttt{adacarbaarb} \\ \texttt{arbaarbadac} \\ \texttt{arbadacarba} \\ \texttt{baarbadacar} \\ \texttt{badacarbaar} \\ \texttt{carbaarbada} \\ \texttt{dacarbaarba} \\ \texttt{rbaarbadaca} \\ \texttt{rbadacarbaa} \end{bmatrix}$$

$$y^N = \texttt{bdbcarraaaa}$$

$$\mathrm{R}(\widehat{x}^N) = 5$$

2. the MTF coding with initial list $L = (\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{r})$:

$$z^N = 24244512111.$$

We assume in the following that positive integer n is encoded to a binary codeword by a universal code of positive integers, in which the bit length of codeword can be upper bounded by a function $f(n)$ satisfying the following properties.

Property 1:

1. $0 < f(s) < \infty$ for any real number $s > 0$.
2. $f(s)$ is concave.
3. $f(s)$ is monotonically increasing.

We note that we can easily find such functions for many useful universal codes of positive integers. For instance, Elias δ-code[12] has $f(s) = \log s + 2\log(1 + \log s) + 1$.

3. Asymptotic Performance of Block Sorting Method

We evaluate the asymptotic performance of the BS method theoretically in this section.

Let $\{X_i\}_{i=1}^{\infty}$ be a stationary, ergodic k-th order Markov process which takes values in a finite discrete ordered alphabet $\mathcal{A}$. The context set in the k-th order Markov process can be denoted as $\mathcal{C} = \{c : \; c \in \mathcal{A}^k\}$. Then we have for any $i > k$ that

$$\begin{aligned} &\Pr(X_i = \widehat{x}_i | X_{i-1}X_{i-2} \cdots X_2X_1 = \widehat{x}^{i-1}) \\ &\quad = \Pr(X_i = \widehat{x}_i | X_{i-1}X_{i-2} \cdots X_{i-k} = \widehat{x}_{i-k}^{i-1}), \\ &\quad \widehat{x}_i \in \mathcal{A}, \;\; \widehat{x}_{i-k}^{i-1} = x_{i-1}x_{i-2} \cdots x_{i-k} \in \mathcal{C}. \end{aligned} \tag{2}$$

Before analyzing the performance of the BS method theoretically, we consider the properties of matrix $\widetilde{M}(\widehat{x}^N)$.

We first note that each symbol of $\mathcal{A}$ has the same frequency in every row and every column of $\widetilde{M}(\widehat{x}^N)$ as $\widehat{x}^N$ because each row of $\widetilde{M}(\widehat{x}^N)$ consists of the cyclic shift of $\widehat{x}^N$. Furthermore, for any $\widehat{x}_i$ and its context $c = \widehat{x}_{i-k}^{i-1}$ in $\widehat{x}^N$, there exists a row that has c as the prefix and $\widehat{x}_i$ as the last element. Since $\widetilde{M}(\widehat{x}^N)$ is sorted in the lexicographical order, rows having the same context c are grouped together in $\widetilde{M}(\widehat{x}^N)$. Hence, the rows of $\widetilde{M}(\widehat{x}^N)$ can be classified by contexts $\mathcal{C} = \{c_j\}$ as shown in Fig. 1. Furthermore, since we use the reversed sequence $\widehat{x}^N$, the probability $\mathrm{P}(y_j|c_l)$ can easily be obtained from the stationary probability of the Markov source as $\mathrm{P}(y_j|c_l) = \Pr(X_i = y_j | X_{i-1}X_{i-2} \cdots X_{i-k} = c_l)$.

Since each row in $\widetilde{M}(\widehat{x}^N)$ is one of the cyclically shifted sequences of $\widehat{x}^N$, sequence $x_lx_{l-1} \cdots x_1x_Nx_{N-1} \cdots x_{N-k+l+1}$, $0 \leqq l \leqq k-1$, becomes a context (prefix) of some row. This context is not an actual context of x_{l+1}, and this edge effect may worsen the performance. However, when the block size N is sufficiently large, such degradation of the performance can become negligible because the number of such rows is only k in the case of the k-th order Markov source and, hence, the degradation caused by the edge effect is $\mathrm{O}(\frac{k}{N})$ in the average codeword length. In the following argument, we neglect the edge effect for the sake of simplicity. For more details of the edge effect, see Note 1 in this section.

For context $c \in \mathcal{C}$ and symbol $a \in \mathcal{A}$, let $N(c)$ and $N(a, c)$ denote the frequencies of c and ac in x^N, respectively. Then the next theorem holds.

Theorem 1: When the sequence y^N is encoded by the

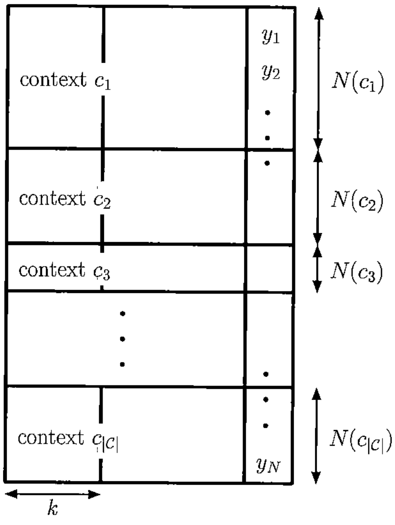

Fig. 1 $\widetilde{M}(\widehat{x}^N)$.

MTF coding scheme and a universal code of positive integers, which satisfies Property 1, the codeword length per symbol is bounded by

$$L(y^N) \leqq \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \frac{N(a,c)}{N} f\left(\frac{N(c)+A}{N(a,c)}\right), \tag{3}$$

where $A = |\mathcal{A}|$. (When $N(a,c)=0$, we use a convention $0f(\infty)=0$.)

We note that this theorem is an extension of the coding theorem for the MTF coding scheme[10, Theorem 1].

Proof: As shown in Fig. 1, symbols with the same context are grouped in y^N. Let $y^N = y_{c_1}^{N(c_1)} y_{c_2}^{N(c_2)} \cdots y_{c_{|C|}}^{N(c_{|C|})}$ and, for each $c \in \mathcal{C}$, $y_c^{N(c)} = y_{c1} y_{c2} \cdots y_{cN(c)}$, where y_{cl}'s are the symbols that are classified into a group of context c. Assume that $y_c^{N(c)}$ are encoded to a sequence of positive integers $z_c^{N(c)} = n_1 n_2 \cdots n_j \cdots n_{N(c)}$ by the MTF coding scheme, and symbol $a \in \mathcal{A}$ appears at indexes $j = t_1, t_2, \cdots, t_{N(a,c)}$ in $y_c^{N(c)}$. Then we have

$$n_j \leqq \begin{cases} A, & \text{for } j = 1 \\ t_j - t_{j-1}, & \text{for } 2 \leqq j \leqq N(a, c_l) \end{cases}$$

because n_j obtained by the MTF coding scheme is bounded above by the size of list L and the index interval of the same kind of symbols.

Since n_j is encoded to a binary sequence by a universal code of integers, the codeword length of which is bounded by $f(n_j)$, the sum of codeword length of the same symbol a occurred in $y_c^{N(c)}$ is given by

$$\begin{aligned} & L_N(a|y_c^{N(c)}) \\ & \leqq f(A) + \sum_{i=2}^{N(a,c)} f(t_i - t_{i-1}) \\ & = N(a,c)\left\{\frac{1}{N(a,c)} f(A) \right. \\ & \qquad \left. + \frac{1}{N(a,c)} \sum_{i=2}^{N(a,c)} f(t_i - t_{i-1})\right\} \\ & \overset{a)}{\leqq} N(a,c) f\left(\frac{1}{N(a,c)}\left\{A + \sum_{i=2}^{N(a,c)} (t_i - t_{i-1})\right\}\right) \\ & = N(a,c) f\left(\frac{t_{N(a,c)} - t_1 + A}{N(a,c)}\right) \\ & \overset{b)}{\leqq} N(a,c) f\left(\frac{t_{N(a,c)} + A}{N(a,c)}\right) \\ & \overset{c)}{\leqq} N(a,c) f\left(\frac{N(c) + A}{N(a,c)}\right), \end{aligned} \tag{4}$$

where the inequalities hold by the following reasons.

a) Jensen's inequality and the concavity of $f(s)$.

b) $f(s)$ is monotonically increasing.

c) $t_{N(a,c)} \leqq N(c)$.

Hence, using the convention of $0f(\infty) = 0$, (4) holds for any $a \in \mathcal{A}$ and $c \in \mathcal{C}$. The codeword length per symbol is given by

$$\begin{aligned} L(y^N) & \leqq \frac{1}{N} \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} L_N(a|y_c^{N(c)}) \\ & \leqq \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \frac{N(a,c)}{N} f\left(\frac{N(c)+A}{N(a,c)}\right). \end{aligned} \tag{5}$$

Q.E.D.

Note that it is difficult to derive the probability distribution of y^N because of the sorting procedure. However, the upper bound of $L(y^N)$ given by (3) is described only with $N(a)$ and $N(a,c)$ which can easily be known from $\widehat{x}^N$. Hence we can evaluate $L(y^N)$ by using the probability distribution of $\widehat{x}^N$.

Now we give the main theorem.

Theorem 2: Let $EL_{BS}(\widehat{X}^N)$ be the expected codeword length of $\widehat{x}^N$ for the stationary ergodic finite order Markov source with the entropy $H(X) = \lim_{l \to \infty} \frac{1}{l} H(X^l)$. Then for any $\varepsilon > 0$ and sufficiently large N,

$$EL_{BS}(\widehat{X}^N) \leqq H(X) + \mathrm{O}(\log H(X)) + \varepsilon. \tag{6}$$

where $\mathrm{O}(\log s) = 2\log(s+1) + 1$.

Proof: Let $P(c)$ and $P(a|c)$ be the probability distributions of context c and symbol a under context c, respectively, for the stationary ergodic k-th order Markov process. Then the set of typical sequences $\mathcal{T}^{(N)}$ is defined as

$$\begin{aligned} \mathcal{T}^{(N)} = \Bigg\{ \widehat{x}^N : & \left| \frac{N(c)}{N-k+1} - P(c) \right| \leqq \varepsilon_0, \\ & \left| \frac{N(a,c)}{N-k} \cdot \frac{N-k+1}{N(c)} - P(a|c) \right| \leqq \varepsilon_0 \\ & \text{for all } a \in \mathcal{A} \text{ and } c \in \mathcal{C} \Bigg\}. \end{aligned}$$

Then we have from the law of large numbers that

$$\Pr\{\mathcal{T}^{(N)}\} \geqq 1 - \delta, \tag{7}$$

where δ can tend to zero by letting N sufficiently large for any $\varepsilon_0 > 0$[1].

Since k is fixed, $\mathcal{T}^{(n)}$ can be described as

$$\begin{aligned} \mathcal{T}^{(N)} = \Bigg\{ \widehat{x}^N : & \left| \frac{N(c)}{N} - P(c) \right| \leqq \varepsilon_1, \\ & \left| \frac{N(a,c)}{N(c)} - P(a|c) \right| \leqq \varepsilon_1 \\ & \text{for all } a \in \mathcal{A} \text{ and } c \in \mathcal{C} \Bigg\}, \end{aligned}$$

where $\varepsilon_1 \to 0$ when $\varepsilon_0 \to 0$.

We first evaluate the codeword length $L_{BS}(\widehat{x}^N)$ of the BS method for the case where $\widehat{x}^N \in \mathcal{T}^{(N)}$.

Since y^N and $\mathrm{R}(\widehat{x}^N)$ can be encoded with $L(y^N)$ and $\lceil \log N \rceil / N$ bits per symbol[†], respectively, letting $\mathcal{C}^* = \{c : c \in \mathcal{C} \text{ and } P(c) > 0\}$ and $\mathcal{A}^*(c) = \{a : a \in \mathcal{A} \text{ and } P(a|c) > 0\}$, we have

$$\begin{aligned}
&L_{BS}(\widehat{x}^N) \\
&= L(y^N) + \frac{\lceil \log N \rceil}{N} \\
&\leqq \sum_{c \in \mathcal{C}^*} \frac{N(c)}{N} \sum_{a \in \mathcal{A}^*(c)} \frac{N(a,c)}{N(c)} f\left(\frac{N(c)+A}{N(a,c)}\right) \\
&\quad + \frac{\lceil \log N \rceil}{N} \\
&\leqq \sum_{c \in \mathcal{C}^*} \frac{N(c)}{N} \sum_{a \in \mathcal{A}^*(c)} \frac{N(a,c)}{N(c)} f\left(\frac{N(c)+A}{N(a,c)}\right) \\
&\quad + \frac{\log N + 1}{N} \\
&\leqq \sum_{c \in \mathcal{C}^*} (P(c) + \varepsilon_1) \sum_{a \in \mathcal{A}^*(c)} (P(a|c) + \varepsilon_1) \\
&\quad \times f\left(\frac{1}{P(a|c) - \varepsilon_1} + \frac{A}{N(a,c)}\right) + \frac{\log N + 1}{N} \\
&= \sum_{c \in \mathcal{C}^*} P(c) \sum_{a \in \mathcal{A}^*(c)} P(a|c) f\left(\frac{1}{P(a|c) - \varepsilon_1} + \frac{A}{N(a,c)}\right) \\
&\quad + \varepsilon_1 \left[\sum_{c \in \mathcal{C}^*} P(c) \sum_{a \in \mathcal{A}^*(c)} f\left(\frac{1}{P(a|c) - \varepsilon_1} + \frac{A}{N(a,c)}\right) \right. \\
&\quad + \sum_{c \in \mathcal{C}^*} \sum_{a \in \mathcal{A}^*(c)} P(a|c) f\left(\frac{1}{P(a|c) - \varepsilon_1} + \frac{A}{N(a,c)}\right) \\
&\quad \left. + \varepsilon_1 \sum_{c \in \mathcal{C}^*} \sum_{a \in \mathcal{A}^*(c)} f\left(\frac{1}{P(a|c) - \varepsilon_1} + \frac{A}{N(a,c)}\right) \right] \\
&\quad + \frac{\log N + 1}{N}. \qquad (8)
\end{aligned}$$

From Property 1, $f(s)$ is a continuous function and $f\left(\frac{1}{P(a|c)}\right)$ is finite for any $P(a|c) > 0$. Furthermore, $N(a,c) \to \infty$ as $N \to \infty$ for $P(c)P(a|c) > 0$. Hence we have

$$\begin{aligned}
&L_{BS}(\widehat{x}^N) \\
&\leqq \sum_{c \in \mathcal{C}^*} P(c) \sum_{a \in \mathcal{A}^*(c)} P(a|c) f\left(\frac{1}{P(a|c)}\right) + \varepsilon_2 \\
&= \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) f\left(\frac{1}{P(a|c)}\right) + \varepsilon_2, \qquad (9)
\end{aligned}$$

where $\varepsilon_2 \to 0$ as $N \to \infty$, and the last equality follows from the convention $0f(\infty) = 0$.

If we use Elias δ-code[12], which satisfies $f(s) = \log s + \mathrm{O}(\log\log s)$, (9) becomes

$$\begin{aligned}
&L_{BS}(\widehat{x}^N) \\
&\leqq \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) \log \frac{1}{P(a|c)} \\
&\quad + \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) \mathrm{O}\left(\log\log \frac{1}{P(a|c)}\right) \\
&\quad + \varepsilon_2. \qquad (10)
\end{aligned}$$

Applying Jensen's inequality to the summation of $\mathcal{A}$ and $\mathcal{C}$, (10) is further bounded by

$$\begin{aligned}
&L_{BS}(\widehat{x}^N) \\
&\leqq \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) \log \frac{1}{P(a|c)} \\
&\quad + \mathrm{O}\left(\log \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) \log \frac{1}{P(a|c)}\right) \\
&\quad + \varepsilon_2 \\
&= H(X) + \mathrm{O}(\log H(X)) + \varepsilon_2, \qquad (11)
\end{aligned}$$

where

$$H(X) \triangleq \sum_{c \in \mathcal{C}} P(c) \sum_{a \in \mathcal{A}} P(a|c) \log \frac{1}{P(a|c)}. \qquad (12)$$

Next we consider the case where $\widehat{x}^N \notin \mathcal{T}^{(N)}$. Since n_j obtained by the MTF coding is always bounded by $n_j \leqq A$, we have

$$\begin{aligned}
L_{BS}(\widehat{x}^N) &\leqq f(A) + \frac{\lceil \log N \rceil}{N} \\
&\leqq f(A) + \varepsilon_3 \\
&= \log A + \mathrm{O}(\log\log A) + \varepsilon_3, \qquad (13)
\end{aligned}$$

where $\varepsilon_3 \to 0$ as $N \to \infty$.

From (7), (11), and (13), the expected codeword length $EL_{BS}(\widehat{X}^N)$ is given by

$$\begin{aligned}
EL_{BS}(\widehat{X}^N) &= \sum_{\widehat{x}^N \in \mathcal{A}^N} P(\widehat{x}^N) L_{BS}(\widehat{x}^N) \\
&= \sum_{\widehat{x}^N \in \mathcal{T}^{(N)}} P(\widehat{x}^N) L_{BS}(\widehat{x}^N) \\
&\quad + \sum_{\widehat{x}^N \notin \mathcal{T}^{(N)}} P(\widehat{x}^N) L_{BS}(\widehat{x}^N) \\
&\leqq H(X) + \mathrm{O}(\log H(X)) + \varepsilon_2 \\
&\quad + \delta(\log A + \mathrm{O}(\log\log A) + \varepsilon_3) \\
&\leqq H(X) + \mathrm{O}(\log H(X)) + \varepsilon_4,
\end{aligned}$$

where $\varepsilon_4 \to 0$ as $N \to \infty$. Q.E.D.

Note 1: k rows with prefix $x_l \cdots x_1 x_N \cdots x_{N-k+l+1}$, $0 \leqq l \leqq k-1$, in $\widetilde{M}(\widehat{x}^N)$ have the so-called edge effect

[†] If block size N is not fixed, N also must be encoded by a universal code of positive integers. However, this additional rate can be bounded by $f(N)/N$, which becomes $(\log N + 2\log(1 + \log N) + 1)/N$ if Elias δ-code is used.

as mentioned before in this section. But y_j of any row can be encoded with at most $f(A)$ bits. Furthermore, if $y_l = y_i \neq y_j$ for $l < i < j$, then n_j obtained from y_j is not affected by y_l in the MTF coding scheme. In other words, each y_j affects each kind of symbol only once and, hence, each y_j affects at most A symbols of y^N. Therefore, the total edge effect can be bounded by $\frac{k}{N}(A+1)f(A)$, which is $\mathrm{O}(\frac{k}{N})$.

Note 2: We proved Theorem 2 for the case where the reversed block $\widehat{x}^N$, instead of x^N, is encoded by the BS coding. However, for the case of the non-reversed x^N, we have the same bound

$$EL_{BS}(X^N) \leqq H(X) + \mathrm{O}(\log H(X)) + \varepsilon \tag{14}$$

because of

$$\begin{aligned} H(X) &= \lim_{N\to\infty} \frac{1}{N} H(X_1X_2\cdots X_N) \\ &= \lim_{N\to\infty} \frac{1}{N} H(X_NX_{N-1}\cdots X_1). \end{aligned}$$

Note 3: We note that the bounds (6) and (14) do not attain the entropy since they have an additional term $\mathrm{O}(\log H(X))$. However, this term can be removed by the so-called symbol extension. Letting $u_j = x_{m(j-1)+1}^{mj}$ for x^N with $N = mL$, we have $H_m(U) = \lim_{L\to\infty} \frac{1}{L} H(U_1U_2\cdots U_L) = \lim_{N\to\infty} \frac{m}{N} H(X_1X_2\cdots X_N) = mH(X)$. Hence, encoding U^L by the BS coding, the expected codeword length of U^L is bounded by

$$\begin{aligned} EL_{BS}(U^L) &\leqq H_m(U) + \mathrm{O}(\log H_m(U)) + \varepsilon \\ &= mH(X) + \mathrm{O}(\log mH(X)) + \varepsilon \\ &\leqq mH(X) + \mathrm{O}(\log mA) + \varepsilon. \end{aligned}$$

This means that the expected codeword length per symbol x is bounded by

$$\begin{aligned} EL_{BS}^{(m)}(X^N) &\triangleq \frac{1}{m} EL_{BS}(U^L) \\ &\leqq H(X) + \mathrm{O}\left(\frac{\log mA}{m}\right) + \varepsilon \\ &\leqq H(X) + \varepsilon', \end{aligned}$$

where $\varepsilon' \to 0$ as $N \to \infty$ and $m \to \infty$.

Theorem 3: Let $EL_{BS}^{(m)}(X^N)$ be the expected codeword length of the m-symbol extension of $x^N = x^{mL}$ for the stationary ergodic finite order Markov source. Then for any $\varepsilon > 0$ and sufficiently large m and N we have

$$EL_{BS}^{(m)}(X^N) \leqq H(X) + \varepsilon. \tag{15}$$

4. Concluding Remarks

We proved that the BS method with symbol extension is asymptotically optimal for stationary ergodic finite order Markov source.

We note that the Context Sorting (CS) method proposed by Yokoo and Takahashi [7] uses the similar coding technique to the BS method and it is proved that the CS method with symbol extension is asymptotically optimal. The CS method is a sequential algorithm and, hence, symbols with the same context are sequentially encoded. But, since the BS method is a blockwise algorithm, symbols even with the same context may be disordered by the sorting. Therefore, the proving technique of the CS method cannot be applied to the BS method.

The asymptotic optimality of the BS method (and the CS method [7]) are attained by the symbol extension. However, it is known that the BS and CS methods can attain high performance without the symbol extension in practical uses [2]–[7]. Hence, it might be possible to prove the asymptotic optimality without the symbol extension. We note that Sadakane [9] proved the asymptotic optimality, though his analysis lacks accuracy, for the case where the B-W transform output y^N of the stationary ergodic finite order Markov source is encoded to a binary sequence by two path arithmetic code without the MTF encoding.

References

[1] T.M. Cover and J.A. Thomas, "Elements of information theory," John Wiley & Sons, Inc., New York, 1991.

[2] M. Burrows and D.J. Wheeler, "A block-sorting lossless data compression algorithm," SRC Research Report 124, Digital Systems Research Center, Palo Alto, CA., May 1994.

[3] P. Fenwick, "Block sorting text compression," Proc. 19th Australasian Computer Science Conference, Melbourne, Australia, Jan. 1996.

[4] M. Nelson, "Data compression with the Burrows-Wheeler transform," Dr. Dobb's Journal, pp.46–50, Sept. 1996.

[5] M. Schindler, "A fast block-sorting algorithm for lossless data compression," Proc. Data Compression Conference (DCC97), Snowbird, Utah, p.469, March 1997.

[6] Z. Arnavut, "Block sorting and compression," Proc. Data Compression Conference (DCC97), Snowbird, Utah, pp.181–190, March 1997.

[7] H. Yokoo and M. Takahashi, "Data compression by context sorting," IEICE Trans. Fundamentals, vol.E78-A, no.5, pp.681–686, May 1995.

[8] M. Arimura and H. Yamamoto, "A performance analysis of block sorting algorithm," Proc. Simposium on Information Theory and Its Applications (SITA97), Matsuyama, Japan, pp.493–496, Dec. 1997.

[9] K. Sadakane, "On optimality of variants of block sorting compression," Proc. Simposium on Information Theory and Its Applications (SITA97), Matsuyama, Japan, pp.357–360, Dec. 1997.

[10] J.L. Bentley, D.D. Sleator, R.E. Tarjan, and V.K. Wei, "A locally adaptive compression scheme," Commun. ACM, vol.29, no.4, pp.320–330, April 1986.

[11] 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.

[12] P. Elias, "Universal codeword sets and representations of the integers," IEEE Trans. Inf. Theory, vol.IT-21, no.2, pp.194–203, March 1975.

[13] H. Yamamoto and H. Ochi, "A new asymptotically optimal code for the positive integers," IEEE Trans. Inf. Theory, vol.37, no.5, pp.1420–1429, Sept. 1991.

[14] T. Amemiya and H. Yamamoto, "A new class of the universal representation for the positive integers," IEICE Trans. Fundamentals, vol.E76-A, no.3, pp.447–452, March 1993.

Mitsuharu Arimura was born in Kagoshima, Japan, on December 17, 1970. He received the B.E. degree in mathematical engineering in 1994, and the M.E. degree in information engineering in 1996, from University of Tokyo, Tokyo, Japan. He is currently a graduate student of the doctor course in the department of information engineering, University of Tokyo. His research interests include Shannon theory and its applications to source coding problems. Mr. Arimura is a member of the IEEE.

Hirosuke Yamamoto was born in Wakayama, Japan, on November 15, 1952. He received the B.E. degree from Shizuoka University, Shizuoka, Japan, in 1975 and the M.E. and Dr.E. degrees from the University of Tokyo, Tokyo, Japan, in 1977 and 1980, respectively, all in electrical engineering. In 1980 he joined Tokushima University, Tokushima, Japan. He was an Associate Professor at Tokushima University from 1983 to 1987, and at University of Electro-Communications, Tokyo, Japan, from 1987 to 1993. Since 1993 he has been an Associate Professor in the Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Japan. In 1989–90, he was a Visiting Scholar at the Information Systems Laboratory, Stanford University. His research interests are in Shannon theory, coding theory, cryptography, and communication theory. Dr. Yamamoto is a member of the IEEE.