A Source Model with Probability Distribution over Word Set and Recurrence Time Theorem

Masayuki GOTO, Toshiyasu MATSUSHIMA, and Shigeichi HIRASAWA

SUMMARY Nishiara and Morita defined an i.i.d. word-valued source which is defined as a pair of an i.i.d. source with a countable alphabet and a function which transforms each symbol into a word over finite alphabet. They showed the asymptotic equipartition property (AEP) of the i.i.d. word-valued source and discussed the relation with source coding algorithm based on a string parsing approach. However, their model is restricted in the i.i.d. case and any universal code for a class of word-valued sources isn't discussed. In this paper, we generalize the i.i.d. word-valued source to the ergodic word-valued source which is defined by an ergodic source with a countable alphabet and a function from each symbol to a word. We show existence of entropy rate of the ergodic word-valued source and its formula. Moreover, we show the recurrence time theorem for the ergodic word-valued source with a finite alphabet. This result clarifies that Ziv-Lempel code (ZL77 code) is universal for the ergodic word-valued source.

key words: word-valued source, word set, word sequences, recurrence time, Ziv-Lempel code

1. Introduction

The source coding theorem and universality of coding have been extended from the case of independently and identically distributed (i.i.d.) sources to the cases of stationary ergodic sources, stationary sources, and general sources [4], [8], [10], [11]. The stationary source is defined by a source whose probability structure does not change for time shift per a symbol unit in a source alphabet. It has an entropy rate that is a limit of the occurrence time, Ziv-Lempel code.

At first, we give the mathematical definition of the ergodic word-valued source.

Let \( Y = Y_1 Y_2 Y_3 \cdots \) be an ergodic source with countable alphabet \( Y \). Let \( \mathcal{X} \) be a finite alphabet and \( \mathcal{X}^\ast \) be the set of all finite sequences over \( \mathcal{X} \). Considering a mapping \( \phi : \mathcal{Y} \to \mathcal{X}^\ast \), we define \( w = \phi(y) \), \((w \in \mathcal{X}^\ast)\) as a word. Here, The range of \( \phi = \phi(y) \), \( y \in \mathcal{Y} \) is denoted by \( \mathcal{W} \), that is \( w \in \mathcal{W} \).
Example 1: Let $\mathcal{Y}$ be $\mathcal{Y} = \{0, 1, 2, 3\}$, and $\mathcal{X}$ be $\mathcal{X} = \{0, 1\}$. And let $\phi : \mathcal{Y} \rightarrow \mathcal{X}^*$ be defined by $\phi(0) = 0$, $\phi(1) = 10$, $\phi(2) = 110$, $\phi(3) = 111$. Then, $\mathcal{W} = \{0, 10, 110, 111\}$. Figure 1 shows the word tree representing the word set.

Let $|w|$ be the length of a word $w \in \mathcal{W}$. For example, if $w = 1010$ then $|w| = 4$.

Let $\mathbf{X} = X_1X_2X_3\cdots$ be a source which is the target of data compression. The source $\mathbf{X} = X_1X_2X_3\cdots$ is defined as a concatenation of sequences $W_1 = \phi(Y_1)$, $W_2 = \phi(Y_2)$, $W_3 = \phi(Y_3)$, $\cdots$ for $\mathbf{Y} = Y_1Y_2Y_3\cdots$. We call $\mathbf{X}$ the ergodic word-valued source. $W_1W_2W_3\cdots$ and $\phi(Y_1)\phi(Y_2)\phi(Y_3)\cdots$ are also denoted by $\mathbf{W}$ and $\phi(\mathbf{Y})$ respectively. Then $\mathbf{W} = \phi(\mathbf{Y})$. The data sequence emitted from the source $\mathbf{X}$, that is a realization value, is denoted by $x = x_1x_2x_3\cdots$. For each finite number $n \in \mathbb{Z}^+ = \{1, 2, 3, \cdots\}$, we define $X^n = X_1X_2X_3\cdots X_n$ and $x^n = x_1x_2x_3\cdots x_n$. Similarly, for each finite number $m \in \mathbb{Z}^+$, we define $Y^m = Y_1Y_2Y_3\cdots Y_m$, $y^m = y_1y_2y_3\cdots y_m$, $W^m = W_1W_2W_3\cdots W_m$ and $w^m = w_1w_2w_3\cdots w_m$. The mapping $\phi : \mathcal{Y}^m \rightarrow \mathcal{W}^m$ is also denoted by $\phi(Y^m)$ which is concatenating the sequences $\phi(Y_1)$, $\phi(Y_2)$, $\phi(Y_3)$, $\cdots$, $\phi(Y_m)$. Then $W^m = \phi(Y^m)$. A word sequence $w^m = w_1w_2\cdots w_m$ from a source is just as it is regarded as a data sequence $x^n$, that is $x^n = w^m$ if $n = |w_1| + |w_2| + \cdots + |w_m|$ for some given $m$.

The probability distributions of $\mathbf{Y}$ and $\mathbf{W}$ are denoted by

$$P_{Y^m}(y^m) = Pr\{Y_1 = y_1, Y_2 = y_2, \cdots, Y_m = y_m\},$$

and

$$P_{W^m}(w^m) = Pr\{W_1 = w_1, W_2 = w_2, \cdots, W_m = w_m\},$$

respectively. The relation between $P_{Y^m}(y^m)$ and $P_{W^m}(w^m)$ is given by

$$P_{W^m}(w^m) = \sum_{y^m : w^m = \phi(y^m)} P_{Y^m}(y^m).$$

When $m = 1$, we briefly denote $P_W(w) = P_W(w)$.

The probability distribution of the target source sequence $\mathbf{X}$ is denoted by

$$P_{X^n}(x^n) = Pr\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}. \quad (4)$$

When $n = 1$, we briefly denote $P_X(x) = P_X(x)$. Using notations as $X^n_i = X_iX_{i+1}\cdots X_j$ and $x^n_i = x_ix_{i+1}\cdots x_j$ for $i < j$, we also define

$$P_{X_i^n}(x^n_i) = Pr\{X_i = x_i, X_{i+1} = x_{i+1}, \cdots, X_j = x_j\}. \quad (5)$$

If $i = 1$ for $X^n_i$ then we abbreviate $X^n_1$ by $X^n$ as $X^n = X_1X_2\cdots X_n$.

Definition 1: (A prefix word set) If each word $\forall w \in W$ is not the prefix of other words $\forall w' \in \mathcal{W} (w' \neq w)$, then we call $\mathcal{W}$ a prefix word set.

Example 2: Consider a simple example as $\mathcal{W} = \{00, 01, 10, 11\}$. Let $\mathbf{Y}$ be an i.i.d. source, that is $\mathbf{W}$ is also an i.i.d. source. Denoting $P_W(00) = \theta_1$, $P_W(01) = \theta_2$, $P_W(10) = \theta_3$, we can calculate the probabilities of events, for example $P_{X^n}(01000110) = \theta_1\theta_2\theta_3$.

In this case, if $t$ is an odd number then $P_{X^n}(0) = \theta_1 + \theta_2$, else if $t$ is an even number then $P_{X^n}(0) = \theta_1 + \theta_3$. Therefore, the source is periodic but not stationary.

If there exists

$$H(\mathbf{X}) = \lim_{n \to \infty} \frac{1}{n} H_n(X^n)$$

$$= \lim_{n \to \infty} \left[ -\frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log P_{X^n}(x^n) \right], \quad (6)$$

then we call $H(\mathbf{X})$ as the entropy rate of $\mathbf{X}$. Then, an entropy rate of the word sequence $\mathbf{W}$ is given by

$$H(\mathbf{W}) = \lim_{m \to \infty} \frac{1}{m} H_m(W^m)$$

$$= \lim_{m \to \infty} \left[ -\frac{1}{m} \sum_{w^m \in \mathcal{W}^m} P_{W^m}(w^m) \log P_{W^m}(w^m) \right]. \quad (7)$$

Let $L_i$ be the length of $W_i$, i.e., $L_i \overset{\text{def}}{=} |W_i| = |\phi(Y_i)|$. The expected word length rate $E[|W|] = E[|\phi(\mathbf{Y})|]$ is defined by

$$E[|W|] = \lim_{m \to \infty} \frac{1}{m} E \left[ \sum_{i=1}^{m} L_i \right], \quad (8)$$

Fig. 1 An example of the prefix word set.
where $E[\cdot]$ means an expectation.

Nishiara and Morita [16] showed the asymptotic equipartition property (AEP) of the word valued source when $W$ is an i.i.d. source.

**Lemma 1** (AEP of an i.i.d. word-valued source [16]): Let $Y$ be an i.i.d. source. If $X = \phi(Y)$ such that $H(Y) < \infty$ and $E[|W|] < \infty$, we have

$$\limsup_{n \to \infty} \frac{1}{n} \left[ -\log P_{X^n}(X^n) \right] \leq \frac{H(Y)}{E[|W|]}, \ \text{a.s.} \ (9)$$

and

$$\limsup_{n \to \infty} \frac{1}{n} E \left[ -\log P_{X^n}(X^n) \right] \leq \frac{H(Y)}{E[|W|]}, \ (10)$$

Furthermore, if $W$ is a prefix word set, then we have

$$\lim_{n \to \infty} \frac{1}{n} \left[ -\log P_{X^n}(X^n) \right] = H(X), \ \text{a.s.} \ (11)$$

and

$$H(X) = \frac{H(W)}{E[|W|]}, \ (12)$$

where $H(W)$ and $E[|W|]$ are given by

$$H(W) = - \sum_{w \in W} P_W(w) \log P_W(w), \ (13)$$

and

$$E[|W|] = \sum_{w \in W} |w| P_W(w), \ (14)$$

from the definitions (7) and (8) respectively on this case.

In the following section, we generalize the i.i.d. word-valued source to the ergodic word-valued source which is defined by an ergodic source with a countable alphabet and a function from each symbol into a word. We show the entropy rate of the ergodic word-valued source. Here, cases exist in which the information about a pair of a source $Y$ and a mapping $\phi$ is previously unknown in the practical case. Therefore, it is important to construct a universal code whose mean codelength converges to the entropy rate. In Sect. 4, we show the recurrence time theorem. From this theorem, we can see that the ZL77 code is universal for the ergodic word-valued source.

3. Main Result I: The Entropy Rate of the Ergodic Word-Valued Source

3.1 Main Theorem

At first, we show the entropy rate of the ergodic word-valued source. If we know the probability structure, i.e. a set of a source $Y$, a prefix word set $W$, and a mapping $\phi$, then we can encode $X$ with a mean codelength which converges to the lower bound $H(X)$ almost surely.

**Theorem 1**: For a prefix word set $W$, let the probability distribution $P_{W^n}(W^n)$ be a stationary ergodic with respect to $m$. If $H(W) < \infty$ and $E[|W|] < \infty$, then $H(X)$ is given by

$$H(X) = \frac{H(W)}{E[|W|]} \ , \quad (15)$$

Moreover,

$$- \frac{1}{n} \log P_{X^n}(X_1 X_2 \cdots X_n) \to \frac{H(W)}{E[|W|]}, \ \text{a.s.} \ (16)$$

when $n \to \infty$.

(Proof) See Appendix A.

From Theorem 1, we have the following corollary which gives the minimum coding rate of the ergodic word-valued sources.

**Corollary 1**: For a prefix word set $W$, let the probability distribution $P_{W^n}(W^n)$ be a stationary ergodic with respect to $m$. If $H(W) < \infty$ and $E[|W|] < \infty$,

$$\lim_{n \to \infty} \left[ -\frac{1}{n} \log P_{X^n}(X_1 X_2 \cdots X_{n-1}) \right] = H(X) = \frac{H(W)}{E[|W|]}, \ \text{a.s.} \ (17)$$

for $\forall i \in \{1, 2, 3, \cdots\}$.

(Proof) See Appendix B.

3.2 Discussion

From the theorem, if the entropy rate of the word sequence, $H(W)$, and the expected word length, $E[|W|]$, are given, then the entropy rate of $X$ is given by $H(X) = \frac{H(W)}{E[|W|]}$. The qualitative explanation of the result is as follows: A word sequence $W$ can be compressed by $H(W)$ per a word and the expected number of symbols over $X$ concatenated in a word is given by $E[|W|]$. Therefore, a source sequence $X$ can be compressed by $H(X) = \frac{H(W)}{E[|W|]}$. Corollary 1 means that the lower bound on the compression rate of the ergodic word-valued source is also given by $H(X) = \frac{H(W)}{E[|W|]}$ regardless of the time point we begin to compress a data sequence.

Nishiara and Morita [16] showed the AEP of the i.i.d. word-valued source. Theorem 1 is a general version of their result [16]. However, Nishiara and Morita discussed the asymptotic properties of the i.i.d. word-valued source with a non-prefix word set. The discussion about the asymptotic properties in the case of a non-prefix word set is future work.

4. Main Result II: Estimation of Entropy Rate by a Recurrence Time Theorem

In practical cases, adaptive methods of data compression are useful [3].
Because a probability structure may be unknown in practice, we must estimate the probability structure by a past data sequence to use a word-valued source model for various practical problems. Then, we can consider the estimation problems for the word-valued source in this section. If the probability structure of a word-valued source can be estimated by a past data sequence, we can construct a universal code for it.

If a word set is known but the probability over the word set is previously unknown, then the problem is reduced to usual estimation of ergodic source. If a word set is previously unknown, the estimation using the probability model is computationally difficult in the following meaning:

i) The probability structure must be estimated by the word unit. Even if we can previously fix the upper limit of the maximum length of words for the unknown target word set, the number of considerable word set is an exponential order with respect to the maximum length of words.

ii) If we assume the cases such that the two candidate word sets \( W_1 \) and \( W_2 \) are applied to a data sequence \( X \), the gaps between words do not synchronize between the two word sets.

In [21], adaptive methods are introduced for settings such that the source distribution is known to be stationary and ergodic, but no other information is available. In that way, the time between events, called recurrence time, is essential. Then we consider the estimation of the entropy rate based on the recurrence time which is a foundation of Ziv-Lempel code [21]. If we can estimate the entropy rate of a word-valued source by observing a recurrence time similar to Ziv-Lempel code, we can avoid the problems i) and ii).

Again we use a notation as \( X_i = X_iX_{i+1} \cdots X_{j} \), \( Y_i = Y_iY_{i+1} \cdots Y_{j} \), and \( W_i = W_iW_{i+1} \cdots W_{j} \) for \( j \geq i \) and \( i, j \in \mathbb{Z} \), where \( \mathbb{Z} = \{ -1, 0, 1, 2, \cdots \} \). In this section, in order to consider the recurrence time, we redefine the infinite source sequence by

\[
X = \cdots X_{-2}X_{-1}X_0X_1X_2 \cdots .
\]  

Let \( Y = \cdots Y_{-2}Y_{-1}Y_0Y_1Y_2 \cdots \) be an ergodic source with a finite alphabet \( \mathcal{Y} \). In Sect.3, we let \( \mathcal{Y} \) be a countable alphabet. However we restrict \( \mathcal{Y} \) to be a finite alphabet in order to discuss universality. Similarly to Sect.3, the mapping \( \phi: \mathcal{Y} \rightarrow \mathcal{W} \) is denoted by \( W_i = \phi(Y_i) \) for \( i \in \mathbb{Z} \). Concatenating the sequence \( \cdots , \phi(Y_{-2}), \phi(Y_{-1}), \phi(Y_0), \phi(Y_1), \phi(Y_2), \cdots \), we define the sequence \( \cdots , W_{-2}W_{-1}W_0W_1W_2 \cdots = \cdots , \phi(Y_{-2})\phi(Y_{-1})\phi(Y_0)\phi(Y_1)\phi(Y_2) \cdots \) which is denoted by \( \phi(Y) \). We also denote as \( W = \cdots W_{-2}W_{-1}W_0W_1W_2 \cdots \phi(Y) \). The source sequence \( X \) which is the object of source coding is defined by rewriting \( W \) using symbols in \( \mathcal{X} \) and \( X_1X_2X_3 \cdots = W_1W_2 \cdots \). That is, the gap between \( W_0 \) and \( W_1 \) is located between \( X_0 \) and \( X_1 \).

### 4.1 Main Theorem

Considering the data sequence \( X_{i+1} = X_iX_{i+1}X_{i+2} \cdots \) with length \( l \) for \( i \in \mathbb{Z} \), we define \( N_i \) to be the time of the first recurrence of \( X_{i+1} \). That is, \( N_i \) is the smallest integer \( N \in \mathbb{Z}^+ = \{ 1, 2, \cdots \} \) such that \( X_{i+1} = X_{i+1+N} \).

\[
N_i = \min \{ N \geq 1 | X_{i+1} = X_{i+1+N} \} \tag{19}
\]

For the conventional ergodic source, only the case of \( i = 0 \) was considered [17], [19], [21]. This is because a source sequence is emitted by the symbol unit in \( \mathcal{X} \) with a stationary probability and \( i \) is meaningless for this source model. However, when a word set is unknown, the gaps between words in a data sequence \( x_\infty = x_i x_{i+1} x_{i+2} \cdots \) cannot be found from only \( x_\infty \). Therefore, we must construct a universal code which works with no problem for the sequence which does not necessarily start at a gap between words.

Then, we can show the following theorem which is a generalized version of the conventional recurrence time theorem [17], [19], [21]. The conventional recurrence time theorem was shown for the class of the ergodic sources. The following theorem says that the recurrence time theorem is satisfied for a more broad model class.

**Theorem 2** (A generalized recurrence time theorem): Let \( \mathcal{W} \) be a finite prefix word set. Let the probability distribution \( P_{W^m}(W^m) \) over \( \mathcal{W} \) be stationary and ergodic with respect to \( m \). Assuming \( P_{W}(W_1 = w) > 0 \) for \( \forall w \in \mathcal{W} \), we have

\[
\lim_{l \to \infty} \frac{\log N_i}{l} = \frac{H(W)}{E[|W|]} = H(X), \quad a.s.
\]  

for \( \forall i \in \mathbb{Z} \).

**Proof** See Appendix C.

Although \( N_i \) in Theorem 2 is the time of the first recurrence in a future sequence, the following reverse variable is useful for source coding:

\[
\hat{N}_i = \min \{ N \geq 1 | X_{i+1} = X_{i+N+1} \} \tag{21}
\]

for \( \forall i \in \mathbb{Z} \).

Then the following theorem obviously holds from Theorem 2.

**Theorem 3**: Let \( \mathcal{W} \) be a finite prefix word set. Let the probability distribution \( P_{W^m}(W^m) \) over \( \mathcal{W} \) be stationary and ergodic with respect to \( m \). Assuming \( P_{W}(W_1 = w) > 0 \) for \( \forall w \in \mathcal{W} \), we have

\[
\lim_{l \to \infty} \frac{\log \hat{N}_i}{l} = \frac{H(W)}{E[|W|]} = H(X), \quad a.s.
\]  

for \( \forall i \in \mathbb{Z} \). 

\( \square \)
4.2 Universal Code for Word-Valued Sources

From this theorem, we can construct a FV-type universal code for the word-valued sources, which is a simplified variant of the Ziv-Lempel code.

Let \( \tilde{N}_i(X_{i+1}^t) \) be the smallest integer \( N \geq 1 \) such that \( X_{i+1}^t = X_i^{i+N-1} \). We append the integer encoding of the pointer \( \tilde{N}_i(X_{i+1}^t) \) in order to encode \( X_{i+1}^t \). Then we can encode \( X_{i+1}^t \) with codelength

\[
L(X_{i+1}^t | X_i) = \log \tilde{N}_i(X_{i+1}^t) + O(\log \log \tilde{N}_i(X_{i+1}^t)).
\]

For example, if we use the Elias code \( \omega^* \), then its codelength is upper bounded by \( L \leq \sum_{i=1}^{\infty} \log \tilde{N}_i(X_{i+1}^t) + 7 \).

Theorem 3 says that the above code is asymptotically optimal for ergodic word-valued sources.

When the true word set is unknown, the above FV-code does not generally synchronize with the corresponding word sequence. That is, the gaps between encoded blocks in \( x \) do not generally correspond to those between words. However, because an integer \( i \) in the above algorithm is arbitrary, even if we cannot specify the gaps between words in a data sequence to be encoded, the algorithm is always universal for the ergodic word-valued sources. That is, the pattern matching algorithm which is a basis of the Ziv-Lempel code is effective not only for stationary and ergodic sources but also for word-valued sources.

4.3 Discussions

From the viewpoint of asymptotic property, the length of a source sequence \( X^n \) emitted from the source is the \( E[|W|] \) times of that of the word sequence \( W^m \) which corresponds to \( X^n \). The recurrence time in \( X \) is also the \( E[|W|] \) times of that in the word sequence \( W \).

If we can assume a good parametric model class for unknown sources, we can use the Laplace estimator to estimate. When the good probability model class cannot be assumed, the estimation using the recurrence time is very useful. Of course, when the suitable parametric model can be set for the unknown source, its performance of the estimation would be better for a finite data size.

5. Conclusion

In this paper, we propose a new source model class, called an ergodic word-valued source, and show the time recurrence theorem. As a result, we show that the Ziv-Lempel 77 code is universal for this model class. Analysis of convergence speed of the universal coding for the proposed model class and relation the ergodic word-valued and AMS sources will be future work.

Acknowledgement

The authors wish to acknowledge M. Nishiara and H. Morita for their useful comments and introduction of their papers. One of the authors, M. Goto, wish to thank M. Kobayashi, T. Ishida, and all member of Hirasawa lab. and Matsushima lab, for their helpful discussion for this paper.

References


Appendix A: Proof of Theorem 1

Since $W$ is stationary and ergodic, when $m \to \infty$

$$-\frac{1}{m} \log P_{W^m}(W_1 W_2 \cdots W_m) \to H(W), \quad a.s. \quad (A.1)$$

holds from the ergodic theorem (Shannon-McMillan-Breiman’s theorem [1], [7]).

For convenience, we define $N_m \overset{\text{def}}{=} N(W^m)$. That is

$$N_m \overset{\text{def}}{=} N(W^m) = \sum_{i=1}^{m} L_i,$$

and $N_m$ is a random variable depending on $W^m$. Rewriting the word sequence $W_1 W_2 \cdots W_m$ for $\forall m \in \mathbb{Z}^+$, we have

$$W_1 W_2 W_3 \cdots W_m = X_1 X_2 X_3 \cdots X_{N_m}. \quad (A.2)$$

From (A.1), when $m \to \infty$

$$-\frac{1}{m} \log P_{X^{N_m}}(X_1 X_2 \cdots X_{N_m}) \to H(W), \quad a.s. \quad (A.3)$$

holds. This means

$$-\frac{N_m}{m} \log P_{X^{N_m}}(X_1 X_2 \cdots X_{N_m}) \to H(W), \quad a.s. \quad (A.4)$$

when $m \to \infty$.

On the other hand, since $W$ is stationary and ergodic, when $m \to \infty$

$$\frac{N_m}{m} \to E[|W|], \quad a.s. \quad (A.5)$$

from the ergodic theorem.

Consider $X^n$ to the contrary for an arbitrary $n$ ($n = 1, 2, \cdots$). Let $M_n$ be the minimum length of $W^m$ such that $N_m \geq n$ for given $n \in \mathbb{Z}^+$. That is,

$$M_n \overset{\text{def}}{=} \min_{m \geq 1} \{ m | N_m \geq n \}.$$ 

Then

$$N_{M_n-1} < n \leq N_{M_n} < n + L_{M_n}. \quad (A.6)$$

There exists some $n \in \mathbb{Z}^+$ satisfying $N_m < n$ for all $m \in \mathbb{Z}^+$ because $N_{M_n} < n$ for $\forall m \in \mathbb{Z}^+$. This means that $m$ satisfying $N_m < n$ can take an arbitrarily large value if $n \to \infty$. Because $N_m < n$ is equivalent to $M_n > m$, we have

$$\lim_{n \to \infty} M_n = \infty, \quad (A.7)$$

for all sample sequences. The above convergence was shown by Nishiara and Morita in [16]. We have therefore

$$-\frac{1}{M_n} \log P_{X^{N_{M_n}}}(X_1 X_2 \cdots X_{N_{M_n}}) \to H(W), \quad a.s. \quad (A.8)$$

when $n \to \infty$ from (A.3) and (A.7).

From the definition, the inequality $N_{M_n} - L_{M_n} \leq n \leq N_{M_n}$ is satisfied $\forall n \in \mathbb{Z}^+$. We have therefore

$$\frac{M_n}{N_{M_n}} \leq \frac{n}{N_{M_n} - L_{M_n}}, \quad (A.9)$$

where $\forall n \in \mathbb{Z}^+$ for every sample sequence. Here we have

$$\frac{M_n}{N_{M_n}} \to \frac{1}{E[|W|]}, \quad a.s. \quad (A.10)$$

when $n \to \infty$. This is because $M_n \to \infty$ when $n \to \infty$ for every sample sequence and $\lim_{m \to \infty} \frac{m}{N_m} = E[|W|], \quad a.s.$ On the other hand, we have

$$\frac{M_n}{N_{M_n} - L_{M_n}} = \frac{M_n}{N_{M_n}} \cdot \frac{N_{M_n} - L_{M_n}}{N_{M_n} - L_{M_n}} = \frac{M_n}{N_{M_n} 1 - \frac{L_{M_n}}{N_{M_n}}}. \quad (A.11)$$

Here

$$\frac{L_{M_n}}{M_n} \to 0, \quad a.s. \quad (A.12)$$

when $m \to \infty$ because $\lim_{m \to \infty} \frac{m}{N_m} = E[|W|], \quad a.s.$ and $N_{M_n} = \sum_{i=1}^{m} L_i$. Because $N_m \geq m$, we have

$$\frac{L_m}{N_m} \to 0, \quad a.s. \quad (A.13)$$

when $m \to \infty$. Since $M_n \to \infty$ when $n \to \infty$, (A.13) means

$$\frac{1}{1 - \frac{L_m}{N_m}} \to 1, \quad a.s. \quad (A.14)$$

when $n \to \infty$. We have therefore
when $n \to \infty$. From (A·9), (A·10), and (A·15), we have
\[
\frac{M_n}{n} \to \frac{1}{E[|W|]}, \quad a.s. \quad (A·16)
\]
when $n \to \infty$.

At last, we shall complete the proof. From $N_{M_{n-1}} < n \leq N_{M_n}$, we have
\[
-\frac{1}{n} \log P_{X^{N_{M_{n-1}}}}(X^{N_{M_{n-1}}}) \\
\leq -\frac{1}{n} \log P_{X^{N_{M_n}}}(X^{N_{M_n}}).
\]
Here we have
\[
-\frac{1}{n} \log P_{X^{N_{M_n}}}(X^{N_{M_n}}) \\
= -\frac{M_n}{n} \log P_{X^{N_{M_n}}}(X^{N_{M_n}}) \\
\to \frac{H(W)}{E[|W|]}, \quad a.s. \quad (A·18)
\]
when $n \to \infty$ from (A·8) and (A·16). On the other hand,
\[
-\frac{1}{n} \log P_{X^{N_{M_{n-1}}}}(X^{N_{M_{n-1}}}) \\
= -\frac{M_n}{n} \frac{M_n - 1}{M_n - 1} \log P_{X^{N_{M_{n-1}}}}(X^{N_{M_{n-1}}}) \\
\to \frac{H(W)}{E[|W|]}, \quad a.s. \quad (A·19)
\]
when $n \to \infty$ from (A·8), (A·16), and $\frac{M_n - 1}{M_n} \to 1$ when $n \to \infty$.

We have therefore
\[
-\frac{1}{n} \log P_{X^*}(X_1X_2 \cdots X_n) \to \frac{H(W)}{E[|W|]}, \quad a.s. \quad (A·20)
\]
when $n \to \infty$. This is the last half of the theorem.

From the bounded convergence theorem [6], we have
\[
-\frac{1}{n} E[\log P_{X^*}(X_1X_2 \cdots X_n)] \to \frac{H(W)}{E[|W|]}.
\]
This means
\[
H(X) = \frac{H(W)}{E[|W|]}, \quad (A·22)
\]
\[\square\]

### Appendix B: Proof of Corollary 1

Because the proof of Corollary 1 is similar with that of Theorem 1, we give an outline roughly.

From the identical discussion with (A·7) in the proof of Theorem 1, we have
\[
M_{k+n} \to \infty \quad (A·23)
\]
when $n \to \infty$ for all fixed $k \in \mathbb{Z}^+$ and all sample sequences. Therefore, we have
\[
\frac{1}{M_{k+n} - M_{k+1}} \log P_{X^{N_{M_{k+n}}}}(X^{N_{M_{k+n}}}) \\
\to H(W), \quad a.s. \quad (A·24)
\]
and
\[
\frac{M_n - M_{k+1}}{N_{M_n} - N_{M_{k+1}}} \to \frac{1}{E[|W|]}, \quad a.s. \quad (A·25)
\]
when $n \to \infty$ for all fixed $k \in \mathbb{Z}^+$ from (A·23) and the ergodic theorem.

On the other hand, from the definition of $M_n$, we have
\[
N_{M_{i+n-1}} < i + n - 1 \leq N_{M_{i+n-1}} \quad (A·26)
\]
and
\[
N_{M_{1+n}} + 1 \leq i \leq N_{M_{1+n}} \quad (A·27)
\]
for all $i, n \in \mathbb{Z}^+$. We have therefore
\[
-\frac{1}{n} \log P_{X^{N_{M_{1+n}}}}(X^{N_{M_{1+n}}}) \\
\leq -\frac{1}{n} \log P_{X^{N_{M_{1+n}}}}(X^{N_{M_{1+n}}}) \\
\leq -\frac{1}{n} \log P_{X^{N_{M_{1+n}}}}(X^{N_{M_{1+n}}}) \quad (A·28)
\]

Using (A·23) and (A·24), we can show that the right and left sides of (A·28) converge to $\frac{H(W)}{E[|W|]}$ almost surely, then the proof is complete. \[\square\]

### Appendix C: Proof of Theorem 2

At first, consider a word sequence $W_{j+j'} = W_{j+1}\cdots W_{j+j'}$ with length $l'$ for $\forall j \in \mathbb{Z}$. Letting $M_l'$ be a first recurrence time of $W_{j+j'}$ measured per word unit which is given by
\[
M_l' = \min\{N \geq 1 | W_{1+j'} = W_{1+j'+N} \}, \quad (A·29)
\]
we have
\[
\lim_{l' \to \infty} \frac{\log M_l'}{l'} = H(W), \quad a.s. \quad (A·30)
\]
for $\forall j \in \mathbb{Z}$ from the recurrence time theorem
which is shown in [21], p. 2046, because \(W_{j+1}^{j+\ell'} = W_{j+1}W_{j+2} \cdots W_{j+\ell'}\) is a stationary, ergodic, finite-alphabet source.

Let \(L = L(W_{j+1}^{j+\ell'}) = L_{j+1} + L_{j+2} + \cdots + L_{j+\ell'}\). Since \(W\) is a stationary, ergodic, finite-alphabet source, when \(\ell' \to \infty\) we have
\[
\frac{L}{\ell'} = \frac{L_{j+1} + L_{j+2} + \cdots + L_{j+\ell'}}{\ell'} \to E[|W|], \quad \text{a.s.} \tag{A-31}
\]
for \(\forall j \in \mathbb{Z}\).

On the other hand, we consider the recurrence time of \(X_{i+1}^{l+1}\) measured by the symbol unit in \(X\). We rewrite \(X_{i}X_{i+1}X_{i+2} \cdots X_{i+1}X_{i+2} \cdots \) by \(W_{j}W_{j+1}W_{j+2} \cdots W_{j+\ell}W_{j+\ell+1}W_{j+\ell+2} \cdots\), where the word sequence \(W_{j+\ell}^{j+\ell'}\) includes \(X_{i+1}^{l+1}\) in its interior. That is, there exist some \(\alpha\) and \(\beta\) \((\alpha, \beta \in \{0, 1, 2, \cdots\})\) such that
\[
W_{j+\ell}^{j+\ell'} = X_{i+1}^{\alpha+1+j}X_{i+1}^{\alpha+2}X_{i+1}^{\alpha+3} \cdots X_{i+1}^{\alpha+1+\ell}X_{i+1}^{\alpha+2+\ell}X_{i+1}^{\alpha+3+\ell}X_{i+1}^{\alpha+4+\ell},
\]
where we can set that \(\alpha\) and \(\beta\) are bounded because the word set is finite. (For example, \(X_{i+1}^{l+1} = 0001110\) is included in \(W_{j+1}^{j+\ell'} = 010001110111\) where \(l = 7, \ell' = 3, w_{j+1} = 0100, w_{j+2} = 011,\) and \(w_{j+3} = 101\). \(\alpha = 2\) and \(\beta = 3\) in this case.) That is, \(L = L_{j+1} + L_{j+2} + \cdots + L_{j+\ell'}\) is \(\ell + \alpha + \beta \geq 1\), and the correspondence from \(X_{i+1}^{l+1}\) to \(W_{j+\ell}^{j+\ell'}\) is not unique.

Letting \(N_W = N(W_{j+1}^{j+\ell'}\) \(= L_{j+1} + L_{j+2} + \cdots + L_{j+\ell'})\), we have
\[
X_{i+1}X_{i+2} \cdots X_{i+\ell} = X_{i+N_W}X_{i+N_W+1}X_{i+N_W+2} \cdots X_{i+N_W+\ell},
\]
become \(W_{i+1}^{i+j} = W_{i+1}^{i+j+M_i} 1_{i+1+j+M_i}\) and \(W_{i+1}^{i+j'} = W_{i+1}^{i+j'+M_i} 1_{i+1+j'+M_i}\) includes \(X_{i+1}^{l+1}\) in its interior.

However, the recurrence time of \(X_{i+1}X_{i+2} \cdots X_{i+\ell}\) in \(X\) measured per symbol unit of \(X\) may be smaller than \(N_W\). This is because \(M_i\) is a first recurrence time of \(W_{j+1}^{i+j}\) measured per word unit, that is all of the subsequences \(X_{i+N_W+1}^{i+j}, n \in \{1, 2, \cdots, N\}\) which contain subsequences straddled words are not considered to find \(M_i\). That is, the recurrence time of \(X_{i}X_{i+1}X_{i+2} \cdots X_{i+l}\) measured per symbol unit of \(X\) can be selected with no relation to gaps between words. Therefore, letting the recurrence time of \(X_{i+1}X_{i+2} \cdots X_{i+l}\) measured per symbol unit of \(X\) be \(N_i\),
\[
\frac{N_i}{M_i} \leq \frac{N_W}{M_i} = \frac{L_{j+1} + L_{j+2} + \cdots + L_{j+M_i}}{M_i} \to E[|W|], \quad \text{a.s.} \tag{A-34}
\]
from (A-31).

Therefore, since \(\ell' \to \infty\) when \(\ell \to \infty\), we have
\[
\frac{\log N_i}{l} \leq \frac{\log N_W}{l} = \frac{\log M_i + \log \frac{N_W}{M_i}}{l} = \frac{L - \alpha - \beta}{l} \left\{ \log M_i + \log \frac{N_W}{M_i} \right\}
\]
\[
\to H(W) = H(X), \quad \text{a.s.} \tag{A-35}
\]
for \(\forall i \in \mathbb{Z}\). Here, the convergence (A-35) holds because \(\frac{N_W}{M_i} \to E[|W|] < \infty, \text{a.s.}, \frac{\log M_i}{l} \to H(W), \text{a.s.}\), and \(\alpha\) and \(\beta\) are bounded and (A-31) holds. (A-35) means
\[
\limsup_{\ell \to \infty} \frac{\log N_i}{\ell} \leq H(X), \quad \text{a.s.} \tag{A-36}
\]

Because \(\limsup_{\ell \to \infty} \frac{\log N_i}{\ell} \leq H(X) \quad \text{a.s.}\) from the above discussion, then the code exists whose mean codelength is less than or equal to \(H(X)\) [8], [21]. Precisely, we can construct a code whose codelength satisfies \(\frac{\log N_i}{l} + o(1)\) when \(l \to \infty\). How to construct such code is shown in Sect. 4.2.

Here, let \(l(P_n)\) be the codelength of a variable length noiseless code \(\{\phi_n, \phi_n^{-1}\}\). Then, we can find a noiseless code \(\{\phi_n, \phi_n^{-1}\}\) satisfying
\[
\limsup_{n \to \infty} \frac{1}{n} l(P_n) \leq H(X), \quad \text{a.s.} \tag{A-37}
\]
from (A-36). Oppositely, for a general source
\[
l(P_n) \geq - \log P_X(x^n) - \log n - 2 \log \log n, \quad \text{a.s.} \tag{A-38}
\]
is satisfied [2], [15]. Therefore we have\(^\dagger\)
\[
\liminf_{n \to \infty} \frac{1}{n} l(P_n) \geq H(X), \quad \text{a.s.} \tag{A-39}
\]
This also means that
\[
\liminf_{l \to \infty} \frac{\log N_l}{l} \geq H(X) \quad \text{a.s.} \tag{A-40}
\]
because we can construct a code whose codelength\(^\dagger\)

\(^\dagger\)If the code \(\{\phi_n, \phi_n^{-1}\}\) satisfies (A-37), then
\[
\lim_{n \to \infty} \frac{1}{n} l(P_n) = H(X), \quad \text{a.s.}
\]
for ergodic word valued sources from (A-39).
\( l(\phi_n) \) satisfies \( l(\phi_n) = \frac{\log N_i}{n} + o(1) \) when \( n \to \infty \) (See Sect. 4.2).

We have therefore

\[
\lim_{l \to \infty} \frac{\log N_i}{l} \to H(X), \ a.s.
\]

for \( \forall i \in \mathcal{Z} \) from the equations (A·36) and (A·40), the proof is complete. \( \square \)

Masayuki Goto was born in Tokyo, Japan, on Jan. 1, 1969. He received his B.E. and M.E. degrees from Musashi Institute of Technology, Tokyo, Japan, in 1992 and 1994, respectively. He received Dr.E. degree in Industrial Engineering and Management from Waseda University, Tokyo, Japan, in 2000. From 2000 to 2002, he was a research associate in Environmental and Ocean Engineering at the University of Tokyo, Tokyo, Japan. He is now an associate professor at Musashi Institute of Technology, Tokyo, Japan. His research interests include information theory, machine learning theory, model selection, Bayesian statistics, and Industrial Engineering. He is a member of the Society of Information Theory and Its Applications, the Japan Industrial Management Association, Business Model Association, the Operations Research Society of Japan and the Japan Society for Artificial Intelligence.

Toshiyasu Matsushima was born in Tokyo, Japan, on Nov. 26, 1955. He received the B.E. degree, M.E. degree and Dr.E. degree in Industrial Engineering and Management from Waseda University, Tokyo, Japan, in 1978, 1980 and 1991, respectively. From 1980 to 1986, he was with the Nippon Electric Corporation, Kanagawa, Japan. From 1986 to 1992, he was a lecture to the Department of Management Information, Yokohama College of Commerce. From 1993, he was an associate professor and since 1996 has been a professor of School of Science and Engineering, Waseda University, Tokyo, Japan. His research interests are information theory and its application, statistics and artificial intelligence. He is a member of the Society of Information Theory and Its Applications, the Japan Society for Quality Control, the Japan Industrial Management Association, the Japan Society for Artificial Intelligence, and IEEE.

Shigeichi Hirasawa was born in Kobe, Japan, on Oct. 2, 1938. He received the B.S. degree in mathematics and the B.E. degree in electrical communication engineering from Waseda University, Tokyo, Japan, 1961 and 1963, respectively, and the Dr.E. degree in electrical communication engineering from Osaka University, Osaka, Japan, in 1975. From 1963 to 1981, he was with the Mitsubishi Electric Corporation, Hyogo, Japan. Since 1981, he has been a professor of School of Science and Engineering, Waseda University, Tokyo, Japan. In 1979, he was a Visiting Researcher in the Computer Science Department at the University of California, Los Angeles, CA. He was a Visiting Researcher at the Hungarian Academy of Science, Hungary, in 1985, and at the University of Trieste, Italy, in 1986. In 2002, he was also a Visiting Faculty at CSD, UCLA. From 1987 to 1989, he was the Chairman of Technical Group on Information Theory of IEICE. He received the 1993 Achievement Award, and the 1993 Kobyashi-Memorial Achievement Award from IEICE. In 1996, he was the President of the Society of Information Theory and Its Applications (Soc. of ITA). His research interests are information theory and its applications, and information processing systems. He is an IEEE Fellow, and a member of Soc. of ITA, the Operations Research Society of Japan, the Information Processing Society of Japan, the Japan Industrial Management Association, and Informs.