SUMMARY

The coding rate of a one-shot Tunstall code for stationary and memoryless sources is investigated in non-universal situations so that the probability distribution of the source is known to the encoder and the decoder. When studying the variable-to-fixed length code, the average coding rate has been defined as (i) the codeword length divided by the average block length. We define the average coding rate as (ii) the expectation of the pointwise coding rate, and prove that (ii) converges to the same value as (i).

**key words:** lossless data compression, VF code, Tunstall code, definition of average coding rate, pointwise redundancy rate

1. Introduction

In this paper, we investigate the coding rate of the Tunstall code [1], which is a variable-to-fixed length (VF) lossless code for stationary and memoryless sources with a finite alphabet.

First we discuss the definition of the average coding rate of the variable-to-variable length (VV) lossless code with $X$ as the source alphabet and $B$ as the code alphabet. The union of all finite products of $X$ and $B$ are written as $X^+ \equiv \cup_{k \in \mathbb{N}} X^k$ and $B^+ \equiv \cup_{k \in \mathbb{N}} B^k$, where $\mathbb{N}$ represents the set of positive integers and $X^k$ means the $n$-th product set of $X$. Subsequently, $X_k^* = \{x^* \} \subseteq X^+$ is a message set, and $X_k^*$ is a random variable which takes values in $X_k^*$ with the probability distribution $P_k$. The length of a finite string $s$ is denoted by $|s|$, and the cardinality of a finite set $S$ is denoted by $|S|$. The bases of log and exp are assumed to be 2 in this paper.

Subsequently, a one-shot lossless VV code is defined by the sequence of the pairs $\{(\varphi_k, \psi_k)\}_{k \in \mathbb{N}}$, where $\varphi_k : X_k^* \rightarrow B^k$ is an encoder and $\psi_k : B^k \rightarrow X_k^*$ is a decoder. Each $x^* \in X_k^*$ of the length $|x^*|$ is encoded to the codeword $\varphi_k(x^*)$ of the length $|\varphi_k(x^*)|$. The average coding rate of this code can be defined in two ways, which are

$$E\left[\frac{|\varphi_k(X_k^*)|}{|X_k^*|}\right] \quad \text{and} \quad \frac{E[|\varphi_k(X_k^*)|]}{E[|X_k^*|]},$$

where $E[Y]$ signifies the expectation of a random variable $Y$. We are interested in the limit of these average coding rates for the case where the message-set size satisfies $\lim_{k \rightarrow \infty} |X_k^*| = \infty$ (the parameter $k$ refers to the extension step of the message set).

We treat the fixed-to-variable length (FV) code and the VF code as special cases of the VV code. For the FV code, the two definitions of the average coding rate of the VV code have no significance since the denominator (block length) is constant. On the other hand, for the VF code, the message set $X_k^*$ is mapped to $\varphi_k(X_k^* \equiv \{\varphi_k(x^* : x^* \in X_k^*\}$, which is encoded in $\log |X_k^*|$ bits (note that $|\varphi_k(X_k^*)| = |X_k^*|$ since the code is lossless). Subsequently, the two definitions of the average coding rate are rewritten as

$$\log |X_k^*| \times \frac{1}{|X_k^*|}$$

and

$$\frac{\log |X_k^*|}{E[|X_k^*|]},$$

respectively, which may result in different achievable coding rates.

When studying VF codes [1]–[10], definition (2) has been used. We can give a reasonable interpretation of this definition for the multi-shot VF coding, as follows. If we were to parse a given source sequence to $j$ words and encode them, the total codeword length is $j \log |X_k^*|$, and the expectation of the total block length is equal to $j E[|X_k^*|]$. Therefore, the total average coding rate can be defined as (2). Moreover, the Tunstall code [1] is an optimal code in that it attains the maximum average block length in the code class of the same message-set size; thus, this code attains the optimal coding rate when we apply definition (2).

On the other hand, since (1) is the expectation of the pointwise coding rate, this definition has an advantage in that we can discuss the convergence in probability or almost sure convergence of the pointwise coding rate with (1). Despite having such merits, we do not know whether (1) converges to the entropy of the source, even after the convergence of (2) to the entropy of the source has been confirmed. This is due to the fact that (1) is larger than (2) as a result of Jensen’s inequality.

Therefore, in this paper, we evaluate the asymptotic coding rate of the Tunstall code using definition (1). We prove that in the one-shot Tunstall coding of stationary and memoryless sources, the average coding rate (1) converges to the same value as the limit of (2), which was proved to be the entropy of the source. Therefore, our results confirm a connection between the pointwise coding/redundancy rate...
and the average coding rate defined by (2).

This paper is organized as follows. In Sect. 2, we give the definitions of the information source and the Tunstall algorithm. Section 3 reviews the evaluation of the pointwise redundancy rate [2] of the Tunstall code. Using this result, we prove the convergence of the coding rate of the Tunstall code for stationary and memoryless sources in Sect. 4. To obtain these results, we use definition (1) as the average coding rate. In Sect. 5, definition (1) is compared with definition (2), and we conclude that these two average coding rates converge to the entropy of the source.

2. Tunstall Code for Stationary and Memoryless Sources

For any random variable $X$, the probability distribution of $X$ is denoted by $P_X$. The sequences $x_1, x_2, x_3, \ldots, x_n$ and $(X_1, X_2, \ldots, X_n)$ are denoted by $x^n$ and $X^n$, and $x'_i$ represents a substring $x_{i+1}, \ldots, x_j$ of $x^n$ ($1 \leq i \leq j \leq n$).

Next we define an information source. Let $X = \{X^n\}_{n \in \mathbb{N}}$ be a stationary and memoryless information source, where $X^n$ is a sequence of independent and identically distributed random variables. Each $X_i$ takes values in a finite alphabet $X$ ($|X| = A < \infty$) with probability distribution given as $P_{X_i} = P_X$. The entropy rate $H(X)$ of the source $X$ is

$$H(X) = H(P_X) \defeq -\sum_{a \in X} P_X(a) \log P_X(a).$$

The maximum and minimum probabilities of $P_X$ are written as

$$P = \max_{a \in X} P_X(a), \quad p = \min_{a \in X} P_X(a).$$

Subsequently, we define a parsing tree and a probability distribution on the leaf set of the tree. Let $T$ be a complete $A$-ary tree of depth $1$. Each edge has a label $\nu$, which is defined by $\nu \in \{X^n\}_{n \in \mathbb{N}}$ be a stationary and memoryless information source, where $X^n$ is a sequence of independent and identically distributed random variables. Each $X_i$ takes values in a finite alphabet $X$ ($|X| = A < \infty$) with probability distribution given as $P_{X_i} = P_X$. The entropy rate $H(X)$ of the source $X$ is

$$H(X) = H(P_X) \defeq -\sum_{a \in X} P_X(a) \log P_X(a).$$

The maximum and minimum probabilities of $P_X$ are written as

$$P = \max_{a \in X} P_X(a), \quad p = \min_{a \in X} P_X(a).$$

Subsequently, we define a parsing tree and a probability distribution on the leaf set of the tree. Let $T$ be a complete $A$-ary tree such that each inner node of $T$ has $A$ children. Each edge has a label $a \in X$. The leaf set of $T$, which is the set of nodes having no child, is written as $\mathcal{T}$. For each leaf node $v \in \mathcal{T}$, let $x(v)$ be the label of $v$, which is a concatenation of the labels of the edges from the root node to $v$. Then the depth of the leaf $v$ can be represented by $\|x(v)\|$. The maximum and minimum depths of the leaves are defined as

$$\overline{w}(T) \defeq \max_{v \in \mathcal{T}} \|x(v)\|, \quad \underline{w}(T) \defeq \min_{v \in \mathcal{T}} \|x(v)\|.$$

We now define a function of a leaf node $v \in \mathcal{T}$ by the probability of the label $x(v)$ as

$$P^v_T(v) \defeq P_{X^{\underline{w}(T)}}(x(v)).$$

Since the set of the labels $\{x(v) : v \in \mathcal{T}\}$ is proper and complete (which means that for any infinite-length string $x^\infty = x_1 x_2 \cdots \in X^\infty$, there exists just one leaf $v \in \mathcal{T}$ such that $x(v)$ is a prefix of $x^\infty$), and the probability distributions $\{P^v_T\}_{v \in \mathcal{T}}$ satisfy the consistency condition

$$\sum_{a \in X} P^{x(a)}_T(x) = P^v_T(x),$$

we have

$$\sum_{v \in \mathcal{T}} P^v_T(v) = 1$$

for any leaf set $\mathcal{T}$. Therefore, we can use $P^v_T$ as a probability distribution on the leaf set $\mathcal{T}$. Let $N_T$ be a random variable which takes values in $\mathcal{T}$ with $P^v_T$. For a tree $T$, the maximum and minimum probabilities of the leaves are denoted by

$$\overline{P}(T) \defeq \max_{v \in \mathcal{T}} P^v_T(v), \quad \underline{P}(T) \defeq \min_{v \in \mathcal{T}} P^v_T(v).$$

Here we describe the encoding algorithm of the Tunstall code. A Tunstall parsing tree is created as follows.

1. At the step 1, let $T_1$ be the complete $A$-ary tree of depth one.
2. At the steps from $i = 2$ to $m$, $T_i$ is created from $T_{i-1}$ as follows (this step is referred to as an extension of the Tunstall tree);
   a. Take any leaf $v \in \mathcal{T}_{i-1}$ which has the largest probability $P^v_{T_{i-1}}(v) = \overline{P}(T_{i-1})$.
   b. Let $T_i$ be the tree created by attaching the root node of $T_{i-1}$ to the leaf $v$.

Since the size of the leaf set $\mathcal{T}_m$ is $|\mathcal{T}_m| = (A - 1)m + 1$, the label $x(v)$ of each leaf $v \in \mathcal{T}_m$ can be encoded in $\lceil \log |\mathcal{T}_m| \rceil$ bits by the fixed-to-fixed length lossless code.

In the usual application of the Tunstall code, we encode a string $x^n$ using the parsing tree $T_m$ as follows. First, we create the parsing tree $T_m$ from the source if we know it, or from the empirical frequency distribution of $x^n$. Next, the string $x^n$ is parsed into the blocks $x_{n_1}^{n_1-1}, x_{n_2}^{n_2-1}, x_{n_3}^{n_3-1}, \ldots, (n_1 = 1)$. At the step $i$, we find the prefix $x_{n_i}^{n_i-1}$ of $x^n$ in the labels of the leaf set $\mathcal{T}_m$, and encode it in $\lceil \log |\mathcal{T}_m| \rceil$ bits. The concatenation of the codewords for all blocks is the codeword for $x^n$.

In the following sections, we investigate the compression performance of the first block $x_{n_1}^{n_1-1}$ as the extension step $m$ becomes large, under the condition where the encoder knows the probability distribution of the information source $X$.

3. Pointwise Redundancy Rate of the Tunstall Code

In this section, we review the result [5] on the pointwise redundancy rate of the Tunstall code.

The pointwise coding rate of the label $x(v)$ of a node $v \in \mathcal{T}_m$ is defined by

$$\frac{1}{\|x(v)\|} \lfloor \log |\mathcal{T}_m| \rfloor,$$

and the self-information rate, or the symbol-wise ideal codeword length of $x(v)$ with respect to the probability $P^v_T$, can
be written by
\[
\frac{1}{\|x(v)\|} \log \frac{1}{P_{T_m}(v)}. \tag{4}
\]
Using these values, we define the pointwise redundancy rate as follows.

**Definition 1:** (Pointwise Redundancy Rate of the Tunstall Code [5]). The pointwise redundancy rate \(r(v, T_m, X)\) of a leaf \(v \in T_m\) of the Tunstall tree \(T_m\), with respect to the source \(X\), is defined by
\[
r(v, T_m, X) \triangleq \frac{1}{\|x(v)\|} \left( \log |T_m| - \log \frac{1}{P_{T_m}(v)} \right).
\]

Consequently, the following theorem holds.

**Theorem 1:** For any step \(m\), the pointwise redundancy rate of the Tunstall code with respect to \(X\) is bounded as follows;
\[
\begin{align*}
\max_{v \in T_m} r(v, T_m, X) &\leq \frac{1 - \log P}{\log |T_m| + \log P} \log \frac{1}{P}, \tag{5} \\
\min_{v \in T_m} r(v, T_m, X) &\geq \frac{-\log P}{\log |T_m| - \log P} \log \frac{1}{P}. \tag{6}
\end{align*}
\]

From this theorem, we can derive the following corollary.

**Corollary 1 ([5]):** For any stationary and memoryless source \(X\),
\[
\begin{align*}
\limsup_{m \to \infty} \max_{v \in T_m} r(v, T_m, X) &\leq 0, \\
\liminf_{m \to \infty} \min_{v \in T_m} r(v, T_m, X) &\geq 0.
\end{align*}
\]

Moreover, the order of the convergence rate of the redundancy rate is
\[
r(v, T_m, X) = \Theta \left( \frac{1}{\log |T_m|} \right).
\]

To prove Theorem 1, we use the following lemmas.

**Lemma 1 ([5]):** For any \(m\), the minimum and the maximum depths of the leaves are bounded as follows;
\[
\begin{align*}
\frac{\log |T_m|}{\log P} &\leq \frac{1}{w(T_m)} \leq \frac{\log P}{\log |T_m|}, \\
\frac{1}{\log P} &\leq \frac{\log |T_m|}{\log w(T_m)} \leq \frac{1}{\log P}.
\end{align*}
\]

These inequalities are equivalent to the following inequalities;
\[
\begin{align*}
\log \frac{P}{w(T_m)} &\leq \log \frac{P}{|T_m|} \leq \log \frac{P}{w(T_m)}, \tag{7} \\
\log \frac{1}{w(T_m)} &\leq \log \frac{1}{|T_m|} \leq \log \frac{1}{w(T_m)}. \tag{8}
\end{align*}
\]

**Lemma 2 ([2]):**
\[
\frac{P'(T_m)}{P(T_m)} \leq \frac{1}{P}.
\]

Using these lemmas, we can bound the maximum depth \(\overline{w}(T_m)\) and the minimum depth \(\underline{w}(T_m)\), respectively, using the leaf-set size \(|T_m|\) of the Tunstall tree. From the inequality
\[
\frac{P'(T_m)}{P(T_m)} \leq \frac{1}{|T_m|} \leq \frac{P'}{P},
\]
and Lemma 2, \(P'(T_m)\) and \(P'(\overline{T}_m)\) are bounded using \(|\overline{T}_m|\) as follows;
\[
\frac{1}{|\overline{T}_m|} \leq \frac{P'(T_m)}{P(T_m)} \leq \frac{1}{|\overline{T}_m|P}, \tag{9}
\]
\[
\frac{P}{|\overline{T}_m|} \leq \frac{P'(T_m)}{P(T_m)} \leq \frac{1}{|\overline{T}_m|}. \tag{10}
\]

Applying (9) and (10) to Lemma 1, we derive the following lemma.

**Lemma 3:** The minimum and maximum depths \(w(T_m)\), \(\overline{w}(T_m)\) of the Tunstall tree \(T_m\) are bounded using the number of the leaves \(|T_m|\) as follows;
\[
\frac{\log |T_m|}{\log P} - 1 \leq w(T_m) \leq \overline{w}(T_m) \leq \frac{\log |T_m|}{\log P} + C(X),
\]
where
\[
C(X) \triangleq \log \frac{P}{\log P}.
\]

From this lemma, we can see that the depth of the Tunstall tree increases in proportion to the logarithm of the number of leaves of the tree, and the ratio of them are bounded by \(-\log P\) and \(-\log P\).

Using these lemmas, Theorem 1 is proved in Appendix A.

## 4. Coding Rate of the Tunstall Code

In the previous section, we have identified the pointwise redundancy rate of the Tunstall code; however, we have yet to establish the relationships between the information source and the pointwise self-information rate defined in (4).

In this section, we show that the self-information rate (4) converges to the entropy \(H(X)\) in probability and in an average sense. Using this result, we show that the coding rate of the Tunstall code converges to the entropy \(H(X)\).

### 4.1 Convergence of the Coding Rate in Probability

In this subsection, we show that the self-information rate, or the symbol-wise ideal codeword length, and the coding rate converge to the entropy rate of the source in probability.

First, we show the following lemma, which is proved in Appendix B.
Lemma 4 (Bernstein): For any stationary and memoryless source $X$ with a finite alphabet of size $A$, it holds that

$$\Pr\left\{ \frac{1}{n} \log P_x(x^n) - H(X) \geq d - \sqrt{2d \log \frac{2d}{A}} \leq \exp\left\{ -n \left( d - \frac{A \log(n + 1)}{n} \right) \right\} \right\}$$

for sufficiently small $d > 0$.

Using this lemma (AEP theorem), we prove the following theorem, which is an AEP theorem for a sequence of the sets of variable-length blocks. Recall that $N_{T_n}$ denotes a random variable which takes values in $\mathcal{T}$ with probability $P_{T_n}$.

Theorem 2: For any stationary and memoryless source $X$,

$$\frac{1}{\|x^*(N_{T_n})\|} \log P_{T_n} - H(X) \rightarrow H(X) \text{ in probability,}$$

as $m$ goes to infinity.

**Proof.** For a sufficiently small $\varepsilon > 0$ and sufficiently large $n$, define $\varepsilon_1(\varepsilon) > 0$ and $\varepsilon_2(\varepsilon, n) > 0$ as

$$\varepsilon_1(\varepsilon) \equiv \varepsilon - \sqrt{2\varepsilon \log \frac{2\varepsilon}{A}},$$

$$\varepsilon_2(\varepsilon, n) \equiv \varepsilon - \frac{A \log(n + 1)}{n}.$$

Note that $\varepsilon_1(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Consequently, we can bound the probability of the non-typical set as follows;

$$\Pr\left\{ \frac{1}{\|x^*(N_{T_n})\|} \log P_{T_n} - H(X) \geq \varepsilon_1(\varepsilon) \right\}$$

$$= \sum_{\ell=1}^{\mathcal{T}(T_n)} \Pr\left\{ \|x^*(N_{T_n})\| = \ell \right\} \cdot \left( \frac{1}{\ell} \log P_{X^\ell} - H(X) \right) \geq \varepsilon_1(\varepsilon) \right\}$$

$$\leq \sum_{\ell=1}^{\mathcal{T}(T_n)} \exp\left\{ -\ell \varepsilon_2(\varepsilon, \ell) \right\}$$

$$\leq \mathcal{T}(T_m) \exp\left\{ -\varepsilon_2(\varepsilon, \mathcal{T}_m) \right\}$$

$$= \mathcal{T}(T_m) \exp\left\{ -\varepsilon_1(\varepsilon, \mathcal{T}_m) \right\}$$

$$= (\mathcal{V}(T_m) + 1)^{d+1} \exp\left\{ -\varepsilon_1(\varepsilon, \mathcal{T}_m) \right\} =: D.$$

From Lemma 3, it holds that $(\mathcal{V}(T_m) + 1)^{d+1} = O((\log |\mathcal{T}_m|)^{d+1})$ and $\varepsilon \exp\left\{ -\varepsilon \mathcal{W}(\mathcal{T}_m) \right\} = O\left( \frac{1}{\mathcal{T}_m} \right)$. Using them, we have $D = O\left( \frac{1}{\log |\mathcal{T}_m|} \right)$. Therefore, the probability of the non-typical set goes to zero as $|\mathcal{T}_m|$ becomes large for any fixed $A$ and $\varepsilon > 0$.

From this theorem, the coding rate of the Tunstall code also converges to the same value as follows;

**Theorem 3:** For any stationary and memoryless source $X$,

$$\frac{\log |\mathcal{T}_m|}{\|x^*(N_{T_n})\|} \rightarrow H(X) \text{ in probability,}$$

as $m$ goes to infinity.

Since the codeword length and the entropy are independent of $v \in \mathcal{T}_m$, this theorem signifies that the depth of the Tunstall tree $\|x^*(v)\|$ becomes almost uniform. In Lemma 3, the upper and lower bounds for all of the leaves are given. This theorem confirms that the ratio of $\log |\mathcal{T}_m|$ and $\|x^*(v)\|$ of most leaves are nearly equal to the entropy of the source.

**Proof.** From Definition 1, we have

$$\frac{\log |\mathcal{T}_m|}{\|x^*(v)\|} = r(v, T_m, X) + \frac{1}{\|x^*(v)\|} \log P_{T_m}(v).$$

Corollary 1 implies that the coding rate converges to the same value as the self-information, and the limit is the entropy of the source from Theorem 2.

4.2 Average Coding Rate of the Tunstall Code

In this subsection, we define the average coding rate of the Tunstall code as the expectation of the pointwise redundancy rate, and prove that the average coding rate converges to the entropy of the source.

First, we show upper and lower bounds of the pointwise self-information rate.

**Lemma 5:** For any stationary and memoryless source $X$ and any Tunstall tree $T_m$, the self-information rate of the label $x^*(v)$ is bounded as follows;

$$\max_{v \in \mathcal{T}_m} \frac{1}{\|x^*(v)\|} \log P_{T_m}(v) \leq \frac{1}{1 - \frac{1}{\log |\mathcal{T}_m|} \log P_{T_m}(v)} \leq \frac{1}{\log P_{T_m}(v)}$$

$$\leq \left( 1 - \frac{1}{\log |\mathcal{T}_m|} \log P_{T_m}(v) \right) \log P_{T_m}(v) \leq \frac{1}{\log |\mathcal{T}_m|} \log P_{T_m}(v) \leq (1 + \epsilon) \log P_{T_m}(v).$$

From this lemma, we can see that for an arbitrarily small $\varepsilon > 0$ and sufficiently large $m$, the self-information rate is bounded as follows;

$$\max_{v \in \mathcal{T}_m} \frac{1}{\|x^*(v)\|} \log P_{T_m}(v) \leq (1 + \varepsilon) \log P_{T_m}(v). \quad (\textbf{12})$$
Proof. From Lemma 2, we have
\[
\Pr^*(\mathcal{T}_m) \leq \frac{P^*(\mathcal{T}_m)}{P}.
\]
Taking the logarithm of both sides,
\[
\log \Pr^*(\mathcal{T}_m) \leq \log \frac{P^*(\mathcal{T}_m)}{P} - \log P
\]
is satisfied. Dividing both sides by \(\log P^*(\mathcal{T}_m)(< 0)\), we have
\[
\frac{\log \Pr^*(\mathcal{T}_m)}{\log P^*(\mathcal{T}_m)} \geq 1 + \frac{1}{\log P^*(\mathcal{T}_m)} \log \frac{1}{P} = 1 - \frac{1}{\log P^*(\mathcal{T}_m)} \log \frac{1}{P} \geq 1 - \frac{1}{\log \mathcal{P}} \log \frac{1}{P}.
\]
Using this inequality, the pointwise self-information rate is bounded from above and below as follows;
\[
\frac{1}{\|x^*(v)\|} \log \frac{1}{P_{r_{\mathcal{T}_m}}(v)} \leq \frac{1}{\|x^*(v)\|} \log \frac{1}{P^*(\mathcal{T}_m)} \leq \frac{\log P}{\log \Pr^*(\mathcal{T}_m)} \log \frac{1}{P_{r_{\mathcal{T}_m}}(v)} = \frac{\log \Pr^*(\mathcal{T}_m)}{\log \Pr^*(\mathcal{T}_m)} \log \frac{1}{P_{r_{\mathcal{T}_m}}(v)} \leq \frac{1}{\log \mathcal{P}} \log \frac{1}{P}.
\]
Since these bounds are independent of \(v\), the bounds for the maximum and minimum self-information rates are the same values as those in the above equations. \(Q.E.D.\)

Next, we prove the convergence of the average self-information rate to the entropy rate of the source using Theorem 2 and Lemma 5.

**Theorem 4:** For any stationary and memoryless source \(X\),
\[
\lim_{m \to \infty} \mathbb{E} \left[ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} \right] = H(X).
\]
**Proof.** We bound the expectation of the self-information rate for the typical set and the non-typical set separately.

First, the average self-information rate is upper-bounded for sufficiently large \(m\) as follows;
\[
\mathbb{E} \left[ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} \right] \leq \Pr \left\{ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} - H(X) \right\} < d \cdot (H(X) + d) + \Pr \left\{ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} - H(X) \right\} \geq d \cdot \max_{v \in \mathcal{P}} \left\{ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} \right\} \leq H(X) + d + O \left( \frac{\log |\mathcal{T}_m|}{|\mathcal{T}_m|^c} \right) (1 + \varepsilon) \log \frac{1}{P}
\]
where \(\varepsilon\) is defined as (12).

Next, the average self-information rate is lower-bounded for sufficiently large \(m\) as follows;
\[
\mathbb{E} \left[ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} \right] \geq \Pr \left\{ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} - H(X) \right\} < d \cdot (H(X) - d) \geq (H(X) - d) \left[ 1 - O \left( \frac{\log |\mathcal{T}_m|}{|\mathcal{T}_m|^c} \right) \right].
\]

Therefore, taking an arbitrarily small \(d > 0\) and sufficiently large \(|\mathcal{T}_m|\), both the upper and lower bounds converge to the entropy of the source. \(Q.E.D.\)

Finally, we show that the average coding rate, which is defined by the expectation of the pointwise coding rate, converges to the entropy of the source.

**Definition 2** (Average Coding Rate 1 of Tunstall Code): We define the average coding rate of the one-shot Tunstall code with the parsing tree \(\mathcal{T}_m\) for the source \(X\) as
\[
\bar{R}_t(\mathcal{T}_m, X) \overset{\text{def}}{=} \mathbb{E} \left[ \frac{\log |\mathcal{T}_m|}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \right].
\]

**Theorem 5:** For any stationary and memoryless source \(X\),
\[
\lim_{m \to \infty} \bar{R}_t(\mathcal{T}_m, X) = H(X).
\]
**Proof.** From Definition 1, we have
\[
\mathbb{E} \left[ \frac{\log |\mathcal{T}_m|}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \right] = \mathbb{E} \left[ r(N_{r_{\mathcal{T}_m}}, \mathcal{T}_m, X) \right] + \mathbb{E} \left[ \frac{1}{\|x^*(N_{r_{\mathcal{T}_m}})\|} \log \frac{1}{P^*(N_{r_{\mathcal{T}_m}})} \right].
\]
Using Corollary 1 and the inequality
\[
\min_{v \in T_m} r(v, T_m, X) \leq \mathbb{E}[r(Nr_m, T_m, X)] \\
\leq \max_{v \in T_m} r(v, T_m, X),
\]
the limit of the average coding rate is the same as that of the average self-information rate. From Theorem 4, the limit is the entropy of the source. Q.E.D.

5. Definition of the Average Coding Rate

In this section, we discuss the definition of the average coding rate of the Tunstall code.

In Theorem 5, we have shown that the average coding rate of the Tunstall code converges to the entropy of the stationary and memoryless source as the extension step \( m \) becomes larger. In this theorem, the average coding rate is defined by (14) as the expectation of the pointwise coding rate.

In the literature, however, the average coding rate of the VF code is defined as the (fixed) codeword length divided by the average block length.

Definition 3: (Average Coding Rate 2 of the Tunstall Code [1]): The average coding rate of the Tunstall code with the parsing tree \( T_m \) for the source \( X \), is defined by
\[
\bar{R}_2(T_m, X) \overset{\text{def}}{=} \frac{\mathbb{E}[\log |T_m|]}{\mathbb{E}[\|x^*(Nr_m)\|]}
\]

For these average coding rates, the following relationship holds generally.

Lemma 6:
\[
\bar{R}_2(T_m, X) \leq \bar{R}_1(T_m, X), \quad \forall m \in \mathbb{N}.
\]

Proof. Applying Jensen’s inequality [16] to \( f(x) = 1/x \), we have
\[
\frac{1}{\mathbb{E}[\|x^*(Nr_m)\|]} \leq \mathbb{E}\left[\frac{1}{\|x^*(Nr_m)\|}\right],
\]
which directly implies (15). Q.E.D.

Since \( f(x) = 1/x \) is strictly convex, the equality in (15) holds only for the uniform-depth Tunstall tree, which cannot be generated from the sources with non-uniform probability distribution, generally. In Theorem 8, we show that this equality holds asymptotically and that these two definitions for the coding rate make no difference, for the class of stationary and memoryless sources, which includes the ones with non-uniform probability distributions.

Using Lemma 6, we derive the following theorem for the upper bound.

Theorem 6 ([2]): For any stationary and memoryless source \( X \),
\[
\lim_{m \to \infty} \bar{R}_2(T_m, X) \leq H(X).
\]

On the other hand, we cannot prove the next theorem using the results we have established in this paper.

Theorem 7 ([2]): For any stationary and memoryless source \( X \),
\[
\lim_{m \to \infty} \bar{R}_1(T_m, X) \geq H(X).
\]

In their efforts to prove Theorems 6 and 7, Jelinek and Schneider [2] use Lemma 2 and the following lemma.

Lemma 7 ([2]):
\[
H(Nr_m) = \mathbb{E}[\|x^*(Nr_m)\|]H(X).
\]

In this paper, Lemma 1 is a counterpart of Lemma 7 for evaluating upper and lower bounds of the pointwise redundancy of the Tunstall code. The values \( -\log P_x \), \( -\log P_x(T_m) \), and \( P_x(T_m) \) in Lemma 1 are upper (lower) bounds of \( H(X) \). They are \( H(Nr_m) \) and \( \mathbb{E}[\|x^*(Nr_m)\|] \) in Lemma 7, respectively.

From Theorems 5, 7, and Lemma 6, we derive the following theorem.

Theorem 8: For any stationary and memoryless source \( X \),
\[
\lim_{m \to \infty} \bar{R}_1(T_m, X) = \lim_{m \to \infty} \bar{R}_2(T_m, X) = H(X).
\]

Therefore, the definitions of \( \bar{R}_1(T_m, X) \) and \( \bar{R}_2(T_m, X) \) make no difference asymptotically for stationary and memoryless sources. Of course, this does not hold for other sources which do not have the asymptotic equipartition property (cf. [16]).

6. Concluding Remarks

One of the advantages of definition (1) is that it is a direct consequence of the pointwise coding rate. This result can be directly extended to the VF coding of general sources [15] with the consistency condition as defined in [12].

Since the two definitions of the average coding rate have been proved to have the same limit, our result establishes a relationship between the pointwise coding/redundancy rate and the average coding rate as defined in (2) via (1).

Acknowledgments

The author wishes to express his gratitude to Prof. Hiroosuke Yamamoto and Prof. Ken-ichi Iwata for their valuable input in the preparation of this paper.

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 20760250.

References

Using Lemma 3 and (7) in Lemma 1, \( r(\nu, T_m, X) \) is bounded from the above as follows:

\[
\begin{align*}
r(\nu, T_m, X) &\leq \frac{1}{w(T_m)} \left\{ \log |T_m| + 1 + \log \frac{P'(T_m)}{|T_m|} \right\} \\
&= \frac{\log |T_m| + 1 + \log \frac{P'(T_m)}{|T_m|}}{w(T_m)} \\
&\leq \frac{\log |T_m| + 1 + \log \frac{P'(T_m)}{|T_m|}}{\log |T_m| - 1 + \log \frac{P}{P'}} \\
&= \frac{1 - \log P}{\log |T_m| + \log P'} \log \frac{1}{P'}.
\end{align*}
\]

This upper bound is independent of \( \nu \), resulting in (5).

In the same manner, the lower bound follows from Lemma 3 and (8) in Lemma 1 as follows:

\[
\begin{align*}
r(\nu, T_m, X) &\geq \frac{1}{w(T_m)} \left\{ \log |T_m| + \log \frac{P'(T_m)}{|T_m|} \right\} \\
&= \frac{\log |T_m| + \log \frac{P'(T_m)}{|T_m|}}{w(T_m)} \\
&\geq \frac{\log |T_m| + \log \frac{P}{P'}}{\log |T_m| - \log \frac{1}{P'}} \log \frac{1}{P'}.
\end{align*}
\]

Since this lower bound is independent of \( \nu \), (6) is satisfied.

**Appendix B: Proof of Lemma 4**

In order to prove Lemma 4, a stationary and memoryless source is denoted by the probability distribution \( P \). We use the following lemmas.

**Lemma 8** (Pinsker, See Lemma 11.6.1 of [16]): For any probability distributions \( P \) and \( Q \) on \( X \),

\[
D(P||Q) \geq \frac{1}{2} \sum_{a \in X} (P(a) - Q(a))^2,
\]

where \( D(P||Q) \) is the Kullback-Leibler divergence.

**Lemma 9** (Lemma 2.7 of [13]): For any probability distributions \( P \) and \( Q \) on the alphabet \( X \) (|X| = A), if

\[
\sum_{a \in X} |P(a) - Q(a)| \leq \Theta < \frac{1}{2},
\]

then

\[
|H(P) - H(Q)| \leq -\Theta \log \frac{\Theta}{A}.
\]

The type and the divergence-typical set are defined as follows;

\[
P_{\nu}(a) \overset{\text{def}}{=} \frac{N(a|x^n)}{n},
\]

\[
T^n_\delta(P) \overset{\text{def}}{=} \{ x^n \in X^n : D(P_{\nu}||P) \leq \delta \},
\]

where \( N(a|x^n) \) is the occurrence number of the symbol \( a \) in the string \( x^n \). Then we have the following lemmas.

**Lemma 10** (Theorem 11.2.1 of [16]):

\[
\Pr[x^n \in X^n : D(P_{\nu}||P) > \epsilon] \leq \exp \left\{ -n \left( \epsilon - \frac{A \log(n+1)}{n} \right) \right\}.
\]

**Lemma 11** (Theorem 2.7 of [14]): For any \( \delta \leq 1/8 \), any \( x^n \in T^n_\delta(P) \) satisfies

\[
\left| \frac{1}{n} \log P_{\nu}(x^n) - H(P) \right| \leq \delta - \sqrt{2\delta} \log \frac{\sqrt{2}}{A}.
\]

**Proof.** For any \( x^n \in T^n_\delta(P) \), Lemma 8 implies that
\[
\sum_{a \in X} |P_{x^n}(a) - P(a)| \leq \sqrt{\frac{2D(P_{x^n}||P)}{\log e}} \leq \sqrt{\frac{2\delta}{\log e}} \leq \sqrt{2\delta}.
\]

Therefore, if \( \delta \leq 1/8 \), the above inequality satisfies the assumption of Lemma 9, so that

\[
|H(P_{x^n}) - H(P)| \leq -\sqrt{2\delta} \log \frac{\sqrt{2\delta}}{A}.
\]

Using this inequality, we have

\[
\left| -\frac{1}{n} \log P_{X^n}(x^n) - H(P) \right|
\]

\[
= \left| -\frac{1}{n} \log \prod_{a \in X} P(a)^{N(a)} - H(P) \right|
\]

\[
= \left| -\sum_{a \in X} P_{x^n}(a) \log P(a) - H(P) \right|
\]

\[
= \left| \sum_{a \in X} P_{x^n}(a) \log \frac{P_{x^n}(a)}{P(a)} + \sum_{a \in X} P_{x^n}(a) \log \frac{1}{P_{x^n}(a)} - H(P) \right|
\]

\[
\leq D(P_{x^n}||P) + |H(P_{x^n}) - H(P)|
\]

\[
\leq \delta - \sqrt{2\delta} \log \frac{\sqrt{2\delta}}{A} \quad \text{Q.E.D.}
\]

Using Lemmas 10 and 11, the following inequality is derived for \( \delta \leq 1/8 \);

\[
\Pr\left\{ \left| -\frac{1}{n} \log P_{X^n}(x^n) - H(P) \right| > \delta - \sqrt{2\delta} \log \frac{\sqrt{2\delta}}{A} \right\}
\]

\[
\leq \Pr\{x^n \in X^n : D(P_{x^n}||P) > \delta\}
\]

\[
\leq \exp\left\{ -n \left( \delta - A \log(n+1) \right) \right\}.
\]

\[\text{Mitsuharu Arimura was born in Kagoshima, Japan, on December 17, 1970. He received his B.E. degree in Mathematical Engineering in 1994, and M.E. and Ph.D. degrees in Information Engineering in 1996 and 1999, respectively, all from the University of Tokyo, Tokyo, Japan. From 1999 to 2004, he was a Research Associate in the Graduate School of Information Systems at the University of Electro-Communications, Tokyo, Japan. Since 2004, he has been at Shonan Institute of Technology, Kanagawa, Japan, starting as a lecturer in the Department of System and Communication Engineering, the name of which was changed to the Department of Applied Computer Sciences in 2006. His research interests include Shannon theory and data compression algorithms. Dr. Arimura is a member of the IEEE.}\]