PAPER *Special Section on Information Theory and Its Applications*

On the Universality of Channel Decoders Constructed from Source Encoders for Finite-State Channels

Tomohiko UYEMATSU[†a)], ***Regular Member and*** **Saad M. KISRA**[†], ***Student Member***

SUMMARY A universal channel decoder for a given family of channels is a decoder that can be designed without prior knowledge of the characteristics of the channel. Nevertheless, it still attains the same random coding error exponent as the optimal decoder tuned to the channel. This paper investigates the duality between universal channel decoders and universal source encoders. First, for the family of finite-state channels, we consider a sufficient condition for constructing universal channel decoders from universal source encoders. Next, we show the existence of a universal channel code that does not depend on the choice of the universal decoder.

key words: *channel coding, finite-state channel, universal decoder, universal source code, reliability function*

1. Introduction

The concept of universal channel decoders was first proposed by Goppa [1]. Later, by combining a fixed composition code with maximum mutual information (MMI) decoder, Csiszár, Körner and Marton [2] (see also [3]) proved the existence of a universal channel code for the family of discrete memoryless channels (DMC's). In 1985, Ziv [4] proposed a universal decoder for the family of unifilar finite-state channels (FSC's), i.e., FSC's with deterministic transitions, by using the Lempel-Ziv (LZ) incremental parsing [5]. As pointed out in [6], Ziv's work implicitly showed the existence of a universal code for the family of unifilar FSC's. Recently, Lapidoth and Ziv [6] extended these results to the general family of FSC's. For this purpose, they used the valuable results obtained by Feder and Lapidoth [7] who discovered the sufficient condition for the existence of universal codes for a given family of channels. However, these results are meaningful only for particular universal decoders such as MMI decoder or Ziv's decoder.

In this paper, we will introduce a new approach to the universal channel coding problem that will lead to the construction of an infinite number of *new* universal decoders from already existing universal source encoders. First, we consider the family of FSC's with state-known-at-receiver, that is, channels of which state is completely known at the receiver. We show a sufficient condition of the universal source encoder for finite-state sources in order that they can be utilized as universal channel decoders for this family of channels. We also show the existence of a universal channel encoder which does not depend on the choice of a particular universal decoder. Then, we extend all these results to the general family of FSC's, i.e. the family of FSC's without any restriction except the initial state. Further, we clarify that universal channel decoders can be constructed from universal codes for memoryless sources. Throughout the paper, we assume that both logarithm and exponent are understood to the base two.

Manuscript received January 19, 2001.
Manuscript revised April 11, 2001.

[†]The authors are with the Department of Communications and Information Systems, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
a) E-mail: uematsu@it.ss.titech.ac.jp

2. Preliminaries

2.1 Basic Definitions

Consider a family of channels defined over the common finite input alphabet $\mathcal{X}$ and the common finite output alphabet $\mathcal{Y}$. Let Θ denote an index set. In this paper, we assume that the channel in use is unknown to the receiver designer, who only knows that the channel belongs to the family of channels $\{p_\theta(\boldsymbol{y}|\boldsymbol{x}), \theta \in \Theta\}$, where the law $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ maps every input sequence $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ to a corresponding probability law on $\mathcal{Y}^n$. We also consider the codebook C_n with rate R and code length n such as

$$C_n = \{\boldsymbol{x}(1), \cdots, \boldsymbol{x}(\lfloor 2^{nR} \rfloor)\} \subset \mathcal{X}^n.$$

where $\lfloor x \rfloor$ denotes the maximum integer which is smaller or equal to x. Generally, a decoder ϕ is a mapping

$$\phi : \mathcal{Y}^n \to \{1, \cdots, \lfloor 2^{nR} \rfloor\},$$

and maps every received sequence $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ to an index i (and by transition to $\boldsymbol{x}(i)$), or declares an error when no appropriate codeword exists.

Without loss of generality, assume that all the codewords in a codebook C_n are used over the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ with equal probability. Then, the average probability of decoding error for the decoder ϕ is given by

$$P_{\theta,\phi}(\text{error}|C) \triangleq \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\{\boldsymbol{y}:\phi(\boldsymbol{y}) \neq i\}} p_\theta(\boldsymbol{y}|\boldsymbol{x}(i)).$$

Now, consider random coding, that is, the codebook C_n is drawn randomly by choosing its codewords independently and uniformly over the input set B^n. For the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$, $\overline{P}_{\theta,\phi}(\text{error})$ denotes the average (over messages and codebooks) probability of error which is incurred when such a random codebook is used over the channel and the decoder is ϕ.

On the other hand, given the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ and the codebook C_n, it is easy to see that for equally probable messages the decoder that minimizes the average probability of error is the maximum-likelihood (ML) decoder. The ML decoder designed for the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ declares that given the received sequence $\boldsymbol{y}$, the transmitted codeword is $\boldsymbol{x}(i)$ only if

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}(i)) = \max_{1\le j\le \lfloor 2^{nR}\rfloor} p_\theta(\boldsymbol{y}|\boldsymbol{x}(j)). \tag{1}$$

For the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$, $P_{\theta,ML}(\text{error}|C)$ denotes the average probability of error incurred when the codebook C is used and ML decoding tuned to θ is employed. Similarly, we use the expression $\overline{P}_{\theta,ML}(\text{error})$ to denote the average (over messages and codebooks) probability of error for a randomly chosen codebook. We can now proceed by defining some terminology used in [7].

Definition 1: A sequence of decoders $\{u_n\}$ is said to be random-coding *weakly* universal for the family of channels $\{p_\theta(\boldsymbol{y}|\boldsymbol{x}) : \theta \in \Theta\}$ and the input-set sequence $\{B_n\}$, $B_n \subset \mathcal{X}^n$ if

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\overline{P}_{\theta,u_n}(\text{error})}{\overline{P}_{\theta,ML}(\text{error})} = 0, \quad \forall \theta \in \Theta. \tag{2}$$

Furthermore, if the convergence in (2) is uniform over Θ, i.e., if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,u_n}(\text{error})}{\overline{P}_{\theta,ML}(\text{error})} = 0,$$

the sequence of decoders $\{u_n\}$ is said to be random-coding *strongly* universal.

Definition 2: A sequence of decoders $\{u_n\}$ is said to be deterministic-coding *weakly* universal for the family of channels $\{p_\theta(\boldsymbol{y}|\boldsymbol{x}) : \theta \in \Theta\}$ and the input-set sequence $\{B_n\}$, $B_n \subset \mathcal{X}^n$, if there exists a sequence of codebooks $\{C_n\}$ with rate R, such that

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P_{\theta,u_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})} = 0, \quad \forall \theta \in \Theta. \tag{3}$$

Also, if the convergence in (3) is uniform over Θ, i.e., if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,u_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})} = 0,$$

the sequence of decoders $\{u_n\}$ is said to be deterministic-coding *strongly* universal.

Throughout this paper, we will deal with the *strong* version of universality. Further, we restrict ourselves to the random coding where codewords are drawn uniformly over the input set $B_n = \mathcal{X}^n$.

2.2 Strong Separable Family of Channels

First, we describe the concept of strong separable family of channels introduced by Feder and Lapidoth [7].

Definition 3: A family of channels $\{p_\theta(\boldsymbol{y}|\boldsymbol{x}),\ \theta \in \Theta\}$ defined over common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is said to be *strongly separable* for the input sets $\{B_n\}$, if there exists some $M > 0$ that upper-bounds the error exponents in the family, i.e. that satisfies

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} -\frac{1}{n} \log \overline{P}_{\theta,ML}(\text{error}) < M$$

such that for every $\delta > 0$ and code length n, there exists a subexponential number $K(n, M, \delta)$ of channels $\{\theta_k^{(n)}\}_{k=1}^{K(n,M,\delta)} \subset \Theta$

$$\lim_{n\to\infty} \frac{1}{n} \log K(n, M, \delta) = 0$$

that well approximate any $\theta \in \Theta$ in the following sense: For any $\theta \in \Theta$, there exists $\theta_{k^*}^{(n)} \in \Theta$, $1 \le k^* \le K(n, M, \delta)$, so that

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) \le 2^{n\delta} p_{\theta_{k^*}^{(n)}}(\boldsymbol{y}|\boldsymbol{x}),$$
$$\forall (\boldsymbol{x}, \boldsymbol{y}) : p_\theta(\boldsymbol{y}|\boldsymbol{x}) > 2^{-n(M+\log|\mathcal{Y}|)},$$

and

$$p_{\theta_{k^*}^{(n)}}(\boldsymbol{y}|\boldsymbol{x}) \le 2^{n\delta} p_\theta(\boldsymbol{y}|\boldsymbol{x}),$$
$$\forall (\boldsymbol{x}, \boldsymbol{y}) : p_{\theta_{k^*}^{(n)}}(\boldsymbol{y}|\boldsymbol{x}) > 2^{-n(M+\log|\mathcal{Y}|)}.$$

The next lemma shows that the family of all FSC's, which will be mainly treated in this paper, is strongly separable.

Lemma 1 [7, Lemma 12]: Consider the family of all FSC's defined over common input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, S$, and characterized by the initial state $s_0 \in S$ and the transition probability

$$\begin{aligned} p_\theta(\boldsymbol{y}|\boldsymbol{x}) &= \sum_{\boldsymbol{s}\in S^n} p_\theta(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}, s_0) \\ &= \sum_{\boldsymbol{s}\in S^n} \prod_{i=1}^{n} p_\theta(y_i, s_i|x_i, s_{i-1}), \end{aligned} \tag{4}$$

where $\boldsymbol{s} = (s_1, s_2, \cdots, s_n) \in S^n$ denotes the state sequence of the channel and S is the finite set of states. Then, this family of channels is strongly separable for any input sets $\{B_n\}$.

The next lemma clarifies the relation between random-coding strong universality and deterministic-coding strong universality for the strongly separable

family of channels.

Lemma 2 [7, Lemma 6]: If the family of channels $\{p_\theta,\ \theta \in \Theta\}$ is strongly separable then random-coding strong universality implies deterministic-coding universality.

By combining above two lemmas, we can conclude that for the family of FSC's random-coding strong universality implies deterministic-coding universality. In the following sections, we often use this observation to simplify the proof.

3. Construction of Universal Decoders for the Family of FSC's with State-Known-at-Receiver

Before investigating the general case of FSC's in Sect. 4, we will treat the special case of FSC with state-known-at-receiver. This is because the proof is more insightful and much simpler by virtue of the *theory of type* [3]. It should be noted that this family of channels is still very large and includes a considerable number of commonly used channels such as DMC's and channels subject to intersymbol interference.

For this family of channels, we assume that the state of the channel is known to the receiver. In other words, we assume that the finite set of states S, the initial state $s_0 \in S$, the next-state function $q : \mathcal{Y} \times S \to S$, and

$$s_i = q(y_i, s_{i-1}) \qquad \text{for } i = 1, \cdots, n$$

are given to the receiver. By this fact, the transition probability can be reduced to the form

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p_\theta(y_i|x_i, s_{i-1}), \tag{5}$$

where $s_i \in S$ $(i = 1, \cdots, n)$ denotes the state of the channel $\theta \in \Theta$ at the ith instant.

Next, we define the asymptotical optimality of a sequence of source codes for *finite-state sources* (FSS's). In what follows, for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$, we use the notation $(\boldsymbol{x}, \boldsymbol{y})$ to mean $(x_1y_1, x_2y_2, \cdots, x_ny_n)$, and we may write $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^n$.

Definition 4: For a given next state function q, a sequence of binary source codes $\{f_n\}$ (where $f_n : (\mathcal{X} \times \mathcal{Y})^n \to \{0,1\}^*$) is said to be asymptotically optimum for FSS's if there exists a sequence $\{\epsilon_n\}$ such that $\epsilon_n \geq 0$, $\lim_{n\to\infty} \epsilon_n = 0$ and

$$\ell(f_n(\boldsymbol{x}, \boldsymbol{y})) \leq nH(\boldsymbol{xy}|\boldsymbol{s}) + n\epsilon_n, \quad \forall(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^n \tag{6}$$

where $\ell(\cdot)$ denotes the length function, $\boldsymbol{s} = (s_0, s_1, \cdots, s_{n-1})$ denotes the sequence of states, and $H(\boldsymbol{xy}|\boldsymbol{s})$ is the conditional empirical entropy determined by the joint type of sequences defined as

$$P_{\boldsymbol{xy}}(a, b, s) \triangleq \frac{1}{n}|\{i : (x_i, y_i, s_{i-1}) = (a, b, s), 1 \leq i \leq n\}|, \quad \forall(a, b, s) \in \mathcal{X} \times \mathcal{Y} \times S.$$

It should be noted that $H(\boldsymbol{xy}|\boldsymbol{s})$ is the empirical entropy of the unifilar source when the next-state function q and the initial state s_0 are known. Hence, it is easy to see that almost all universal source codes for FSS's, such as Lempel-Ziv 78 code [5] and adaptive arithmetic codes [8], [9] using a model determined by q are asymptotically optimum.

Also, in order to prove our theorem, we require that the length of the codeword satisfies the condition

$$\ell(f_n(\boldsymbol{x}, \boldsymbol{y})) \geq nH(\boldsymbol{xy}|\boldsymbol{s}), \quad \forall(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^n. \tag{7}$$

However, this requirement is not significant because even if some codewords do not satisfy it, we can always modify their length by adding some 0's at the end so that (7) is satisfied. We can easily show that adaptive arithmetic codes [8], [9] using a model determined by q satisfy this condition. Also, we can verify that (7) holds for any *regular code* [8] satisfying

$$\ell(f_n(\boldsymbol{x}, \boldsymbol{y})) \geq -\log P^n(\boldsymbol{x}, \boldsymbol{y}), \quad \forall(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^n$$

where $P^n(\boldsymbol{x}, \boldsymbol{y})$ is some probability function defining the FSS.

The next definition describes the method for constructing a sequence of decoders $\{\hat{f}_n\}$ from a sequence of asymptotically optimum source codes $\{f_n\}$.

Definition 5: For a codebook $C_n \subset \mathcal{X}^n$ and a given received sequence $\boldsymbol{y} \in \mathcal{Y}^n$, the decoder $\{\hat{f}_n\}$ declares that the transmitted codeword is $\boldsymbol{x}(i)$ only if

$$\ell(f_n(\boldsymbol{x}(i), \boldsymbol{y})) \leq \ell(f_n(\boldsymbol{x}(j), \boldsymbol{y})), \qquad \forall j \neq i. \tag{8}$$

In other words, this decoder outputs the codeword that produces the shortest codeword when encoded jointly with the received sequence. This decoder is universal since it does not require any knowledge of the channel.

The next theorem compares the performance of the proposed decoder with the optimal performance obtainable by the ML decoder.

Theorem 1: For a family of FSC's characterized by (5), the sequence of universal decoders $\{\hat{f}_n\}$ are random-coding *strongly* universal, i.e.

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,\hat{f}_n}(\text{error})}{\overline{P}_{\theta,ML}(\text{error})} = 0, \tag{9}$$

and deterministic-coding *strongly* universal, i.e. there exists a sequence of deterministic codebooks $\{C_n\}$ with

rate R so that

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\frac{1}{n}\log\frac{P_{\theta,\hat{f}_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})}=0. \tag{10}$$

Remark 1: The difference between Ziv's result [4, Sect. II D] and Theorem 1 can be summarized as follows: (i) Ziv dealt with the family of DMC's, but Theorem 1 deals with more general family of channels, i.e. the family of FSC's with state-known-at-receiver. (ii) Ziv only considered the MMI decoder, but Theorem 1 shows that almost all universal source encoders can be used as channel decoder. Hence, Theorem 1 enriches the variety of universal channel decoders.

Before we prove the theorem, we first introduce a very important lemma proved by Ziv [4].

Lemma 3 [4, Corollary 1]: For any channel over common finite input and output alphabet $\mathcal{X}$, $\mathcal{Y}$ where the codewords are uniformly chosen from the input set $B_n\subset\mathcal{X}^n$, we have

$$\frac{\overline{P}_{\theta,\hat{f}_n}(\text{error})}{2\overline{P}_{\theta,ML}(\text{error})}\le\max_{\boldsymbol{x}\in B_n,\boldsymbol{y}\in\mathcal{Y}^n}\frac{|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|}{|S_{ML}(\boldsymbol{x},\boldsymbol{y})|}+1,$$

where

$$S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})\triangleq\{\hat{\boldsymbol{x}}\in B_n:\ell(f_n(\hat{\boldsymbol{x}},\boldsymbol{y}))\le\ell(f_n(\boldsymbol{x},\boldsymbol{y}))\},$$

$$S_{ML}(\boldsymbol{x},\boldsymbol{y})\triangleq\{\hat{\boldsymbol{x}}\in B_n:p_\theta(\boldsymbol{y}|\hat{\boldsymbol{x}})\ge p_\theta(\boldsymbol{y}|\boldsymbol{x})\}.$$

Proof of Theorem 1: According to the results of Feder and Lapidoth described in Sect. 2.2, we only prove the first half of Theorem 1, i.e. the random-coding universality. In what follows, all the information measures like the empirical entropy and the conditional empirical entropy that are induced by the joint type $P_{\boldsymbol{xy}}$ will be denoted by $H(\boldsymbol{xy})$, $H(\boldsymbol{x}|\boldsymbol{ys})$, etc.

For any two pairs of sequences $(\boldsymbol{x},\boldsymbol{y})$ and $(\hat{\boldsymbol{x}},\boldsymbol{y})$ both in $(\mathcal{X}\times\mathcal{Y})^n$, $P_{\boldsymbol{xy}}=P_{\hat{\boldsymbol{x}}\boldsymbol{y}}$ implies $p_\theta(\boldsymbol{y}|\boldsymbol{x})=p_\theta(\boldsymbol{y}|\hat{\boldsymbol{x}})$. Hence,

$$\begin{aligned}|S_{ML}(\boldsymbol{x},\boldsymbol{y})|&\ge|\{\hat{\boldsymbol{x}}\in\mathcal{X}^n:P_{\hat{\boldsymbol{x}}\boldsymbol{y}}=P_{\boldsymbol{xy}}\}|\\&\overset{(a)}{=}\prod_{b\in\mathcal{Y},s\in S}\frac{\{n\sum_{a\in\mathcal{X}}P_{\boldsymbol{xy}}(a,b,s)\}!}{\prod_{a\in\mathcal{X}}(nP_{\boldsymbol{xy}}(a,b,s))!}\\&\overset{(b)}{\ge}\frac{\exp\{nH(\boldsymbol{x}|\boldsymbol{ys})\}}{(n+1)^{|\mathcal{X}||\mathcal{Y}||S|}},\end{aligned} \tag{11}$$

where (a) comes from the fact that the next state s_i is a function of y_i and s_{i-1}, and (b) comes from the standard method of type [3].

On the other hand, for the given class of channels, the output sequence $\boldsymbol{y}\in\mathcal{Y}^n$ is known to the receiver and the sequence of states can be determined uniquely. This implies that $H(\boldsymbol{y}|\boldsymbol{s})$ does not depend on $\boldsymbol{x}$. From (6), (7) and $H(\boldsymbol{xy}|\boldsymbol{s})=H(\boldsymbol{x}|\boldsymbol{ys})+H(\boldsymbol{y}|\boldsymbol{s})$, we can show that $\ell(f_n(\hat{\boldsymbol{x}},\boldsymbol{y}))\le\ell(f_n(\boldsymbol{x},\boldsymbol{y}))$ implies that

$$H(\hat{\boldsymbol{x}}|\boldsymbol{ys})\le H(\boldsymbol{x}|\boldsymbol{ys})+\epsilon_n.$$

By using these relations, we can bound $|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|$ as follows

$$\begin{aligned}&|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|\\&=|\{\hat{\boldsymbol{x}}\in\mathcal{X}^n:\ell(f_n(\hat{\boldsymbol{x}},\boldsymbol{y}))\le\ell(f_n(\boldsymbol{x},\boldsymbol{y}))\}|\\&\le|\{\hat{\boldsymbol{x}}\in\mathcal{X}^n:H(\hat{\boldsymbol{x}}|\boldsymbol{ys})\le H(\boldsymbol{x}|\boldsymbol{ys})+\epsilon_n\}|\\&\le(n+1)^{|\mathcal{X}||\mathcal{Y}||S|}\exp\{n(H(\boldsymbol{x}|\boldsymbol{ys})+\epsilon_n)\}.\end{aligned} \tag{12}$$

Combining (11) and (12), we obtain for any $(\boldsymbol{x},\boldsymbol{y})$

$$\begin{aligned}&\frac{1}{n}\log\left(\frac{|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|}{|S_{ML}(\boldsymbol{x},\boldsymbol{y})|}+1\right)\\&\overset{(c)}{\le}\frac{1}{n}\log\left\{(n+1)^{2|\mathcal{X}||\mathcal{Y}||S|}\exp(n\epsilon_n)+1\right\}\\&=\frac{1}{n}+\epsilon_n+\frac{2|\mathcal{X}||\mathcal{Y}||S|\log(n+1)}{n},\end{aligned}$$

where (c) comes from the inequality $\log(1+x)\le1+\log x$ for $x\ge1$. This upper bound vanishes as n tends to infinity, and does not depend on $\theta\in\Theta$. Hence, by using Lemma 3, the first half of Theorem 1 is obtained. Q.E.D.

Next, we strengthen our result by showing the existence of a universal codebook (or encoder) which does not depend on a particular universal decoder.

Theorem 2: Consider a family of channels characterized by (5). For a given sequence $\{\epsilon_n\}$ such that $\epsilon_n\ge0$ and $\lim_{n\to\infty}\epsilon_n=0$, there exists a sequence of codebooks $\{C_n\}$ with rate R such that all sequences of universal decoders $\{\hat{f}_n\}$ obtained by the sequence of universal source codes $\{f_n\}$ satisfying (6) for the given $\{\epsilon_n\}$ above are deterministic-coding *strongly* universal, i. e.

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\frac{1}{n}\log\frac{P_{\theta,\hat{f}_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})}=0, \tag{13}$$

and this sequence of codebooks $\{C_n\}$ does not depend on the sequence of universal decoders $\{\hat{f}_n\}$.

Proof of Theorem 2: Let us consider the universal threshold decoder $\{u_n\}$ defined by the following decoding rule: For a codebook $C_n\subset\mathcal{X}^n$, the given nonnegative sequence $\{\epsilon_n\}$ and a received sequence $\boldsymbol{y}\in\mathcal{Y}^n$, the threshold decoder $\{u_n\}$ declares that the transmitted codeword is $\boldsymbol{x}(i)$ only if

$$H(\boldsymbol{x}(i)\boldsymbol{y}|\boldsymbol{s})\le H(\boldsymbol{x}(j)\boldsymbol{y}|\boldsymbol{s})-\epsilon_n,\qquad\forall j\ne i.$$

Then, in a manner similar to the proof of [4, Corollary 1], we have

$$\frac{\overline{P}_{\theta,u_n}(\text{error})}{2\overline{P}_{\theta,ML}(\text{error})}\le\max_{\boldsymbol{x}\in B_n,\boldsymbol{y}\in\mathcal{Y}^n}\frac{|S_{u_n}(\boldsymbol{x},\boldsymbol{y})|}{|S_{ML}(\boldsymbol{x},\boldsymbol{y})|}+1,$$

where

$$S_{u_n}(\boldsymbol{x},\boldsymbol{y}) \stackrel{\triangle}{=} \{\hat{\boldsymbol{x}} \in \mathcal{X}^n : H(\hat{\boldsymbol{x}}\boldsymbol{y}|\boldsymbol{s}) \le H(\boldsymbol{x}\boldsymbol{y}|\boldsymbol{s}) + \epsilon_n\}.$$

Hence, in a manner similar to the proof of Theorem 1, we can show that there exists a sequence of codebooks $\{C_n\}$ such that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{P_{\theta,u_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})} = 0. \tag{14}$$

This implies that the sequence of decoders $\{u_n\}$ is deterministic-coding universal.

On the other hand, for any sequence of universal decoders $\{\hat{f}_n\}$ obtained by the sequence of asymptotically optimum source codes $\{f_n\}$ satisfying (6) and (7), $H(\boldsymbol{x}(i)\boldsymbol{y}|\boldsymbol{s}) \le H(\boldsymbol{x}(j)\boldsymbol{y}|\boldsymbol{s}) - \epsilon_n$ implies $\ell(f_n(\boldsymbol{x}(i),\boldsymbol{y})) \le \ell(f_n(\boldsymbol{x}(j),\boldsymbol{y}))$. Hence, we have

$$\{\boldsymbol{y} \in \mathcal{Y}^n : u_n(\boldsymbol{y}) = i\} \subset \{\boldsymbol{y} \in \mathcal{Y}^n : \hat{f}_n(\boldsymbol{y}) = i\} \quad \forall i \in \{1,\cdots,\lfloor 2^{nR} \rfloor\}.$$

Therefore,

$$\begin{aligned} & P_{\theta,\hat{f}_n}(\text{error}|C_n) \\ &= \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\{\boldsymbol{y}:\hat{f}_n(\boldsymbol{y})\neq i\}} p_\theta(\boldsymbol{y}|\boldsymbol{x}(i)) \\ &\le \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\{\boldsymbol{y}:u_n(\boldsymbol{y})\neq i\}} p_\theta(\boldsymbol{y}|\boldsymbol{x}(i)) \\ &= p_{\theta,u_n}(\text{error}|C_n). \end{aligned}$$

Combining this with (14), we obtain (13). Q.E.D.

4. Construction of Universal Decoders for the Family of General FSC's

In this section, we will consider the general family of FSC's which are characterized by the initial state $s_0 \in S$ and the transition probability (4).

In Sect. 3, we defined the asymptotical optimality for FSS's, which turned out to be a sufficient condition to construct universal decoders for the family of FSC's with state-known-at-receivers. Similarly, in order to show a sufficient condition to construct the universal decoders for the general family of FSC's, we define the asymptotical optimality of source codes for *memoryless sources* in a little bit different manner as Definition 4.

Definition 6: A sequence of binary source codes $\{f_n\}$ $(f_n : (\mathcal{X}\times\mathcal{Y})^n \to \{0,1\}^*)$ is said to be asymptotically optimum if there exists a sequence $\{\epsilon_n(|\mathcal{X}\times\mathcal{Y}|)\}$ such that $\epsilon_n(|\mathcal{X}\times\mathcal{Y}|) \ge 0$, $\lim_{n\to\infty} \epsilon_n(|\mathcal{X}\times\mathcal{Y}|) = 0$ and

$$|\ell(f_n(\boldsymbol{x},\boldsymbol{y})) - nH(\boldsymbol{x}\boldsymbol{y})| \le n\epsilon_n(|\mathcal{X}\times\mathcal{Y}|), \quad \forall(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \tag{15}$$

where $H(\boldsymbol{x}\boldsymbol{y})$ is the joint empirical entropy determined by the joint type of sequences defined as

$$P_{\boldsymbol{x}\boldsymbol{y}}(a,b) \stackrel{\triangle}{=} \frac{1}{n} \left|\{i : (x_i,y_i) = (a,b), 1 \le i \le n\}\right|, \quad \forall(a,b) \in \mathcal{X}\times\mathcal{Y}. \tag{16}$$

It is easy to see that almost all universal source codes for memoryless sources, such as Lynch-Davisson code [10], [11], adaptive arithmetic codes [8], [9] are asymptotically optimum. It should be emphasized that in Sect. 3 we treated the source codes for FSS's, but now we are considering the source codes for memoryless sources. Further, the sequence of source codes satisfying both (6) and (7) satisfies (15) as well. Hence, the class of source codes treated here is much wider than that in Sect. 3.

In what follows, we assume that k divides n and we parse $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathcal{X}^n$, $\boldsymbol{s} \in S^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ into n/k blocks of length k. We also denote the ith block of $\boldsymbol{x}$ that starts at the $(i-1)k+1$th position and ends at the ikth as $\tilde{\boldsymbol{x}}_i$ where $i = 1,\cdots,n/k$. We also assume that the source code f_n encodes sequences on the extended alphabet $(\mathcal{X}\times\mathcal{Y})^k$. When the sequence $(\boldsymbol{x},\boldsymbol{y}) \in (\mathcal{X}\times\mathcal{Y})^n$ is encoded by f_n over the extended alphabet $(\mathcal{X}\times\mathcal{Y})^k$, we denote the corresponding codeword as $f_n(\boldsymbol{x},\boldsymbol{y},k)$. Further, for the extended alphabet $(\mathcal{X}\times\mathcal{Y})^k$, the asymptotically optimality (15) can be translated to

$$\left|\ell(f_n(\boldsymbol{x},\boldsymbol{y},k)) - \frac{n}{k}H(\boldsymbol{x}\boldsymbol{y},k)\right| \le \frac{n}{k}\epsilon_{n/k}(|\mathcal{X}\times\mathcal{Y}|^k), \quad \forall(\boldsymbol{x},\boldsymbol{y}) \in (\mathcal{X}\times\mathcal{Y})^n \tag{17}$$

where

$$H(\boldsymbol{x}\boldsymbol{y},k) \stackrel{\triangle}{=} -\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^k} \sum_{\tilde{\boldsymbol{y}}\in\mathcal{Y}^k} P_{\boldsymbol{x}\boldsymbol{y}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \log P_{\boldsymbol{x}\boldsymbol{y}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \tag{18}$$

is the empirical entropy associated with the joint type

$$P_{\boldsymbol{x}\boldsymbol{y}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \stackrel{\triangle}{=} \frac{k}{n}\left|\left\{i : (\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{y}}_i) = (\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}), 1 \le i \le \frac{n}{k}\right\}\right|, \quad \forall(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \in \mathcal{X}^k \times \mathcal{Y}^k. \tag{19}$$

From now on, we assume that the subblock length k is specified depending on the code length n and satisfies three conditions

$$\left.\begin{aligned} &\lim_{n\to\infty} k = \infty, \\ &\lim_{n\to\infty} \frac{|\mathcal{X}\times\mathcal{Y}|^k \log(n/k)}{n} = 0, \\ &\lim_{n\to\infty} \epsilon(n,k) = 0, \end{aligned}\right\} \tag{20}$$

where

$$\epsilon(n,k) \stackrel{\triangle}{=} \frac{\epsilon_{n/k}(|\mathcal{X}\times\mathcal{Y}|^k)}{k}.$$

Next, we construct a sequence of universal decoders $\{\hat{f}_n\}$ from a given sequence of asymptotically

optimum source codes $\{f_n\}$ for memoryless sources.

Definition 7: For a codebook $C_n \subset \mathcal{X}^n$ and a given received sequence $\boldsymbol{y} \in \mathcal{Y}^n$, the decoder $\{\hat{f}_n\}$ declares that the transmitted codeword is $\boldsymbol{x}(i)$ only if

$$\ell(f_n(\boldsymbol{x}(i), \boldsymbol{y}, k)) \leq \ell(f_n(\boldsymbol{x}(j), \boldsymbol{y}, k)), \quad \forall j \neq i. \tag{21}$$

The performance obtained for this sequence of universal decoders is clarified in Theorem 3 and Theorem 4.

Theorem 3: For the family of FSC's characterized by (4), the sequence of decoders $\{\hat{f}_n\}$ is random-coding *strongly* universal and deterministic-coding *strongly* universal. Thus,

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,\hat{f}_n}(\text{error})}{\overline{P}_{\theta,ML}(\text{error})} = 0, \tag{22}$$

and there also exists a sequence of codebooks $\{C_n\}$ with rate R so that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{P_{\theta,\hat{f}_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})} = 0. \tag{23}$$

Remark 2: In [6, Theorem 1], Lapidoth and Ziv showed that Ziv's decoder [4] is deterministic-coding universal for the family of FSC's. On the other hand, Theorem 3 shows that decoders constructed from almost all universal source encoders for memoryless sources are deterministic-coding universal for the family of FSC's. This together with Theorem 1 drastically enriches the variety of universal channel decoders.

Proof of Theorem 3: As described in Sect. 2.2, it is sufficient to prove the random-coding universality of the proposed decoder for the family of FSC's.

We first introduce a threshold decoder that is comparable to the ML decoder. For the channel $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ characterized by (4), a threshold decoder ϕ_{TH} with threshold sequence $\{\alpha_n\}, \alpha_n \geq 1$, is defined as follows: For a received sequence $\boldsymbol{y}$, the threshold decoder ϕ_{TH} declares that codeword $\boldsymbol{x}(i)$ is transmitted only if

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}(i)) \geq \alpha_n p_\theta(\boldsymbol{y}|\boldsymbol{x}(j)), \quad \text{for } j \neq i,$$

and declares an error if no such codeword exists.

Lemma 4 [6, Lemma 2]: For the threshold decoder ϕ_{TH}, if the threshold sequence satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = 0, \tag{24}$$

we have

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,\phi_{TH}}(\text{error})}{\overline{P}_{\theta,ML}(\text{error})} = 0. \tag{25}$$

Note that the threshold decoder is not universal and is, in general, inferior to the ML decoder. However, Lemma 4 shows that when the threshold sequence satisfies the condition (24), the threshold decoder is asymptotically optimum. Hence, in order to prove Theorem 3, we only need to show that the proposed universal decoder is asymptotically as good as the threshold decoder. In other words, we need to prove

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta,\hat{f}_n}(\text{error})}{\overline{P}_{\theta,\phi_{TH}}(\text{error})} = 0. \tag{26}$$

To this end, in a manner similar to the proof of [4, Corollary 1], we immediately have

$$\frac{\overline{P}_{\theta,\hat{f}_n}(\text{error})}{2\overline{P}_{\theta,\phi_{TH}}(\text{error})} \leq 1 + \max_{\boldsymbol{x}\in\mathcal{X}^n, \boldsymbol{y}\in\mathcal{Y}^n} \frac{|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|}{|S_{TH}(\boldsymbol{x},\boldsymbol{y})|}, \tag{27}$$

where

$$S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y}) = \{\hat{\boldsymbol{x}} \in \mathcal{X}^n : \ell(f_n(\hat{\boldsymbol{x}},\boldsymbol{y},k)) \leq \ell(f_n(\boldsymbol{x},\boldsymbol{y},k))\},$$

$$S_{TH}(\boldsymbol{x},\boldsymbol{y}) = \{\hat{\boldsymbol{x}} \in \mathcal{X}^n : p_\theta(\boldsymbol{y}|\hat{\boldsymbol{x}}) \geq \alpha_n^{-1} p_\theta(\boldsymbol{y}|\boldsymbol{x})\}.$$

The next step is to bound $|S_{TH}(\boldsymbol{x},\boldsymbol{y})|$ and $|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|$. This is presented in Lemma 5 and Lemma 6, respectively.

Lemma 5: For the threshold decoder with the threshold sequence $\alpha_n = |S|^{n/k}$, we have

$$\frac{1}{n} \log |S_{TH}(\boldsymbol{x},\boldsymbol{y})| \geq \frac{1}{k} H(\boldsymbol{x}|\boldsymbol{y},k) - \frac{1}{k} \log |S|^2 e, \tag{28}$$

where $H(\boldsymbol{x}|\boldsymbol{y},k)$ is the conditional empirical entropy defined as

$$H(\boldsymbol{x}|\boldsymbol{y},k) \triangleq -\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^k} \sum_{\tilde{\boldsymbol{y}}\in\mathcal{Y}^k} P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \log P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}}), \tag{29}$$

while

$$P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}}) \triangleq \frac{P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})}{\sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^k} P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})},$$

and $P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})$ is defined in (19).

Lemma 6: For the sequence of decoders $\{\hat{f}_n\}$ obtained from the sequence of asymptotically optimum source codes defined in Definition 6, we have

$$\frac{1}{n} \log |S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})| \leq \frac{1}{k} H(\boldsymbol{x}|\boldsymbol{y},k) + 2\epsilon(n,k) + \frac{2}{k} + |\mathcal{X}\times\mathcal{Y}|^k \left(\frac{\log(n/k)+1}{n}\right). \tag{30}$$

The proofs of these lemmas are shown in Appendix.

Remark 3: Lapidoth and Ziv proved similar lemmas associated with Ziv's decoder [6, Lemma 4 and Lemma

3], where the term of conditional empirical entropy $H(\boldsymbol{x}|\boldsymbol{y},k)/k$ in (28) and (30) is replaced by the Ziv's decoding function $u(\boldsymbol{x},\boldsymbol{y})$ (see [6, Eq. (11)] for its definition), and the other terms in right hand side of (28) and (30) are replaced by functions of $O(\log\log n/\log n)$.

We are now in a position to prove Theorem 3. First, by combining (28) and (30), we can show that for all $(\boldsymbol{x},\boldsymbol{y})$

$$\begin{aligned}&\frac{1}{n}\log\left(1+\frac{|S_{\hat{f}_n}(\boldsymbol{x},\boldsymbol{y})|}{|S_{TH}(\boldsymbol{x},\boldsymbol{y})|}\right)\\&\le\frac{1}{n}+2\epsilon(n,k)+\frac{2}{k}+|\mathcal{X}\times\mathcal{Y}|^k\left(\frac{\log(n/k)+1}{n}\right)\\&\quad+\frac{1}{k}\log|S|^2e.\end{aligned}$$

In this inequality, the right hand side does not depend on $\theta\in\Theta$ and vanishes as n tends to infinity from the assumption of k (20). By combining this with (27), we obtain (26). Q.E.D.

The next theorem is an extended version of Theorem 2.

Theorem 4: Consider a family of FSC's characterized by (4). For a given sequence $\{\epsilon_n\}$ such that $\epsilon_n\ge 0$ and $\lim_{n\to\infty}\epsilon_n=0$, there exists a sequence of codebooks $\{C_n\}$ with rate R such that all sequences of universal decoders $\{\hat{f}_n\}$ obtained by the sequence of universal source codes $\{f_n\}$ satisfying (15) and $\epsilon(n,k)\le\epsilon_n$ are deterministic-coding *strongly* universal. Thus,

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\frac{1}{n}\log\frac{P_{\theta,\hat{f}_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})}=0. \tag{31}$$

It should be noted that this sequence of codebooks $\{C_n\}$ does not depend on the sequence of universal decoders $\{\hat{f}_n\}$.

Proof of Theorem 4: Consider the following universal decoder $\{\tilde{u}_n\}$: For a codebook $C_n\subset\mathcal{X}^n$ and a given received sequence $\boldsymbol{y}\in\mathcal{Y}^n$, the universal decoder $\{\tilde{u}_n\}$ declares that the transmitted codeword is $\boldsymbol{x}(i)$ only if

$$\frac{n}{k}H(\boldsymbol{x}(i)\boldsymbol{y},k)<\frac{n}{k}H(\boldsymbol{x}(j)\boldsymbol{y},k)-2\epsilon_n,\quad\forall j\ne i.$$

where $H(\boldsymbol{x}(i)\boldsymbol{y},k)$ is defined in (18).

First, we will prove the strong universality of the sequence of decoders $\{\tilde{u}_n\}$. We first consider the set

$$\begin{aligned}S_{\tilde{u}_n}(\boldsymbol{x},\boldsymbol{y})=\Big\{\hat{\boldsymbol{x}}\in\mathcal{X}^n\ :\ &\frac{n}{k}H(\boldsymbol{x}(i)\boldsymbol{y},k)<\\&\frac{n}{k}H(\boldsymbol{x}(j)\boldsymbol{y},k)+2\epsilon_n\Big\}.\end{aligned}$$

Then, in a method similar to the proof of Lemma 6, we have

$$\begin{aligned}&\frac{1}{n}\log|S_{\tilde{u}_n}(\boldsymbol{x},\boldsymbol{y})|\\&\le\frac{1}{k}H(\boldsymbol{x}|\boldsymbol{y},k)+2\epsilon_n\\&\quad+\frac{2}{k}+|\mathcal{X}\times\mathcal{Y}|^k\left(\frac{\log(n/k)+1}{n}\right).\end{aligned}$$

Combining this result with Lemma 5 in a manner similar to the proof of Theorem 3, we can show that $\{\tilde{u}_n\}$ is random-coding universal. Since random-coding universality implies deterministic-coding universality for the family of FSC's, there exists a sequence of deterministic codebooks $\{C_n\}$ such that

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\frac{1}{n}\log\frac{P_{\theta,\tilde{u}_n}(\text{error}|C_n)}{\overline{P}_{\theta,ML}(\text{error})}=0$$

for a sequence of universal decoders $\{\tilde{u}_n\}$.

On the other hand, when $\tilde{u}_n(\boldsymbol{y})=i$, for any universal source code f_n satisfying $\epsilon(n,k)\le\epsilon_n$, we have

$$\begin{aligned}\ell(f_n(\boldsymbol{x}(i),\boldsymbol{y},k))&\le\frac{n}{k}H(\boldsymbol{x}(i)\boldsymbol{y},k)+\epsilon_n\\&<\frac{n}{k}H(\boldsymbol{x}(j)\boldsymbol{y},k)-\epsilon_n\\&\le\ell(f_n(\boldsymbol{x}(j),\boldsymbol{y},k)),\qquad\forall j\ne i\end{aligned}$$

which implies that $\hat{f}_n(\boldsymbol{y})=i$. Therefore, in a manner similar to the proof of Theorem 2, whatever the decoder $\hat{f}_n$ is, we obtain

$$P_{\theta,\hat{f}_n}(\text{error}|C_n)\le p_{\theta,u_n}(\text{error}|C_n).$$

Since the sequence of codebooks $\{C_n\}$ is determined by $\{\tilde{u}_n\}$, all the decoders $\{\hat{f}_n\}$ associated with the source encoders $\{f_n\}$ are also deterministic-coding universal. Q.E.D.

5. Conclusion

We have considered the relationship between source coding and channel coding. We have found a sufficient condition for universal source encoders to be utilized as universal channel decoders. With regard to the abundance of literature concerning universal source encoders, this enables us to construct a considerable number of new universal channel decoders. Next, we have shown the existence of a universal channel encoder that does not depend on the choice of a universal decoder for a family of FSC's.

References

[1] V.D. Goppa, "Nonprobabilitistic mutual information without memory," Problems of Control and Inform., Theory, vol.4, pp.97–102, 1975.
[2] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," Presented at the IEEE Int. Symp. Information Theory, Cornell Univ., Ithaca, NY, 1977.
[3] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic, New York, 1981.

[4] J. Ziv, "Universal decoding for finite-state channels," IEEE Trans. Inf. Theory, vol.31, no.4, pp.453–460, July 1985.
[5] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," IEEE Trans. Inf. Theory, vol.24, no.5, pp.530–536, Sept. 1978.
[6] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," IEEE Trans. Inf. Theory, vol.44, no.5, pp.1746–1755, Sept. 1998.
[7] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," IEEE Trans. Inf. Theory, vol.44, no.5, pp.1726–1745, Sept. 1998.
[8] J. Rissanen, "Complexity of strings in the class of Markov sources," IEEE Trans. Inf. Theory, vol.32, no.4, pp.526–532, July 1986.
[9] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: Basic properties," IEEE Trans. Inf. Theory, vol.41, no.3, pp.653–664, May 1995.
[10] T.J. Lynch, "Sequence time coding for data compression," Proc. IEEE, vol.54, no.12, p.2010, Dec. 1966.
[11] L.D. Davisson "Comments on 'Sequence time coding for data compression'," Proc. IEEE, vol.54, no.10, pp.1490–1491, Oct. 1966.

Appendix: Proof of Lemma 5

The proof is similar to the proof of [6, Lemma 4]. One significant difference is that, as mentioned earlier, we first divide the input, output and state sequences into blocks of fixed length k.

For any state sequence, let $p(\boldsymbol{y}, \sigma|\boldsymbol{x}, s_0)$ denote the probability that for a given initial state s_0 and input sequence $\boldsymbol{x}$, we observe the output sequence $\boldsymbol{y}$ together with the states $s_{ik} = \sigma(i)$ for all $i \in \{1, 2, \cdots, n/k\}$. Then, we have

$$p(\boldsymbol{y}, \sigma|\boldsymbol{x}, s_0) = \prod_{m=1}^{n/k} p_\theta(\tilde{\boldsymbol{y}}_m, \sigma(m)|\tilde{\boldsymbol{x}}_m, \sigma(m-1)), \qquad (\mathrm{A}\cdot 1)$$

where $\sigma(0) = s_0$ and $p(\tilde{\boldsymbol{y}}_m, \sigma(m)|\tilde{\boldsymbol{x}}_m, \sigma(m-1))$ is the probability that the channel will be in the state $\sigma(m)$ at the (n/k)th instant and will produce the output $\tilde{\boldsymbol{y}}_m$ given that the channel is in the state $\sigma(m-1)$ and is fed with the input $\tilde{\boldsymbol{x}}_m$. Since there are $|S|^{n/k}$ kinds of state sequences, we can easily see that there exists a sequence $\sigma_0 \in S^{n/k}$ that satisfies

$$p(\boldsymbol{y}, \sigma_0|\boldsymbol{x}, s_0) \geq \frac{1}{|S|^{n/k}} p_\theta(\boldsymbol{y}|\boldsymbol{x}). \qquad (\mathrm{A}\cdot 2)$$

On the other hand, let $\boldsymbol{x}'$ be the permutation of $\boldsymbol{x}$ such that for some $1 \leq m < m' \leq k/n$, it satisfies

(i) $\tilde{\boldsymbol{y}}_m = \tilde{\boldsymbol{y}}_{m'}$
(ii) $\sigma_0(m-1) = \sigma_0(m'-1)$
(iii) $\sigma_0(m) = \sigma_0(m')$.

For such a permutation, it is clear from (A·1) that

$$p(\boldsymbol{y}, \sigma_0|\boldsymbol{x}', s_0) = p(\boldsymbol{y}, \sigma_0|\boldsymbol{x}, s_0). \qquad (\mathrm{A}\cdot 3)$$

Moreover, we can lower bound the transition probability as follows

$$\begin{aligned} p_\theta(\boldsymbol{y}|\boldsymbol{x}') &= \sum_{\tilde{\boldsymbol{s}} \in S^{n/k}} p(\boldsymbol{y}, \tilde{\boldsymbol{s}}|\boldsymbol{x}', s_0) \\ &\overset{(a)}{\geq} p(\boldsymbol{y}, \sigma_0|\boldsymbol{x}', s_0) \\ &\overset{(b)}{=} p(\boldsymbol{y}, \sigma_0|\boldsymbol{x}, s_0) \\ &\overset{(c)}{\geq} \alpha_n^{-1} p_\theta(\boldsymbol{y}|\boldsymbol{x}), \end{aligned}$$

where (a) is obvious because we consider only one specific state sequence, (b) follows from (A·3), and (c) comes from (A·2). This implies that $\boldsymbol{x}' \in S_{TH}(\boldsymbol{x}, \boldsymbol{y})$.

The next step is to count how many such permutations (satisfying (i), (ii) and (iii)) exist. Before that, we find it useful to define $c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')$ as the number of joint occurrences of $\tilde{\boldsymbol{x}} \in \mathcal{X}^k$ and $\tilde{\boldsymbol{y}} \in \mathcal{Y}^k$ that end in state s and are preceded by state s'. That is,

$$\begin{aligned} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \overset{\triangle}{=} |\{i \in \{1, \cdots, n/k\} : \tilde{\boldsymbol{x}}_i = \tilde{\boldsymbol{x}}, \\ \tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}, \sigma_0(i) = s, \sigma_0(i-1) = s'\}|. \end{aligned}$$

For easier notation, we also define

$$c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \overset{\triangle}{=} \sum_{(s,s') \in S^2} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') = \frac{n}{k} P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}),$$

$$c(\tilde{\boldsymbol{y}}) \overset{\triangle}{=} \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{n}{k} \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} P_{\boldsymbol{xy}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}).$$

Then, by restricting our attention to the permutations mentioned above (which are all included in $S_{TH}(\boldsymbol{x}, \boldsymbol{y})$), we have

$$|S_{TH}(\boldsymbol{x}, \boldsymbol{y})| \geq \prod_{\tilde{\boldsymbol{y}} \in \mathcal{Y}^k} \prod_{(s,s') \in S^2} \frac{\{\sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')\}!}{\prod_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')!}.$$

Taking the logarithms of both sides and using Stirling's formula, we obtain

$$\begin{aligned} &\log\left(|S_{TH}(\boldsymbol{x}, \boldsymbol{y})|\right) \\ &\geq \sum_{\tilde{\boldsymbol{y}}} \sum_{(s,s')} \left[\log\Big\{ \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \Big\}! \right. \\ &\qquad \left. - \sum_{\tilde{\boldsymbol{x}}} \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')! \right] \\ &\geq \sum_{\tilde{\boldsymbol{y}}} \sum_{(s,s')} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \\ &\qquad \times \left\{ \log \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') - \log e \right\} \\ &\qquad - \sum_{\tilde{\boldsymbol{y}}} \sum_{(s,s')} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \\ &\overset{(a)}{\geq} \sum_{\tilde{\boldsymbol{y}}} \sum_{(s,s')} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \log \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \end{aligned}$$

$$
\begin{aligned}
&- \sum_{\tilde{\boldsymbol{y}}} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) - \frac{n}{k} \log e \\
&= \sum_{\tilde{\boldsymbol{y}}} c(\tilde{\boldsymbol{y}}) \sum_{(s,s')} \frac{\sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')}{c(\tilde{\boldsymbol{y}})} \\
&\quad \times \left[\log \frac{\sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')}{c(\tilde{\boldsymbol{y}})} + \log c(\tilde{\boldsymbol{y}}) \right] \\
&\quad - \sum_{\tilde{\boldsymbol{y}}} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) - \frac{n}{k} \log e \\
&\overset{(b)}{\geq} \sum_{\tilde{\boldsymbol{y}}} c(\tilde{\boldsymbol{y}}) \frac{\sum_{(s,s')} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')}{c(\tilde{\boldsymbol{y}})} \\
&\quad \times \log \frac{\sum_{(s,s')} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s')}{c(\tilde{\boldsymbol{y}}) \times |S|^2} \\
&\quad + \sum_{\tilde{\boldsymbol{y}}} c(\tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{y}}) \\
&\quad - \sum_{\tilde{\boldsymbol{y}}} \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) - \frac{n}{k} \log e \\
&= \sum_{\tilde{\boldsymbol{y}}} \Big[- c(\tilde{\boldsymbol{y}}) \log |S|^2 + c(\tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{y}}) \\
&\quad - \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \Big] - \frac{n}{k} \log e \\
&= \sum_{\tilde{\boldsymbol{y}}} \left[c(\tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{y}}) - \sum_{\tilde{\boldsymbol{x}}} c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \log c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \right] \\
&\quad - \frac{n}{k} \log |S|^2 e \\
&= \frac{n}{k} \left[\sum_{\tilde{\boldsymbol{y}}} \frac{c(\tilde{\boldsymbol{y}})}{n/k} \left(\log \frac{c(\tilde{\boldsymbol{y}})}{n/k} + \log \frac{n}{k} \right) - \log |S|^2 e \right. \\
&\quad \left. - \sum_{\tilde{\boldsymbol{y}}} \sum_{\tilde{\boldsymbol{x}}} \frac{c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}{n/k} \left(\log \frac{c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}{n/k} + \log \frac{n}{k} \right) \right] \\
&= \frac{n}{k} H(\boldsymbol{x}|\boldsymbol{y}, k) - \frac{n}{k} \log |S|^2 e,
\end{aligned}
$$

where (a) comes from the fact $c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, s, s') \leq c(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$, and (b) comes from Jensen inequality. Q.E.D.

Proof of Lemma 6

Consider the lossless source coding problem of $\hat{\boldsymbol{x}} \in \mathcal{X}^n$, when the side information $\boldsymbol{y} \in \mathcal{Y}^n$ is known at both encoder and decoder. If we use the Shannon-Fano code based on the conditional probability

$$
P_{\hat{\boldsymbol{x}}\boldsymbol{y}}(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{y}}) = \frac{P_{\hat{\boldsymbol{x}}\boldsymbol{y}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}{\sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} P_{\hat{\boldsymbol{x}}\boldsymbol{y}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})},
$$

the codelength $L(\hat{\boldsymbol{x}}|\boldsymbol{y})$ for $\hat{\boldsymbol{x}}$ satisfies

$$
\begin{aligned}
\frac{1}{n} L(\hat{\boldsymbol{x}}|\boldsymbol{y}) \leq\ & \frac{1}{k} H(\hat{\boldsymbol{x}}|\boldsymbol{y}, k) + \frac{2}{k} \\
& + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right). \qquad (\mathrm{A} \cdot 4)
\end{aligned}
$$

Let $H(\boldsymbol{y}, k)$ denote the empirical entropy associated with the empirical distribution

$$
P_{\boldsymbol{y}}(\tilde{\boldsymbol{y}}) = \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^k} P_{\boldsymbol{x}\boldsymbol{y}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}), \quad \forall \tilde{\boldsymbol{y}} \in \mathcal{Y}^k.
$$

Since $H(\boldsymbol{y}, k)$ does not depend on $\hat{\boldsymbol{x}} \in \mathcal{X}^n$, we can see that $H(\hat{\boldsymbol{x}}\boldsymbol{y}, k) = H(\hat{\boldsymbol{x}}|\boldsymbol{y}, k) + H(\boldsymbol{y}, k)$. Hence, (A·4) can be written as

$$
\begin{aligned}
\frac{1}{n} L(\hat{\boldsymbol{x}}|\boldsymbol{y}) \leq\ & \frac{1}{k} H(\hat{\boldsymbol{x}}\boldsymbol{y}, k) - \frac{1}{k} H(\boldsymbol{y}, k) + \frac{2}{k} \\
& + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right). \qquad (\mathrm{A} \cdot 5)
\end{aligned}
$$

On the other hand, $\ell(f_n(\hat{\boldsymbol{x}}, \boldsymbol{y}, k)) \leq \ell(f_n(\boldsymbol{x}, \boldsymbol{y}, k))$ implies

$$
H(\hat{\boldsymbol{x}}\boldsymbol{y}, k) \leq H(\boldsymbol{x}\boldsymbol{y}, k) + 2\epsilon(n, k) \times k.
$$

By substituting this into (A·5), we obtain

$$
\begin{aligned}
\frac{1}{n} L(\hat{\boldsymbol{x}}|\boldsymbol{y}) \leq\ & \frac{1}{k} H(\boldsymbol{x}\boldsymbol{y}, k) + 2\epsilon(n, k) - \frac{1}{k} H(\boldsymbol{y}, k) \\
& + \frac{2}{k} + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right) \\
=\ & \frac{1}{k} H(\boldsymbol{x}|\boldsymbol{y}, k) + 2\epsilon(n, k) + \frac{2}{k} \\
& + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right).
\end{aligned}
$$

Therefore, for a given $(\boldsymbol{x}, \boldsymbol{y})$, the number of sequences $\hat{\boldsymbol{x}} \in \mathcal{X}^n$ satisfying $\ell(f_n(\hat{\boldsymbol{x}}, \boldsymbol{y}, k)) \leq \ell(f_n(\boldsymbol{x}, \boldsymbol{y}, k))$ can be upper bounded by

$$
\begin{aligned}
& \exp \left\{ n \left[\frac{1}{k} H(\boldsymbol{x}|\boldsymbol{y}, k) + 2\epsilon(n, k) + \frac{2}{k} \right. \right. \\
& \left. \left. + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right) \right] \right\}.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
|S_{\hat{f}_n}(\boldsymbol{x}, \boldsymbol{y})| \leq\ & \exp \left\{ n \left[\frac{1}{k} H(\boldsymbol{x}|\boldsymbol{y}, k) + 2\epsilon(n, k) + \frac{2}{k} \right. \right. \\
& \left. \left. + |\mathcal{X} \times \mathcal{Y}|^k \left(\frac{\log(n/k) + 1}{n} \right) \right] \right\},
\end{aligned}
$$

which completes the proof. Q.E.D.

Tomohiko Uyematsu received the B.E., M.E. and D.E. degrees from Tokyo Institute of Technology in 1982, 1984 and 1988, respectively. From 1984 to 1992, he was with the Department of Electrical and Electronic Engineering of Tokyo Institute of Technology, first as research associate, next as lecturer, and lastly as associate professor. From 1992 to 1997, he was with School of Information Science of Japan Advanced Institute of Science and Technology as associate professor. Since 1997, he returned to the Department of Electrical and Electronic Engineering of Tokyo Institute of Technology as associate professor. In 1992 and 1996, he was a visiting researcher at the Centre National de la Recherche Scientifique, France and Delft University of Technology, Netherlands, respectively. He received Shinohara Memorial Young Engineer Award in 1989, and the Best Paper Award in 1993 and 1996, both from IEICE. His current research interests are in the areas of information theory, especially Shannon theory and multi-terminal information theory. Dr. Uyematsu is a member of IEEE.

Saad M. Kisra was born in Rabat, Morocco in 1975. He attended two terms at l'école préparatoire pour les concours des grandes écoles d'ingénieurs at Rabat. He graduated from the Japanese language center affiliated to Tokyo University for Foreign Studies in 1995. He received his B.E. in International Development of Eng. and M.E. in Electronics and Electrical Eng. from Tokyo Institute of Technology in 1999 and 2001 respectively. Since April 2001, he is with Schlumberger, Co., Ltd., Dept. of Reservoir Evaluation.