Source Coding Using Families of Universal Hash Functions

Hiroki Koga, *Member, IEEE*

***Abstract*—This correspondence is concerned with new connections between source coding and two kinds of families of hash functions known as the families of universal hash functions and N-strongly universal hash functions, where $N \geq 2$ is an integer. First, it is pointed out that such families contain classes of well-known source codes such as bin codes and linear codes. Next, performance of a source coding scheme using either of the two kinds of families is evaluated. An upper bound on the expectation of the decoding error probability is obtained for each family. The expectation of the decoding error probability is analyzed in detail for the cases of discrete memoryless sources and sources without the memoryless assumption under a certain class of decoders.**

***Index Terms*—Bin coding, error exponent, linear coding, strongly universal hash functions, universal hash functions.**

I. Introduction

The family $\mathcal{F}$ of universal hash functions, which was first proposed by Carter and Wegman [2], is a collection of mappings from a finite set $\mathcal{A}$ to another finite set $\mathcal{B}$, where the cardinalities of $\mathcal{A}$ and $\mathcal{B}$, say A and B, respectively, are assumed to satisfy $A \geq B$. Letting f be an arbitrary mapping in $\mathcal{F}$, we say that *collision* occurs if $f(a) = f(a')$ for some distinct $a, a' \in \mathcal{A}$. A requirement on collision is imposed on the family of universal hash functions because collision can cause some problems in applications of hash functions. Different requirements lead to various kinds of families of hash functions [2], [8], [17], [18]. While hash functions are usually used for storage and retrieval of information, families of hash functions are often studied in several contexts in cryptography such as authentication [8], [17], [18], privacy amplification [1] and secret-key agreement [15].

However, families of hash functions rarely appear in Shannon theory or source coding. While Kurosawa and Yoshida [12] construct an identification code based on a certain family of hash functions, coding theorems based on families of hash functions have not been studied. On the other hand, Muramatsu [16] recently proposed a source coding algorithm in which both an encoder and a decoder share randomness and synchronously update respective codebooks according to the randomness. Muramatsu's algorithm suggests encoding using hash functions because the process of encoding can be regarded as a time-varying hashing. We should also note that a relationship between source coding and a linear hash function is clearly mentioned in Mackay's textbook on information theory [14].

The objective of this correspondence is the investigation of new connections between source coding and families of hash functions. We are interested in the families known as families of universal hash functions [2] and N-strongly universal hash functions [18], where N is an integer satisfying $1 \leq N \leq A$. We apply the two families to source coding and evaluate the decoding error probability. Letting $\mathcal{F}$ be a family of (N-strongly) universal hash functions, an encoder and a decoder share a hash function $f : \mathcal{A} \to \mathcal{B}$ in $\mathcal{F}$ randomly chosen subject to the uniform distribution. Given a source output $X \in \mathcal{A}$, the encoder computes a codeword $Y = f(X)$ and transmits Y to the decoder. The decoder can correctly decode Y if no collision occurs, say, the other elements in $\mathcal{A}$ are not mapped to Y. Hence, in order to evaluate the decoding error probability, we need to evaluate the probability of such collision.

These families of hash functions can be regarded as generalizations of certain classes of source codes. In fact, in the random coding arguments for establishing the direct parts of coding theorems, we often use bin coding [3] and linear coding [5]. We point out that, in coding of a single source, the class of linear codes is one of the families of universal hash functions. We also see that the class of bin codes is regarded as a family of N-strong universal hash functions with $N = A$ as well as a family of universal hash functions. In this correspondence, we give two upper bounds on the expectation of the decoding error probability. One of the upper bounds is valid for any family of universal hash functions. That is, both linear coding and bin coding meet the upper bound. The other upper bound is smaller and is valid for any family of N-strongly hash functions under a certain condition on N. In fact, the bin coding meets this smaller upper bound. We analyze the two upper bounds in detail for stationary memoryless sources and obtain an attainable error exponent that is the same as in [5].

The correspondence is organized as follows. In Section II, we define families of universal hash functions and N-strongly universal hash functions. Important examples of the families of hash functions are given. In Section III, we define encoding and decoding of a source output X and evaluate the expectation of the decoding error probability by using a combinatoric argument. Coding of stationary memoryless sources is discussed in Section IV. An achievable error exponent is obtained by using the methods of the types [7]. In Section V, we consider sources without the memoryless assumption and evaluate the expectation of the decoding error probability of a decoder in a wide class. We take an approach from information-spectrum methods [10] and give a sufficient condition under which the expectation vanishes as the blocklength increases.

II. Families of Universal Hash Functions

Throughout the correspondence let $\mathcal{A}$ and $\mathcal{B}$ be finite sets. Denote the cardinalities of $\mathcal{A}$ and $\mathcal{B}$ by A and B, respectively. We assume that $A \geq B$. Let $\mathcal{F}$ be a finite set of mappings $f : \mathcal{A} \to \mathcal{B}$. Carter and Wegman [2] defined a family of universal hash functions as follows.

Definition 1: We call $\mathcal{F}$ a family of universal hash functions (or, simply, universal) if for any distinct $a_1, a_2 \in \mathcal{A}$ it holds that

$$|\{f \in \mathcal{F} : f(a_1) = f(a_2)| \leq \frac{|\mathcal{F}|}{B} \tag{1}$$

where $|\cdot|$ denotes the cardinality of the set.

We give two important families of universal hash functions.

Example 1: Let $\mathcal{F}_1$ be the set of all the mappings from $\mathcal{A}$ to $\mathcal{B}$. Then, $\mathcal{F}_1$ is a family of universal hash functions.

Example 2: Suppose that $\mathcal{A} = (\mathrm{GF}(q))^n$ and $\mathcal{B} = (\mathrm{GF}(q))^k$ for some integers $n \geq k \geq 1$, where GF(q) denotes a finite field with q elements. Then, the set $\mathcal{F}_2$ of all the linear mappings from $\mathcal{A}$ to $\mathcal{B}$ is a family of universal hash functions.

Example 1 is given in Wegman and Carter [18] not as a family of universal hash functions but as a family of N-strongly universal hash functions that will be defined afterward. As is claimed in Proposition 2 below, any family of N-strongly universal hash functions with $N \geq 2$ is universal.

Example 2 is due to Csiszár [5] in which linear coding of two correlated sources is discussed. We can verify that $\mathcal{F}_2$ is universal in the

Manuscript received January 25, 2006; revised April 26, 2007. The material of this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

The author is with Graduate School of Systems and Information Engineering, University of Tsukuba, Ibaraki 305-8573, Japan (e-mail: koga@esys.tsukuba.ac.jp).

Communicated by V. A. Vaishampayan, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2007.903113

following way. Letting $\boldsymbol{x} \in \mathcal{A}$ be a row vector and $\boldsymbol{M} \in (\mathrm{GF}(q))^{nk}$ an n-by-k matrix, consider a mapping $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{M}$, where $\boldsymbol{y} \in \mathcal{B}$ is a row vector. Since $\mathcal{F}_2$ is all the collection of linear mappings, there is a bijection between $\mathcal{F}_2$ and the set of all the n-by-k matrices. This implies that $|\mathcal{F}_2| = q^{nk}$. Next, we evaluate the number of matrices $\boldsymbol{M}$ satisfying $\boldsymbol{x}_1\boldsymbol{M} = \boldsymbol{x}_2\boldsymbol{M}$ for arbitrarily given $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{A}$ satisfying $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$. Since we have $(\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{M} = \boldsymbol{0}$ and $\boldsymbol{x}_1 - \boldsymbol{x}_2 \neq \boldsymbol{0}$, it suffices to evaluate the number of matrices $\boldsymbol{M} = (m_{ij})$ satisfying $\boldsymbol{x}\boldsymbol{M} = \boldsymbol{0}$ for an arbitrarily given nonzero vector $\boldsymbol{x} \in \mathcal{A}$. Notice that $\boldsymbol{x}\boldsymbol{M} = \boldsymbol{0}$ holds by adequate choice of $m_{i^*j}, j = 1, 2, \ldots, k$, where i^* means one of indices of nonzero components of $\boldsymbol{x}$. The other $(n-1)k$ components of $\boldsymbol{M}$ can be chosen arbitrarily. Therefore, it holds that

$$|\{f \in \mathcal{F}_2 : f(\boldsymbol{x}_1) \neq f(\boldsymbol{x}_2)\}| = q^{(n-1)k} = \frac{q^{nk}}{q^k} = \frac{|\mathcal{F}_2|}{B}$$

which shows that $\mathcal{F}_2$ is universal.

Wegman and Carter [18] introduced a family of N-strongly universal hash functions as follows:

Definition 2: Let N be an integer satisfying $1 \leq N \leq A$. We call $\mathcal{F}$ a family of N-strongly universal hash functions (or, simply, N-SU) if for any distinct $a_1, a_2, \ldots, a_N \in \mathcal{A}$ and for any $b_1, b_2, \ldots, b_N \in \mathcal{B}$ it holds that

$$|\{f \in \mathcal{F}\colon f(a_i) = b_i \text{ for all } i = 1, 2, \ldots, N\}| \leq \frac{|\mathcal{F}|}{B^N}. \qquad (2)$$

Wegman and Carter [18] give the following two families as examples of N-SU hash functions.

Example 3: The class $\mathcal{F}_1$ in Example 1 is N-SU for every $1 \leq N \leq A$.

It is easily checked that, if $\mathcal{F} = \mathcal{F}_1$, the left-hand side of (2) equals B^{A-N}. This means that (2) holds with equality.

Example 4: Suppose that both $\mathcal{A} = \mathcal{B} = \mathrm{GF}(q)$. For a fixed $(\alpha_0, \alpha_1, \ldots, \alpha_{N-1}) \in (\mathrm{GF}(q))^N$ we define $f : \mathcal{A} \to \mathcal{B}$ as the mapping

$$f : x \mapsto \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{N-1} x^{N-1}$$

where the above additions and multiplications are the operations of $\mathrm{GF}(q)$. Then, $\mathcal{F}_3 = \{f : (\alpha_0, \alpha_1, \ldots, \alpha_{N-1}) \in \mathrm{GF}(q)^N\}$ is N-SU.

The family $\mathcal{F}_3$ in Example 4 is valid for an arbitrarily fixed $1 \leq N \leq A$ under the requirement of $A = B$. However, it is not clear whether or not there exists a family $\mathcal{F} \neq \mathcal{F}_1$ of N-SU hash functions that is defined for the case of $A > B$ and makes sense for an arbitrarily given $1 \leq N \leq A$. In order to guarantee the existence of such a family, we extend Definition 2. Letting $\mathcal{J}$ be a finite set of indices, let $\mathcal{F} = \{f_j\}_{j \in \mathcal{J}}$ be a collection of mappings $f_j : \mathcal{A} \to \mathcal{B}$. We can use the following definition Instead of Definition 2.

Definition 3: Let N be an arbitrary integer satisfying $1 \leq N \leq A$. A collection of mappings $\mathcal{F} = \{f_j\}_{j \in \mathcal{J}}$ is called a family of N-strongly universal hash functions in the extended sense (N-ESU hash functions for short) if for any distinct $a_1, a_2, \ldots, a_N \in \mathcal{A}$ and for any $b_1, b_2, \ldots, b_N \in \mathcal{B}$ it holds that

$$|\{j \in \mathcal{J}\colon f_j(a_i) = b_i, \text{ for all } i = 1, 2, \ldots, N\}| \leq \frac{|\mathcal{J}|}{B^N}. \qquad (3)$$

Note that, if $\mathcal{F} = \{f_j\}_{j \in \mathcal{J}}$ is N-ESU, then f_j may be equal to $f_{j'}$ for some $j' \neq j$. In other words, a family $\mathcal{F}$ of N-ESU hash functions is N-SU if all $f_j, j \in \mathcal{J}$, are distinct.

We give an example of a family of N-ESU hash functions. A rough idea of this example can be found in [18], though clear definition is not given.

Example 5: Letting n and k be arbitrary integers satisfying $0 < k < n$, we set $\mathcal{A} = \mathrm{GF}(2^n)$ and $\mathcal{B} = \{0, 1\}^k$. Fix an integer N satisfying $1 \leq N \leq 2^n$. For a fixed $(\alpha_0, \alpha_1, \ldots, \alpha_{N-1}) \in \mathrm{GF}(2^n)^N$ we define a mapping $f_j : \mathcal{A} \to \mathcal{B}$ by

$$f_j : x \mapsto \mathrm{LSB}_k[\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{N-1} x^{N-1}]$$

where $j = (\alpha_0, \alpha_1, \ldots, \alpha_{N-1})$ and $\mathrm{LSB}[y]_k$ for $y \in \mathrm{GF}(2^n)$ means the k least significant bits of y when y is expressed in the binary form. Then, $\mathcal{F}_4 = \{f_j\}_{j \in \mathcal{J}}$ is N-ESU with $\mathcal{J} = (\mathrm{GF}(2^n))^N$.

It is easily checked that $\mathcal{F}_4$ Example 5 is N-ESU. Let $a_1, a_2, \ldots, a_N$ be arbitrary distinct elements of $\mathcal{A}$ and $b_1, b_2, \ldots, b_N$ arbitrary elements in $\mathcal{B}$. For each $i = 1, 2, \ldots, N$ we define $\tilde{b}_i$ as an arbitrary element of $\mathrm{GF}(2^n)$ satisfying $b_i = \mathrm{LSB}_k[\tilde{b}_i]$. Since for each i there are 2^{n-k} choices of such $\tilde{b}_i$'s, there are $2^{N(n-k)}$ choices of $\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_N$. It is important to notice that the system of N linear equations

$$\tilde{b}_i = \alpha_0 + \alpha_1 a_i + \cdots + \alpha_{N-1} a_i^{N-1}, \quad i = 1, \ldots, N$$

has a unique solution $(\alpha_0, \alpha_1, \ldots, \alpha_{N-1})$ because the associated Vandermonde matrix is always invertible for any distinct $a_1, a_2, \ldots, a_N$. Thus, the left-hand side of (3) turns out to be equal to $2^{N(n-k)}$, which is equal to the right-hand side of (3) because $|\mathcal{J}| = 2^{Nn}$ and $B = 2^k$.

Hereafter, for notational convenience, we regard families of N-ESU hash functions as an extended (but implicit) interpretation of families of N-SU hash functions. In the following sections we will give results on families of N-SU hash functions. All of such results are valid for families of N-ESU hash functions as well. However, giving rigorous proofs for the results on families of N-ESU hash functions are easy and therefore omitted.

We give a simple proposition on families of N-SU hash functions.

Proposition 1: Let N and N' be integers satisfying $0 < N' < N \leq A$. If $\mathcal{F}$ is N-SU, then $\mathcal{F}$ is N'-SU.

Proof: It suffices to prove that $\mathcal{F}$ is $(N-1)$-SU if $\mathcal{F}$ is N-SU. Let $a_1, a_2, \ldots, a_{N-1}$ be arbitrary distinct elements in $\mathcal{A}$ and $b_1, b_2, \ldots, b_{N-1}$ arbitrary elements in $\mathcal{B}$. The following relationship is a key to the proof

$$\begin{aligned}&\{f \in \mathcal{F} : f(a_i) = b_i \text{ for all } i = 1, 2, \ldots, N-1\}\\ &\qquad = \bigcup_{b_N \in \mathcal{B}} \{f \in \mathcal{F} : f(a_i) = b_i \text{ for all } i = 1, 2, \ldots, N\}. \qquad (4)\end{aligned}$$

Note that all the sets on the right-hand side of (4) are disjoint. Then, it follows from (4) that

$$\begin{aligned}&|\{f \in \mathcal{F} : f(a_i) = b_i \text{ for all } i = 1, 2, \ldots, N-1\}|\\ &\quad = \sum_{b_N \in \mathcal{B}} |\{f \in \mathcal{F} : f(a_i) = b_i \text{ for all } i = 1, 2, \ldots, N\}|\\ &\quad \leq \frac{|\mathcal{F}|}{B^{N-1}}\end{aligned}$$

where the inequality follows from (2). □

We can also establish the following proposition by using the same idea as in the proof of Proposition 1.

Proposition 2: If $\mathcal{F}$ is N-SU for some $N \geq 2$, then $\mathcal{F}$ is universal.

Proof: Proposition 1 guarantees that $\mathcal{F}$ is 2-SU. Hence, for any distinct $a_1, a_2 \in \mathcal{A}$ it holds that

$$\begin{aligned}&|\{f \in \mathcal{F} : f(a_1) = f(a_2)\}|\\ &\quad = \sum_{b \in \mathcal{B}} |\{f \in \mathcal{F} : f(a_1) = f(a_2) = b\}|\\ &\quad \leq \frac{|\mathcal{F}|}{B}\end{aligned}$$

where the inequality holds because $\mathcal{F}$ is 2-SU. □

Stinson [17] gives two families of 2-SU hash functions. Proposition 2 tells us that the two families are universal as well.

III. Source Coding Using Families of Hash Functions

In this section we apply a family $\mathcal{F}$ of universal hash functions or N-SU hash functions to source coding. Let X be a random variable taking values in $\mathcal{A}$. Denote the probability distribution of X by P_X. Let $\mathcal{F}$ be a family of (N-strongly) universal hash functions.

We define encoding and decoding of X. Suppose that an encoder and a decoder share an $f \in \mathcal{F}$ that is randomly chosen subject to the uniform distribution. Let $\mathcal{T}$ be an arbitrary nonempty subset of $\mathcal{A}$ and a_0 an arbitrary element of $\mathcal{A}$ chosen in advance by the decoder. Denote by T the cardinality of $\mathcal{T}$. We use the notation $f^{-1}(b) = \{a \in \mathcal{A} : f(a) = b\}$ for $f \in \mathcal{F}$ and $b \in \mathcal{B}$. The encoder computes a codeword $Y \overset{\text{def}}{=} f(X)$ and transmits Y to the decoder, while the decoder outputs $\hat{X} \in \mathcal{X}$ that is equal to the unique element of $f^{-1}(Y) \cap \mathcal{T}$ if $|f^{-1}(Y) \cap \mathcal{T}| = 1$ and a_0 otherwise.

In the above coding scheme, the coding rate is $\log B$, where throughout this correspondence all the logarithms are to the base 2. Denote by P_e the decoding error probability $\Pr\{X \neq \hat{X}\}$. We have the following theorem on P_e.

Theorem 1:

A) Let $\mathcal{F}$ be an arbitrary family of universal hash functions. Then, for any nonempty subset $\mathcal{T}$ of $\mathcal{A}$ it holds that

$$E[P_e] \leq \Pr\{X \notin \mathcal{T}\} + \Pr\{X \in \mathcal{T}\} \frac{T-1}{B} \tag{5}$$

where $E[\cdot]$ denotes the expectation with respect to the random choice of $f \in \mathcal{F}$ subject to the uniform distribution.

B) Let $\mathcal{F}$ be an arbitrary family of N-SU hash functions. Then, for any subset $\mathcal{T}$ of $\mathcal{A}$ with $1 \leq T \leq N$ it holds that

$$E[P_e] \leq \Pr\{X \notin \mathcal{T}\} + \Pr\{X \in \mathcal{T}\} \left[1 - \left(1 - \frac{1}{B}\right)^{T-1}\right] \tag{6}$$

where $E[\cdot]$ denotes the expectation with respect to the random choice of $f \in \mathcal{F}$ subject to the uniform distribution.

Theorem 1-A) and -B) give upper bounds on $E[P_e]$ that are dependent on P_X and $\mathcal{T}$. Each upper bound consists of two terms. The first term corresponds to the probability that the source outputs an element not belonging to $\mathcal{T}$. The second term corresponds to the probability of X that belongs to $\mathcal{T}$ but is not correctly decoded because of collision.

Let us compare the two upper bounds. Letting $m \geq 1$ be an arbitrary fixed integer, we can easily prove the inequality $1 - (1-u)^m \leq mu$ for any $u \in (0, 1)$. Thus, the upper bound in (6) is smaller than the upper bound in (5) when the same $\mathcal{T}$ is used in decoding. This means that $E[P_e]$ can be smaller if we use a family of N-SU hash functions. However, note that Theorem 1-B) says nothing when we use $\mathcal{T} \subset \mathcal{A}$ with $T > N$. While $\mathcal{T}$ must satisfy $1 \leq T \leq N$ in Theorem 1-B), $\mathcal{T}$ can be an arbitrary subset of $\mathcal{A}$ in Theorem 1-A).

If we encode a source output X by bin coding [3], we first randomly assign an element of $\mathcal{B}$ to every element of $\mathcal{A}$ subject to the uniform distribution. This random assignment corresponds to the random choice of $f \in \mathcal{F}_1$ subject to the uniform distribution. Note that, for the case of $f \in \mathcal{F}_1$ the claim of Theorem 1-B) is valid for any $\mathcal{T} \subset \mathcal{A}$. Theorem 1-B) tells us that the existence of a class of codes with a property similar to the bin code.

Of course, an immediate consequence of Theorem 1-A) (respectively, Theorem 1-B)) is the existence of an $f \in \mathcal{F}$ such that P_e is bounded by the right-hand side of (5) (respectively, (6)).

We use the following lemma in the proof of Theorem 1-A).

Lemma 1: Let $\mathcal{F}$ be an arbitrary family of universal hash functions. Let $\mathcal{T}$ be an arbitrary nonempty subset of $\mathcal{A}$ with the cardinality T. Then, for any $a \in \mathcal{T}$ it holds that

$$\frac{|\{f \in \mathcal{F} : f(a) = f(a') \text{ for some } a' \in \mathcal{T} \text{ and } a' \neq a\}|}{|\mathcal{F}|} \leq \frac{T-1}{B}. \tag{7}$$

Proof: Since (7) is trivial if $T = 1$, we can assume that $T \geq 2$. Fix $\mathcal{F}$ and $\mathcal{T}$ arbitrarily. Define

$$\chi_f(a, a') = \begin{cases} 1, & \text{if } f(a) = f(a') \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\chi_f^*(a) = \begin{cases} 1, & \text{if } f(a) = f(a') \text{ for some } a' \in \mathcal{T}, a' \neq a \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

for $f \in \mathcal{F}$ and $a, a' \in \mathcal{T}$. Then, in view of the definitions of $\chi_f(a, a')$ and $\chi_f^*(a)$, we have

$$\chi_f^*(a) \leq \sum_{a' \in \mathcal{T}, a' \neq a} \chi_f(a, a'), \quad \text{for all } f \in \mathcal{F} \text{ and } a \in \mathcal{T}. \tag{10}$$

In addition, since $\mathcal{F}$ is universal, it holds that

$$\sum_{f \in \mathcal{F}} \chi_f(a, a') = |\{f \in \mathcal{F} : f(a) = f(a')\}| \leq \frac{|\mathcal{F}|}{B}, \quad \text{for all } a, a' \in \mathcal{T} \text{ and } a' \neq a. \tag{11}$$

Hence, in view of (10) and (11) we have

$$\begin{aligned} &|\{f \in \mathcal{F} : f(a) = f(a') \text{ for some } a' \in \mathcal{T} \text{ and } a' \neq a\}| \\ &= \sum_{f \in \mathcal{F}} \chi_f^*(a) \\ &\leq \sum_{f \in \mathcal{F}} \sum_{a' \in \mathcal{T}, a' \neq a} \chi_f(a, a') \\ &\leq \sum_{a' \in \mathcal{T}, a' \neq a} \frac{|\mathcal{F}|}{B} \\ &\leq |\mathcal{F}| \frac{T-1}{B} \end{aligned}$$

which is equivalent to (7) □

We use the following lemma in the proof of Theorem 1-B). We prove this lemma by using a combinatoric argument.

Lemma 2: Let $\mathcal{F}$ be an arbitrary family of N-SU hash functions. Let $\mathcal{T}$ be an arbitrary subset of $\mathcal{A}$ with the cardinality T. If $1 \leq T \leq N$, then for any $a \in \mathcal{A}$ it holds that

$$\frac{|\{f \in \mathcal{F} : f(a) = f(a') \text{ for some } a' \in \mathcal{T} \text{ and } a' \neq a\}|}{|\mathcal{F}|} \leq 1 - \left(1 - \frac{1}{B}\right)^{T-1}. \tag{12}$$

Proof: We consider the case of $T \geq 2$ because (12) is trivial if $T = 1$. Set $\mathcal{A} = \{a_1, a_2, \ldots, a_A\}$ and $\mathcal{B} = \{b_1, b_2, \ldots, b_B\}$. Without loss of generality, we can assume that $\mathcal{T} = \{a_1, a_2, \ldots, a_T\}$ and $a = a_1$. Define

$$\begin{aligned} \mathcal{E} &= \{f \in \mathcal{F} : f(a_1) = f(a_i) \text{ for some } i = 2, 3, \ldots, T\} \\ \mathcal{E}_j &= \{f \in \mathcal{E} : f(a_1) = b_j\}, \quad j = 1, 2, \ldots, B. \end{aligned}$$

Clearly, $\mathcal{E}_j, j = 1, 2, \ldots, B$, form a partition of $\mathcal{E}$. Therefore, it holds that

$$|\mathcal{E}| = \sum_{j=1}^{B} |\mathcal{E}_j|. \tag{13}$$

We evaluate $|\mathcal{E}_1|$ first. Letting $\{\mathcal{T}_j\}_{j=1}^{B}$ be an arbitrary partition of $\mathcal{T}$, define

$$\mathcal{F}\left(\{\mathcal{T}_j\}_{j=1}^{B}\right) = \{f \in \mathcal{F} : f(a) = b_j \text{ for all } a \in \mathcal{T}_j \text{ and } j = 1, 2, \ldots, B\}.$$

Since $\mathcal{F}$ is assumed to be N-SU and therefore T-SU from Proposition 1 and $T \le N$, we have

$$\left|\mathcal{F}\left(\{\mathcal{T}_j\}_{j=1}^{B}\right)\right| \le \frac{|\mathcal{F}|}{B^T} \tag{14}$$

from Definition 3. It is important to notice the fact that $\mathcal{E}_1$ is the union of $\mathcal{F}(\{\mathcal{T}_j\}_{j=1}^{B})$ with respect to the partitions $\{\mathcal{T}_j\}_{j=1}^{B}$ satisfying $a_1 \in \mathcal{T}_1$ and $|\mathcal{T}_1| \ge 2$.

We can enumerate the number of such partitions of $\mathcal{T}$. In fact, a simple observation tells us that the number of such partitions is equal to $B^{T-1} - (B-1)^{T-1}$. To see this, we consider a partition $\{\tilde{\mathcal{T}}_j\}_{j=1}^{B}$ of $\mathcal{T} \backslash \{a_1\}$ and set $\mathcal{T}_1 = \tilde{\mathcal{T}}_1 \cup \{a_1\}$ and $\mathcal{T}_j = \tilde{\mathcal{T}}_j$ for $j = 2, 3, \ldots, B$. Notice that the number of all the partitions $\{\tilde{\mathcal{T}}_j\}_{j=1}^{B}$ is equal to B^{T-1} and the number of all the partitions $\{\tilde{\mathcal{T}}_j\}_{j=1}^{B}$ satisfying $\tilde{\mathcal{T}}_1 = \phi$ is equal to $(B-1)^{T-1}$. Since $|\mathcal{T}_1| \ge 2$ if and only if $|\tilde{\mathcal{T}}_1| \ge 1$, and all the partitions $\{\mathcal{T}_j\}_{j=1}^{B}$ of $\mathcal{T}$ satisfying $a_1 \in \mathcal{T}_1$ and $|\mathcal{T}_1| \ge 2$ are obtained in this way, we can conclude that the number of all the partitions $\{\mathcal{T}_j\}_{j=1}^{B}$ satisfying $a_1 \in \mathcal{T}_1$ and $|\mathcal{T}_1| \ge 2$ is equal to $B^{T-1} - (B-1)^{T-1}$. Hence, it follows from (14) that

$$\begin{aligned} |\mathcal{E}_1| &\le \frac{|\mathcal{F}|}{B^T}[B^{T-1} - (B-1)^{T-1}] \\ &= \frac{|\mathcal{F}|}{B}\left[1 - \left(1 - \frac{1}{B}\right)^{T-1}\right]. \end{aligned} \tag{15}$$

Obviously, the argument that establishes (15) is valid for evaluation of $|\mathcal{E}_j|$ for $j = 2, 3, \ldots, B$. Therefore, we have

$$|\mathcal{E}_j| \le \frac{|\mathcal{F}|}{B}\left[1 - \left(1 - \frac{1}{B}\right)^{T-1}\right], \quad j = 1, 2, \ldots, B. \tag{16}$$

Then, (13) and (16) yield

$$\begin{aligned} |\mathcal{E}| &\le \sum_{j=1}^{B} \frac{|\mathcal{F}|}{B}\left[1 - \left(1 - \frac{1}{B}\right)^{T-1}\right] \\ &= |\mathcal{F}|\left[1 - \left(1 - \frac{1}{B}\right)^{T-1}\right] \end{aligned}$$

which establishes the claim of this lemma. □

Proof of Theorem 1: First, we evaluate $E[P_e]$ in the following form:

$$\begin{aligned} E[P_e] &= \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{A}} P_X(a) \Pr\{\hat{X} \ne X \,|\, X = a\} \\ &\le \Pr\{X \notin \mathcal{T}\} \\ &\quad + \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{T}} P_X(a) \Pr\{\hat{X} \ne X \,|\, X = a\}. \end{aligned} \tag{17}$$

Notice that $\Pr\{\hat{X} \ne X \,|\, X = a\} \le \chi_f^*(a)$ for any given $f \in \mathcal{F}$ and $a \in \mathcal{T}$, where $\chi_f^*(a)$ is defined in (9). Then, the second term on the right-hand side of (17) is evaluated as

$$\begin{aligned} &\sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{T}} P_X(a) \Pr\{\hat{X} \ne X \,|\, X = a\} \\ &\quad \le \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{T}} P_X(a) \chi_f^*(a) \\ &\quad = \sum_{a \in \mathcal{T}} P_X(a) \frac{|\mathcal{F}(a)|}{|\mathcal{F}|} \end{aligned} \tag{18}$$

where $\mathcal{F}(a) \overset{\text{def}}{=} \{f \in \mathcal{F} : f(a) = f(a') \text{ for some } a' \in \mathcal{T} \text{ and } a' \ne a\}$.

Now, suppose that $\mathcal{F}$ is universal. Then, in view of Lemma 1, (18) leads to

$$\sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{T}} P_X(a) \Pr\{\hat{X} \ne X \,|\, X = a\} \le \Pr\{X \in \mathcal{T}\} \frac{T-1}{B}. \tag{19}$$

By combining (17) with (19), we obtain the claim of Theorem 1-(A). Similarly, application of Lemma 2 to the right-hand side of (18) yields the claim of Theorem 1-(B). □

Let us apply Theorem 1 for coding of a stationary discrete memoryless source. Let $\mathcal{X}$ be a finite alphabet and X a random variable on $\mathcal{X}$ subject to a probability distribution P_X. Denote by $H(X)$ the entropy of X. Suppose that $X^n = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ is a sequence of length n generated from the source. We set $\mathcal{A} = \mathcal{X}^n$ and $\mathcal{B} = \{1, 2, \ldots, \lceil 2^{nR} \rceil\}$ for some $R > 0$, where R specifies the coding rate. Then, Theorem 1 yields the following corollary.

Corollary 1: Let $\gamma > 0$ be an arbitrary constant. Let $\mathcal{F}$ be an arbitrary family of universal hash functions defined for all $n \ge 1$. If $R > H(X)$, then there exists an integer n_0 such that $E[P_e^{(n)}] \le \gamma$ for all $n \ge n_0$, where $P_e^{(n)}$ denotes the decoding error probability of X^n.

This corollary is easily obtained from Theorem 1-A). Since $R > H(X)$, we can choose an $\varepsilon > 0$ satisfying $R > H(X) + 2\varepsilon$. We define $\mathcal{T}$ as the typical set $A_\varepsilon^{(n)}$ defined by

$$A_\varepsilon^{(n)} = \left\{x^n \in \mathcal{X}^n : \left|\frac{1}{n} \log \frac{1}{P_{X^n}(x^n)} - H(X)\right| \le \varepsilon\right\}.$$

It is well-known that $\Pr\{X^n \notin A_\varepsilon^{(n)}\} \to 0$ as $n \to \infty$ and $A_\varepsilon^{(n)}$ satisfies $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$ for all $n \ge 1$ [4]. Hence, it follows from Theorem 1-A) that

$$\begin{aligned} E\left[P_e^{(n)}\right] &\le \Pr\left\{X^n \notin A_\varepsilon^{(n)}\right\} + \Pr\left\{X^n \in A_\varepsilon^{(n)}\right\} \frac{\left|A_\varepsilon^{(n)}\right|}{\lceil 2^{nR} \rceil} \\ &\le \Pr\left\{X^n \notin A_\varepsilon^{(n)}\right\} + 2^{-n\varepsilon} \\ &\to 0, \qquad \text{as } n \to \infty. \end{aligned} \tag{20}$$

On the other hand, if $\mathcal{F}$ is $2^{n\nu}$-SU for some $\nu > H(X)$, then we can obtain the following smaller upper bound form Theorem 1-B)

$$\begin{aligned} E\left[P_e^{(n)}\right] &\le \Pr\left\{X^n \notin A_\varepsilon^{(n)}\right\} \\ &\quad + \Pr\left\{X^n \in A_\varepsilon^{(n)}\right\}\left[1 - \left(1 - \frac{1}{\lceil 2^{nR} \rceil}\right)^{|A_\varepsilon^{(n)}|}\right] \end{aligned} \tag{21}$$

which goes to zero as $n \to \infty$ because the right-hand side of (21) is upper bounded by the right-hand side of (20).

Corollary 1 can be extended to wide classes of sources with finite alphabets. If X^n is an output from a stationary ergodic source, we can also choose the typical set as $\mathcal{T}$. For the case that X^n is an output of a general source $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ [10], the claim of Corollary 1 holds if $R > \overline{H}(\boldsymbol{X})$, where

$$\overline{H}(\boldsymbol{X}) = \text{p-}\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P_{X^n}(X^n)} \tag{22}$$

is the spectrum sup-entropy rate [10]. Here, for a sequence of real-valued random variables $\{Z_n\}_{n=1}^{\infty}$ the limsup in probability is defined by

$$\text{p-}\limsup_{n\to\infty} Z_n = \inf\left\{\alpha : \lim_{n\to\infty} \Pr\{Z_n \le \alpha\} = 1\right\}$$

[10]. We can prove the result for a general source similarly to Corollary 1. We actually choose

$$\mathcal{T}_n = \left\{x^n \in \mathcal{X}^n : \frac{1}{n} \log \frac{1}{P_{X^n}(x^n)} \le \overline{H}(\boldsymbol{X}) + \varepsilon\right\}.$$

as $\mathcal{T}$ and use the fact that $|\mathcal{T}_n| \le 2^{n(\overline{H}(\boldsymbol{X})+\varepsilon)}$, where $\varepsilon > 0$ is a constant satisfying $R \ge \overline{H}(\boldsymbol{X}) + 2\varepsilon$. Recall here that $\overline{H}(\boldsymbol{X})$ has the operational meaning as the infimum of achievable rates of fixed-to-fixed length coding with the vanishing decoding error probability [10].

IV. Performance of the Minimum-Entropy Decoder

The encoder treated in the preceding section encodes a source output X to a codeword $Y = f(X)$ by using an $f \in \mathcal{F}$ shared by the encoder and the decoder. While $\mathcal{F}$ does not depend on the probability distribution P_X of X, a subset $\mathcal{T}$ used by the decoder depends on P_X in general for making $E[P_e]$ small. In this section, we consider coding of a discrete memoryless source and evaluate asymptotic behavior of the decoding error probability under a universal decoder.

To this end, we fix a discrete memoryless source with a finite alphabet $\mathcal{X}$ determined by a probability distribution P_X on $\mathcal{X}$. For a blocklength $n \ge 1$ denote the type of $x^n \in \mathcal{X}^n$ by P_{x^n}. Let $\mathcal{F}$ be an arbitrary family of universal hash functions, where every $f \in \mathcal{F}$ is a mapping from $\mathcal{X}^n$ to $\mathcal{B} \overset{\text{def}}{=} \{1, 2, \ldots, \lceil 2^{nR} \rceil\}$.

Suppose that an encoder and a decoder share an $f \in \mathcal{F}$ chosen randomly subject to the uniform distribution. We use the same encoder given in the preceding section. In this section, however, we use the minimum-entropy decoder [6]. That is, for a transmitted codeword $Y_n \overset{\text{def}}{=} f(X^n)$ the minimum-entropy decoder outputs

$$\hat{X}^n = \arg \min_{x^n \in f^{-1}(Y_n)} H(P_{x^n})$$

where $H(P_{x^n})$ denotes the entropy of the type P_{x^n} of x^n. Ties can be broken arbitrarily.

Denoting by $\mathcal{P}$ the set of all the probability distributions on $\mathcal{X}$, we define

$$G(R, P_X) = \min_{Q_X \in \mathcal{P}} [D(Q_X \,\|\, P_X) + |R - H(Q_X)|^+]$$

where $|t|^+ = \max\{t, 0\}$, $H(Q_X)$ denotes the entropy of $Q_X \in \mathcal{P}$ and $D(Q_X \,\|\, P_X)$ denotes the divergence. The function $G(R, P_X)$ is known as an attainable error exponent of the bin coding [3], [6] and the linear coding [5] for a single source. It is easily verified that $G(R, P_X) > 0$ if $R > H(P_Z)$ and $G(R, P_X) = 0$ if $R \le H(P_X)$.

The following theorem claims that the same attainable error exponent appears in coding of a memoryless source based on an arbitrary family of universal hash functions.

Theorem 2: Let $\mathcal{F}$ be an arbitrary family of universal hash functions. Then, for any $n \ge 1$ it holds that

$$E\left[P_e^{(n)}\right] \le (n+1)^{2|\mathcal{X}|} 2^{-nG(R, P_X)} \tag{23}$$

where $E[\,\cdot\,]$ denotes the expectation with respect to the random choice of $f \in \mathcal{F}$ subject to the uniform distribution.

Before giving the proof of Theorem 2, we briefly review well-known properties of the types. See [4], [7] for more details of arguments using the types. Let $\mathcal{P}_n$ denote the set of all the types of elements in $\mathcal{X}^n$. The cardinality of $\mathcal{P}_n$ is known to satisfy

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}. \tag{24}$$

For any $Q_X \in \mathcal{P}_n$, denote the type class of Q_X by $T(Q_X)$. Then, it is known that $|T(Q_X)|$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(Q_X)} \le |T(Q_X)| \le 2^{nH(Q_X)}. \tag{25}$$

Denote the probability that $X^n \in T(Q_X)$ by $P_{X^n}(T(Q_X))$. Then, we have

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q_X \,\|\, P_X)} \le P_{X^n}(T(Q_X)) \le 2^{-nD(Q_X \,\|\, P_X)}. \tag{26}$$

Now, we are ready to prove Theorem 2.

Proof of Theorem 2: For given $f \in \mathcal{F}$ and $x^n \in \mathcal{X}^n$ define

$$\chi_f(x^n) = \begin{cases} 1, & \text{if } f(\tilde{x}^n) = f(x^n) \text{ for some } \tilde{x}^n \in \mathcal{X}^n \\ & \text{satisfying } \tilde{x}^n \ne x^n \text{ and } H(P_{\tilde{x}^n}) \le H(P_{x^n}) \\ 0, & \text{otherwise} \end{cases}$$

Then, it follows that

$$\begin{aligned} E\left[P_e^{(n)}\right] &\le \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \chi_f(x^n) \\ &= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \frac{|\mathcal{F}(x^n)|}{|\mathcal{F}|} \\ &\le \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \min\left\{\frac{|\mathcal{L}_n(x^n)|}{2^{nR}}, 1\right\} \end{aligned} \tag{27}$$

where

$$\mathcal{F}(x^n) = \{f \in \mathcal{F} : f(\tilde{x}^n) = f(x^n) \text{ for some } \tilde{x}^n \in \mathcal{L}_n(x^n) \text{ and } \tilde{x}^n \ne x^n\}$$

$$\mathcal{L}_n(x^n) = \{\tilde{x}^n \in \mathcal{X}^n : H(P_{\tilde{x}^n}) \le H(P_{x^n})\}$$

and the first and the second inequalities in (27) follow from the definition of the minimum entropy decoder and Lemma 1, respectively. Since a standard argument on the type yields

$$\begin{aligned} |\mathcal{L}_n(x^n)| &= \sum_{Q_X \in \mathcal{P}_n : H(Q_X) \le H(P_{x^n})} |T(Q_X)| \\ &\le (n+1)^{|\mathcal{X}|} 2^{nH(P_{x^n})} \end{aligned} \tag{28}$$

(27) and (28) lead to

$$\begin{aligned} &E\left[P_e^{(n)}\right] \\ &\quad \le \sum_{Q_X \in \mathcal{P}_n} \sum_{x^n \in T(Q_X)} P_{X^n}(x^n) \\ &\qquad \times \min\left[\frac{(n+1)^{|\mathcal{X}|} 2^{nH(P_{x^n})}}{2^{nR}}, (n+1)^{|\mathcal{X}|}\right] \\ &\quad \le (n+1)^{|\mathcal{X}|} \sum_{Q_X \in \mathcal{P}_n} 2^{-nD(Q_X \,\|\, P_X)} 2^{-n|R - H(Q_X)|^+} \\ &\quad \le (n+1)^{2|\mathcal{X}|} 2^{-nG(R, P_X)} \end{aligned} \tag{29}$$

where the last inequality in (29) follows from (26). This establishes (23). □

Next, consider the case where an arbitrarily family of $|\mathcal{X}|^n$-SU hash functions, say $\mathcal{F}_1$ in Example 1, is used as $\mathcal{F}$ in Theorem 2. Recall here that Theorem 1-B) holds for any subset $\mathcal{T}$ for such a class. Obviously, we can obtain

$$E\left[P_e^{(n)}\right] \le \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\left[1-\left(1-\frac{1}{\lceil 2^{nR}\rceil}\right)^{|\mathcal{L}_n(x^n)|}\right] \quad (30)$$

by applying Lemma 2 to the last inequality in (27). Then, a natural question arises whether we can obtain an attainable error exponent that is greater than $G(R, P_X)$ or not. Unfortunately, the answer to this question is negative. Letting F_n be the right-hand side of (30), we can actually prove that

$$\lim_{n\to\infty} -\frac{1}{n}\log F_n = G(R, P_X) \quad (31)$$

which means that no attainable error exponent greater than $G(R, P_X)$ is obtained from this approach. See the Appendix for the proof of (31). By taking Examples 2 and 3 into account, we can intuitively understand the reason why the attainable error exponent of bin coding coincides with the attainable error exponent of linear coding.

V. Performance of Other Decoders for a General Source

In the preceding section we have analyzed the expectation of the decoding error probability of the minimum-entropy decoder for a discrete memoryless source. In this section we consider a wide class of decoders including universal decoders under no assumption on the source.

Hereafter, we use terminologies of information-spectrum methods [10]. Let $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ be a general source with a finite alphabet $\mathcal{X}$. Here, a general source is defined as a sequence of probability distributions on $\mathcal{X}^n$ not required to satisfy the consistency condition. Encoding and decoding are defined for each $n \ge 1$. For defining an encoder and a decoder of blocklength n, we arbitrarily fix a family of universal hash functions $\mathcal{F}$. We assume that an encoder and a decoder share an $f \in \mathcal{F}$ that is chosen randomly subject to the uniform distribution. We consider the same encoder as in Sections III and IV. That is, for an output X^n from a source, the encoder outputs a codeword $Y_n = f(X^n) \in \mathcal{B} \overset{\text{def}}{=} \{1, 2, \ldots, \lceil 2^{nR}\rceil\}$ and transmits Y_n to the decoder, where $R > 0$ is a constant that determines the coding rate.

In this section, we define a decoder of blocklength n by using an arbitrary mapping $\varphi_n : \mathcal{X}^n \to \{0,1\}^*$, where $\{0,1\}^*$ means the set of all the binary sequences of finite length. Denote by $l(\varphi_n(x^n))$ the length of $\varphi_n(x^n)$. We assume that φ_n satisfies the Kraft inequality

$$\sum_{x^n \in \mathcal{X}^n} 2^{-l(\varphi_n(x^n))} \le 1. \quad (32)$$

We consider a decoder that outputs

$$\hat{X}^n = \arg\min_{x^n \in f^{-1}(Y_n)} l(\varphi_n(x^n))$$

where ties can be broken arbitrarily.

The following gives an upper bound of $E[P_e^{(n)}]$ that is dependent on P_{X^n}, ψ_n and R.

Theorem 3: Let X^n be an arbitrary random variable taking values in $\mathcal{X}^n$. Let φ_n be an arbitrary mapping from $\mathcal{X}^n$ to $\{0,1\}^*$ satisfying the Kraft inequality (32). Suppose that $\mathcal{F}$ is an arbitrary family of universal hash functions. If we use the encoder and the decoder described above, then the decoding error probability $P_e^{(n)}$ satisfies

$$E\left[P_e^{(n)}\right] \le \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) 2^{-|nR - l(\varphi_n(x^n))|^+} \quad (33)$$

where $E[\,\cdot\,]$ denotes the expectation with respect to the random choice of $f \in \mathcal{F}$ subject to the uniform distribution.

Proof: Basic ideas for the proof of this theorem have already appeared in the proofs of Theorems 1 and 2. In fact, similarly to (27) we can obtain

$$E\left[P_e^{(n)}\right] \le \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \min\left\{\frac{|\mathcal{L}_{\varphi_n}(x^n)|}{2^{nR}}, 1\right\} \quad (34)$$

where

$$\mathcal{L}_{\varphi_n}(x^n) = \{\tilde{x}^n \in \mathcal{X}^n : l(\varphi_n(\tilde{x}^n)) \le l(\varphi_n(x^n))\}.$$

Notice here that we have

$$|\mathcal{L}_{\varphi_n}(x^n)| \le 2^{l(\varphi_n(x^n))}, \quad \text{for all } x^n \in \mathcal{X}^n \quad (35)$$

owing to the assumption that φ_n satisfies the Kraft inequality (32). In fact, it is easily checked that

$$\begin{aligned} 1 &\ge \sum_{\tilde{x}^n \in \mathcal{X}^n} 2^{-l(\varphi_n(\tilde{x}^n))} \\ &\ge \sum_{\tilde{x}^n \in \mathcal{L}_{\varphi_n}(x^n)} 2^{-l(\varphi_n(\tilde{x}^n))} \\ &\ge \sum_{\tilde{x}^n \in \mathcal{L}_{\varphi_n}(x^n)} 2^{-l(\varphi_n(x^n))} \\ &\ge |\mathcal{L}_{\varphi_n}(x^n)|\, 2^{-l(\varphi_n(x^n))} \end{aligned}$$

for every $x^n \in \mathcal{X}^n$, which yields (35). Thus, the combination of (34) and (35) establishes the claim of this theorem. □

Next, we explore a sufficient condition under which $E[P_e^{(n)}]$ in Theorem 3 converges to zero as $n \to \infty$. To this end, for a sequence of mapping $\boldsymbol{\varphi} \overset{\text{def}}{=} \{\varphi_n\}_{n=1}^{\infty}$ we define

$$\overline{L}(\boldsymbol{X}) = \text{p-}\limsup_{n\to\infty} \frac{1}{n} l(\varphi_n(X^n))$$

[11]. Then, Theorem 3 yields the following corollary.

Corollary 2: If $R > \overline{L}(\boldsymbol{X})$, then $E[P_e^{(n)}]$ converges to zero as $n \to \infty$.

Proof: By assumption, there exists a $\gamma_0 > 0$ satisfying $R \ge \overline{L}(\boldsymbol{X}) + 2\gamma_0$. Define

$$\mathcal{V}_n = \left\{x^n \in \mathcal{X}^n : \frac{1}{n} l(\varphi_n(x^n)) \le R - \gamma_0\right\}.$$

Then, in view of the definitions of $\overline{L}(\boldsymbol{X})$ and γ_0 it holds that

$$\Pr\{X^n \in \mathcal{V}_n\} \ge \Pr\left\{\frac{1}{n} l(\varphi_n(X^n)) \le \overline{L}(\boldsymbol{X}) + \gamma_0\right\} \to 1, \quad \text{as } n \to \infty$$

which implies that

$$\lim_{n\to\infty} \Pr\{X^n \notin \mathcal{V}_n\} = 0. \quad (36)$$

By using Theorem 3 we can evaluate $E[P_e^{(n)}]$ in the following manner:

$$\begin{aligned} E\left[P_e^{(n)}\right] &\le \sum_{x^n\in\mathcal{V}_n} P_{X^n}(x^n)2^{-|nR-l(\varphi_n(x^n))|^+} \\ &\quad + \sum_{x^n\notin\mathcal{V}_n} P_{X^n}(x^n)2^{-|nR-l(\varphi_n(x^n))|^+} \\ &\le 2^{-n\gamma_0}\Pr\{X^n\in\mathcal{V}_n\}+\Pr\{X^n\notin\mathcal{V}_n\} \end{aligned} \tag{37}$$

where the second inequality follows because $nR - l(\varphi_n(x^n)) \ge n\gamma_0 > 0$ for all $x^n \in \mathcal{V}_n$ and $|nR - l(\varphi_n(x^n))|^+ \ge 0$ for all $x^n \notin \mathcal{V}_n$. Then, the claim of this corollary is immediate from (36) and (37). □

It is shown in [11] that if $\boldsymbol{\varphi} = \{\varphi_n\}_{n=1}^\infty$ is mean-optimal, i.e., $\boldsymbol{\varphi}$ satisfies

$$\lim_{n\to\infty}\left[\frac{1}{n}E[l(\varphi_n(X^n))]-\frac{1}{n}H(X^n)\right]=0 \tag{38}$$

then we have $\overline{L}(\boldsymbol{X})=\overline{H}(\boldsymbol{X})$, where $\overline{H}(\boldsymbol{X})$ is defined in (22). Note that the assumption on the uniform integrability of $\{\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)}\}_{n=1}^\infty$ is satisfied because $\mathcal{X}$ is a finite alphabet. There are several examples of φ with the mean-optimality. The Shannon-Fano code (e.g., [4]) satisfies (38) even if $\boldsymbol{X}$ is a general source. For the case that $\boldsymbol{X}$ is stationary and ergodic, (38) is satisfied by a universal code such as the LZ78 code [19]. If $\boldsymbol{X}$ is stationary and memoryless, we can use the Lynch-Davisson code [9], [13] as φ_n. In this case, we can obtain a result similar to Theorem 2 because $\frac{1}{n}l(\varphi_n(x^n)) \le H(Q_X)+\frac{|\mathcal{X}|}{n}\log(n+1)+\frac{2}{n}$ for all $x^n \in T(Q_X)$ and $Q_X \in \mathcal{P}_n$.

VI. Conclusion

The objective of this correspondence is the investigation of new connections between source coding and two kinds of families of hash functions known as the families of universal and N-strongly universal hash functions. We have given a coding scheme using one of the two families and have obtained an upper bound on the expectation of the decoding error probability for each family. In particular, for the case of discrete memoryless sources, we have developed an attainable error exponent under the minimum-entropy decoder that coincides with the attainable error exponent by linear coding. We have also discussed coding of sources without the memoryless assumption. A sufficient condition under which the expectation of decoding error probability goes to zero for a decoder in a certain class.

Appendix
Proof of (31)

Proof: We prove (31) by establishing both

$$\liminf_{n\to\infty} -\frac{1}{n}\log F_n \ge G(R,P_X) \tag{39}$$

$$\limsup_{n\to\infty} -\frac{1}{n}\log F_n \le G(R,P_X). \tag{40}$$

Inequality (39) follows immediately. In fact, owing to the inequality $1-(1-u)^m \le mu$ for any integer $m \ge 1$ and $u \in (0,1)$, the right-hand side of (30) is upper bounded by the right-hand side of (27). Hence, Theorem 2 guarantees that $F_n \le (n+1)^{2|\mathcal{X}|}2^{-nG(R,P_X)}$, which implies (39).

We need preparation for establishing (40). Let Q_X^* be the probability distribution on $\mathcal{X}$ that attains the minimum of $G(R,P_X)$. Notice that we can choose a sequence $\{Q_X^{(n)}\}_{n=1}^\infty$ of probability distributions on $\mathcal{X}$ satisfying $Q_X^{(n)} \in \mathcal{P}_n$ for all $n \ge 1$ and $Q_X^{(n)} \to Q_X^*$ as $n \to \infty$. In particular, for each $n \ge 1$ we can choose $Q_X^{(n)}$ satisfying $|Q_X^{(n)}(x) - Q_X^*(x)| \le \frac{1}{n}$ for all $x \in \mathcal{X}$. Then, it holds that $|H(Q_X^{(n)})-H(Q_X^*)| = O(\frac{1}{n})$ under such a choice of $\{Q_X^{(n)}\}_{n=1}^\infty$.

Now, we establish a lower bound of F_n in (31). Since F_n can be written as

$$F_n = \sum_{Q_X\in\mathcal{P}_n}\sum_{x^n\in T(Q_X)} P_{X^n}(x^n)\left[1-\left(1-\frac{1}{\lceil 2^{nR}\rceil}\right)^{|\mathcal{L}_n(x^n)|}\right]$$

and every term in the above summations is nonnegative, we have

$$F_n \ge \sum_{x^n\in T(Q_X^{(n)})} P_{X^n}(x^n)\left[1-\left(1-\frac{1}{\lceil 2^{nR}\rceil}\right)^{|\mathcal{L}_n(x^n)|}\right]. \tag{41}$$

Notice that a standard argument using the types yields

$$\begin{aligned} |\mathcal{L}_n(x^n)| &\ge \left|T(Q_X^{(n)})\right| \\ &\ge \frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(Q_X^{(n)})}, \quad \text{for all } x^n \in T(Q_X^{(n)}) \end{aligned} \tag{42}$$

where the second inequality follows from (25). Then by combining (41), (42), and (26) we have the following lower bound of F_n:

$$\begin{aligned} F_n \ge{}& \frac{1}{(n+1)^{|\mathcal{X}|}}2^{-nD(Q_X^{(n)}\,\|\,P_X)} \\ &\times\left[1-\left(1-\frac{1}{2^{nR+1}}\right)^{\frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(Q_X^{(n)})}}\right]. \end{aligned} \tag{43}$$

Hereafter, we consider the two cases A) $R \ge H(Q_X^*)$ and B) $R < H(Q_X^*)$. Consider case A) first. By using the inequality $(1-u)^m \le \exp[-mu]$ for all $m \ge 0$ and $0 \le u \le 1$, we have

$$\begin{aligned} 1&-\left(1-\frac{1}{2^{nR+1}}\right)^{\frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(Q_X^{(n)})}} \\ &\ge 1-\exp\left[-\frac{1}{(n+1)^{|\mathcal{X}|}}2^{-n(R-H(Q_X^{(n)}))-1}\right] \\ &\ge \frac{1}{(n+1)^{|\mathcal{X}|}}2^{-n(R-H(Q_X^{(n)}))-2} \end{aligned} \tag{44}$$

for all sufficiently large n, where the second inequality follows from the inequality $\exp[-x] \le 1-\frac{x}{2}$ for all $0 \le x \le \ln 2$. Here, notice that $\frac{1}{(n+1)^{|\mathcal{X}|}}2^{-n(R-H(Q_X^{(n)}))-1}$ becomes arbitrarily close to zero even for the case of $R = H(Q_X^*)$ due to the choice of $Q_X^{(n)}$, $n \ge 1$. Therefore, the combination of (43) and (44) yields

$$F_n \ge \frac{1}{(n+1)^{2|\mathcal{X}|}}2^{-n[D(Q_X^{(n)}\,\|\,P_X)+R-H(Q_X^{(n)})]-2}$$

which, together with the continuity of $D(Q_X\,\|\,P_X)+R-H(Q_X)$ with respect to Q_X, establishes (40) for case A).

Next, we consider case B). In this case, since

$$2^{nR+1} \le \frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(Q_X^{(n)})}$$

for all sufficiently large n, it holds that

$$\begin{aligned} 1&-\left(1-\frac{1}{2^{nR+1}}\right)^{\frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(Q_X^{(n)})}} \\ &\ge 1-\left(1-\frac{1}{2^{nR+1}}\right)^{2^{nR+1}} \\ &\to 1-\frac{1}{e}, \qquad \text{as } n\to\infty. \end{aligned}$$

Hence, there exists a constant $C > 0$ such that

$$F_n \geq \frac{C}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q_X^{(n)} \,\|\, P_X)}, \quad \text{for all sufficiently large } n$$

which, together with the continuity of $D(Q_X \,\|\, P_X)$ with respect to Q_X, establishes (40) for case (B). □

Acknowledgment

The author is grateful to the Associate Editor and an anonymous reviewer for their helpful comments. In particular, the Associate Editor's comment on (n, k) linear codes was significant in revision of this correspondence. The author wishes to thank I. Nakano for discussions on the proof of Lemma 2.

References

[1] C. H. Benett, G. Brassard, C. Crépeau, and U. M. Maurer, "Generalized privacy amplification," *IEEE Trans. Inf. Theory*, vol. 41, pp. 1915–1923, 1995.

[2] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *J. Comput. Syst. Sci.*, vol. 18, pp. 143–154, 1979.

[3] T. Cover, "A proof of data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 226–228, 1975.

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

[5] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 585–592, 1982.

[6] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. IT-26, pp. 155–165, 1980.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Source*. New York: Academic, 1981.

[8] H. Krawczyk, "LFSR-based hashing and authentication," in *Proc. Adv. Cryptol.—CRYPTO'95 (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 1994, vol. 963, pp. 129–139.

[9] L. D. Davisson, "Comments on 'sequence time coding for data compression,'" *Proc. IEEE*, vol. 54, p. 2010, 1966.

[10] T. S. Han, *Information-Spectrum Methods in Information Theory*. New York: Springer-Verlag, 2003.

[11] H. Koga and H. Yamamoto, "Asymptotic properties on codeword lengths of an optimal FV code for general sources," *IEEE Trans. Inf. Theory*, vol. 51, pp. 1546–1555, 2005.

[12] K. Kurosawa and T. Yoshida, "Strongly universal hashing and identification codes via channels," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2091–2095, 1999.

[13] T. J. Lynch, "Sequence time coding for data compression," *Proc. IEEE*, vol. 54, pp. 1490–1491, 1966.

[14] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[15] U. Maurer and S. Wolf, "Information-theoretic key agreement: From weak to strong secrecy for free," in *Adv. Cryptol.—EUROCRYPT'00 (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 2000, vol. 1807, pp. 351–368.

[16] J. Muramatsu, "Source coding algorithms using the randomness of a past sequence," *IEICE Trans. Fund.*, vol. E88-A, pp. 1063–1083, 2005.

[17] D. R. Stinson, "Universal hashing and authentication codes," *Des., Codes Cryptogr.*, vol. 4, pp. 369–380, 1994.

[18] M. N. Wegman and J. L. Carter, "New hash functions and their use in authentication and set equality," *J. Comput. Syst. Sci.*, vol. 22, pp. 265–279, 1981.

[19] J. Ziv and A. Lempel, "Compression of individual sequence via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, pp. 530–536, 1978.

On Defining Partition Entropy by Inequalities

Ping Luo, Guoxing Zhan, Qing He,
Zhongzhi Shi, *Senior Member, IEEE*, and Kevin Lü

***Abstract*—*Partition entropy* is the numerical metric of uncertainty within a partition of a finite set, while *conditional entropy* measures the degree of difficulty in predicting a decision partition when a condition partition is provided. Since two direct methods exist for defining conditional entropy based on its partition entropy, the inequality postulates of monotonicity, which conditional entropy satisfies, are actually additional constraints on its entropy. Thus, in this paper partition entropy is defined as a function of probability distribution, satisfying all the inequalities of not only partition entropy itself but also its conditional counterpart. These inequality postulates formalize the intuitive understandings of uncertainty contained in partitions of finite sets. We study the relationships between these inequalities, and reduce the redundancies among them. According to two different definitions of conditional entropy from its partition entropy, the convenient and unified checking conditions for any partition entropy are presented, respectively. These properties generalize and illuminate the common nature of all partition entropies.**

***Index Terms*—Conditional entropy, inequality, partition entropy, uncertainty.**

I. Introduction

Learning is an important cognitive process that allows the making of correct decisions and improves performance. From an information theory point of view, learning can be seen as a reduction of uncertainty and the amount by which the uncertainty is reduced can be an indicator of the speed of learning [1]. Thus, *partition entropy* [2], measuring uncertainty and impurity in a given *partition* of a finite set, is an important concept in cognitive and computer science.

Conditional Entropy [2], defined based on its partition entropy, is another significant concept. It describes the degree of difficulty in predicting a *decision partition* by a *condition partition*. It is also the measure of uncertainty left in a decision partition after a condition partition is provided. This concept is widely used in the field of Machine Learning, as heuristics to guide the greedy search for suboptimal solutions. For example, [3, Algorithm C4.5], which is a popular algorithm for building a decision tree, uses the Shannon conditional entropy as a metric to select the local "optimal" attribute to branch. In the algorithm for attribute reduction of information view [4], the Shannon conditional entropy is also selected as the measure of attribute importance for decision predicting. In these algorithms, the one with the minimal conditional entropy among all available options is chosen to continue the following steps. Thus, only the *relative* magnitude of entropies for

Manuscript received November 23, 2005; revised December 18, 2006. This work is supported by the National Science Foundation of China (No. 60435010, 90604017, 60675010), the 863 Project (No. 2006AA01Z128), National Basic Research Priorities Programme (No. 2003CB317004), and the Nature Science Foundation of Beijing (No. 4052025).

P. Luo is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China (e-mail: ping.luo@gmail.com). He is also with the Graduate University of the Chinese Academy of Sciences.

G. Zhan is with the Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China. He is also with the Graduate University of the Chinese Academy of Sciences.

Q. He and Z. Shi are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China.

K. Lü is with the Brunel University, Uxbridge UB8 3PH, U.K.

Communicated by V. A. Vaishampayan, Associate Editor for At Large.

Digital Object Identifier 10.1109/TIT.2007.903124

0018-9448/$25.00 © 2007 IEEE