Generalized Universal Coding of Integers

Wei Yan, *Member, IEEE* and Yunghsiang S. Han, *Fellow, IEEE*

***Abstract*—Universal coding of integers (UCI) is a class of variable-length code such that the ratio of the expected codeword length to $\max\{1, H(P)\}$ is bounded by a constant factor, where $H(P)$ is the Shannon entropy of the decreasing probability distribution P. However, if we consider the ratio of the expected codeword length to $H(P)$ for UCI, the ratio tends to infinity when $H(P)$ tends to zero. To resolve this issue, we introduce a class of codes, called generalized universal coding of integers (GUCI), where the ratio of the expected codeword length to $H(P)$ is bounded by a constant factor K.**

First, the definition of GUCI is proposed. The coding structure of GUCI is introduced. Next, we propose a class of GUCIs $\mathcal{C}$ to reach the expansion factor $K_{\mathcal{C}} = 2$, and we show that the smallest minimum expansion factor is in the range $1 \le K^* \le 2$. Then, by comparing UCI and GUCI, we show that when the entropy is very large or $P(0)$ is not large, there are also cases where the average codeword length of GUCI is shorter. Finally, the asymptotically optimal GUCI is presented.

***Index Terms*—Universal coding of integers, Source coding, and Elias coding.**

I. Introduction

There are three major categories of lossless source coding: variable-to-fixed length (VF) codes (e.g., Tunstall code [2]), fixed-to-variable length (FV) codes (e.g., Huffman code [3]), and variable-to-variable length (VV) codes (e.g., Khodak code [4, 5]). As their name implies, VF codes encode a variable-length sequence of source symbols into a constant-length codeword. FV codes encode a constant-length sequence of source symbols into a variable-length codeword. In particular, *variable-length codes* map the source symbols to a variable number of bits, which is the most important type of FV code. VF and FV codes are special cases of VV codes; the main study of VV codes has been focused on redundancy rates [4–7].

Universal coding of integers (UCI) is a variable-length code (i.e., a type of FV code) for discrete memoryless sources with infinite alphabets, and the probability distribution of sources does not require prior knowledge. In 1968, Levenshtein [8] proposed the first UCI, although UCI was not yet defined. In 1975, Elias [9] established the fundamental framework of UCI. Elias considered discrete memoryless sources $S = (P, \mathcal{A})$ with a countable alphabet set $\mathcal{A} = \mathbb{N}^+ = \{1, 2, 3, \cdots\}$ and a decreasing probability distribution (DPD) P of $\mathbb{N}^+$ (i.e., $\sum_{n=1}^{\infty} P(n) = 1$, and $P(m) \ge P(m+1) \ge 0$, for all $m \in \mathbb{N}^+$). Let $H(P) = -\sum_{n=1}^{\infty} P(n) \log_2 P(n)$ denote the Shannon entropy of P. Let $\mathcal{C}$ be a variable-length code for the source $S = (P, \mathbb{N}^+)$; it maps the positive integers $\mathbb{N}^+$ onto binary codewords $\{0,1\}^*$. Let $L_{\mathcal{C}}(\cdot)$ denote the length function such that $L_{\mathcal{C}}(m) = |\mathcal{C}(m)|$, for all $m \in \mathbb{N}^+$, where $\mathcal{C}(m)$ is the corresponding codeword of m. Furthermore, $E_P(L_{\mathcal{C}}) = \sum_{n=1}^{\infty} P(n) L_{\mathcal{C}}(n)$ denotes the expected codeword length of $\mathcal{C}$. We say that $\mathcal{C}$ is *universal* if

$$\frac{E_P(L_{\mathcal{C}})}{\max\{1, H(P)\}} \le K_{\mathcal{C}}, \tag{1}$$

for all DPDs P with $H(P) < \infty$. $K_{\mathcal{C}}$ is called the *expansion factor* of UCI $\mathcal{C}$, and $K_{\mathcal{C}}^* \triangleq \inf\{K_{\mathcal{C}} \,|\, \forall \text{ DPD } P \text{ and } H(P) < \infty\}$ is called the *minimum expansion factor* of UCI $\mathcal{C}$. Moreover, $\mathcal{C}$ is called *asymptotically optimal* if $\mathcal{C}$ is universal and there exists a function $R_{\mathcal{C}}(\cdot)$ such that

$$\frac{E_P(L_{\mathcal{C}})}{\max\{1, H(P)\}} \le R_{\mathcal{C}}(H(P)), \tag{2}$$

for all DPDs P with $H(P) < \infty$ and

$$\lim_{H(P) \to +\infty} R_{\mathcal{C}}(H(P)) = 1.$$

UCI has two main categories [10], namely, the *message length strategy* and the *flag pattern strategy*. The γ, δ and ω codes proposed by Elias [9] are associated with the message length strategy. With this strategy, UCI is the main recursive code used to minimize $L_{\mathcal{C}}(m)$ for large $m \in \mathbb{N}$. For example, two classes of UCIs were proposed by Stout [11] to improve ω code for large m. Furthermore, Yamamoto [12] cleverly designed a delimiter with a length greater than 1 to construct a new class of UCIs whose length function satisfies

$$L_{\mathcal{C}}(m) < \log_2 m + \log_2(\log_2 m) + \cdots + \log_2^{t^*(m)} m,$$

where $t^*(m)$ is the largest positive integer t satisfying $\log_2^t m \ge 0$. However, the UCI with the message length strategy should be used in an error-free environment. Instead, the UCI with the flag pattern strategy, first studied by Lakshmanan [13], compensates for this problem, and it has certain resynchronization properties. The family of Fibonacci codes [14] is likely the most famous flag pattern strategy for UCI. However, this approach is not asymptotically optimal and involves complicated encoding and decoding. A UCI with a new flag pattern strategy, proposed by Wang [15], has been improved in the above two aspects. Yamamoto *et al.* [16] further improved upon Wang's coding scheme. Furthermore, Amemiya *et al.* [17] provided a new *group strategy* UCI such that the message length strategy coding can be regarded as a special group strategy coding.

This work was supported by the National Key Research and Development Program of China under Grant 2022YFA1004902.(Corresponding author: Wei Yan)

This work was partially presented at the 2021 IEEE Information Theory Workshop [1].

Wei Yan is with the National University of Defense Technology, Hefei 230037, China, and also with Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, Hefei 230037, China (email: yan.wei2023@nudt.edu.cn).

Yunghsiang S. Han is with Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518000, China (email: yunghsiangh@gmail.com).

Recently, Ávila *et al.* [18] proposed a new family of UCIs whose length function can reach the bounds in [8, 19, 20] with a constant difference. Allison *et al.* [21] focused on the universality of the Wallace tree code. Yan *et al.* [22, 23] first studied the range of the minimum expansion factor of UCIs. If a class of UCIs $\mathcal{C}$ has the smallest minimum expansion factor $K_{\mathcal{C}}^*$, then $\mathcal{C}$ is called *the optimal UCI*.[1] The authors proved that the optimal UCI is in the range $2 \le K_{\mathcal{C}}^* \le 2.5$, where $K_{\mathcal{C}}^* = 2.5$ is achieved by ι code [23]. Today, UCI is used in many applications, such as biological sequencing data compression [24, 25], inverted file index [26], H.264 advanced video compression standard [27], H.265 high efficiency video coding [28],[2] standard MIDI (Musical Instrument Digital Interface) file format [30], evolving secret sharing [31], and unbounded search problems [19, 32].

A. Motivations

1) For a UCI, the ratio of $E_P(L_{\mathcal{C}})$ to $\max\{1, H(P)\}$ is bounded by a constant. However, when $H(P)$ is extremely small, the expected codeword length of a variable-length code is

$$E_P(L_{\mathcal{C}}) = \sum_{n=1}^{\infty} P(n) L_{\mathcal{C}}(n) \ge \sum_{n=1}^{\infty} P(n) \cdot 1 = 1. \quad (3)$$

Therefore, for a UCI, the ratio of $E_P(L_{\mathcal{C}})$ to $H(P)$ cannot be bounded by a constant $K_{\mathcal{C}}$ when $H(P)$ approaches zero. That is, the UCI cannot satisfy the following inequality:

$$\frac{E_P(L_{\mathcal{C}})}{H(P)} \le K_{\mathcal{C}}. \quad (4)$$

We found that the reason the UCI cannot satisfy (4) is that $E_P(L_{\mathcal{C}})$ of variable-length codes have a lower bound (3).

Next, we show that neither FV codes nor VF codes can satisfy the inequality similar to (4). Let $\mathcal{C}_n$ be an FV code for the discrete memoryless source $S = (P, \mathcal{A})$, and it maps $\mathcal{A}^n$ onto binary codewords $\{0,1\}^*$, where n is a fixed positive integer. Let P^n be the probability distribution over $\mathcal{A}^n$. Due to

$$E_{P^n}(L_{\mathcal{C}_n}) \triangleq \sum_{x \in \mathcal{A}^n} P^n(x) L_{\mathcal{C}_n}(x) \ge \sum_{x \in \mathcal{A}^n} P^n(x) = 1,$$

then $\mathcal{C}_n$ cannot satisfy

$$\frac{E_{P^n}(L_{\mathcal{C}_n})}{n} \Big/ H(P) \le K_{\mathcal{C}}$$

when $H(P)$ approaches zero. Since the size of the alphabet $\mathcal{A}$ is infinite, there can be no bijection mapping from an infinite subset $\mathcal{D}$ of $\mathcal{A}^*$ to $\{0,1\}^n$; i.e., no such FV code exists.

Therefore, we introduce VV codes in this universal coding problem to address this issue. First, by introducing VV codes, it is possible to construct a new class of code that satisfies the inequality similar to (4). Second, due to

$$\frac{E_P(L_{\mathcal{C}})}{\max\{1, H(P)\}} \le \frac{E_P(L_{\mathcal{C}})}{H(P)} \le K_{\mathcal{C}},$$

the VV code $\mathcal{C}$ is still a UCI as long as an inequality similar to (4) is satisfied. That is, (4) is a stronger requirement than (1) is.

In short, the aim is to propose a new class of code satisfying the inequality similar to (4).

2) When the study of the problem is extended from variable-length codes to VV codes, the mapping of an individual input symbol to $\{0,1\}^*$ we consider is changed to the mapping of sequences of input symbols to $\{0,1\}^*$. Therefore, we next introduce the related concept of universal source coding (USC) [33, 34].

USC considers the stationary and memoryless source $S = (P, \mathcal{A})$, where $\mathcal{A}$ is the alphabet and P is an unknown probability distribution on $\mathcal{A}$. Let $\{\mathcal{C}_n\}_{n=1}^{\infty}$ be a sequence of prefix-free FV codes, where $\mathcal{C}_n$ is a mapping from $\mathcal{A}^n$ to $\{0,1\}^*$. Let $\mathcal{P}(\mathcal{A})$ denote the set of all probability distributions on $\mathcal{A}$, and $\mathcal{P}_H(\mathcal{A}) \triangleq \{P \in \mathcal{P}(\mathcal{A}) \mid H(P) < \infty\}$. Let $D(P \parallel Q)$ denote the relative entropy between two probability distributions P and Q. Thus, we obtain

$$\begin{aligned}\frac{D(P^n \parallel Q)}{n} &\triangleq \frac{1}{n} \sum_{x \in \mathcal{A}^n} P^n(x) \log_2 \frac{P^n(x)}{Q(x)} \\ &= -\frac{1}{n} \sum_{x \in \mathcal{A}^n} P^n(x) \log_2 Q(x) - H(P) \\ &= \frac{1}{n} \sum_{x \in \mathcal{A}^n} P^n(x) L_{\mathcal{C}_n}(x) - H(P) \\ &= \frac{E_{P^n}(L_{\mathcal{C}_n})}{n} - H(P),\end{aligned}$$

where $L_{\mathcal{C}_n}(x) = -\log_2 Q(x)$ for all $x \in \mathcal{A}^n$.[3] A class $\Upsilon \subseteq \mathcal{P}(\mathcal{A})$ is called *weakly universal* if there exists a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that

$$\sup_{P \in \Upsilon} \lim_{n \to \infty} \left(\frac{E_{P^n}(L_{\mathcal{C}_n})}{n} - H(P) \right) = 0.$$

And is called *strongly universal* if there exists a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that

$$\lim_{n \to \infty} \sup_{P \in \Upsilon} \left(\frac{E_{P^n}(L_{\mathcal{C}_n})}{n} - H(P) \right) = 0.$$

When $\Upsilon = \mathcal{P}_H(\mathcal{A})$ and $\mathcal{A} = \mathbb{N}^+$, there is no weak USC [34, 36].

Therefore, there are two subsequent research directions for [36], as follows. One is to consider the strong universality of the proposed method over sub-collections of distributions [37], and the other is to

[1] The optimal UCI may not exist, but there exists a family of UCIs such that the minimum expansion factor tends to a minimum value. In this paper, we approximate the limit of a family of UCIs to UCIs. We follow the previous research and assume that the optimal UCI exists.

[2] There is a class of UCIs, termed Exp-Golomb codes, which is used in H.264 and H.265. In addition, arithmetic codes [29] are also used in H.264 and H.265.

[3] In [35, P429], $Q(x) = 2^{-L_{\mathcal{C}_n}(x)}$ is the distribution that corresponds to the codeword lengths $L_{\mathcal{C}_n}(x)$. Let us note that to make Q a probability distribution, $\mathcal{C}_n$ ensures that the Kraft inequality is equal.

consider the almost lossless (lossy) source coding over $\Upsilon = \mathcal{P}_H(\mathcal{A})$ [38].

In a broader sense, this work can also be considered a follow-up study for [36]. The sources we consider are arbitrary probability distributions[4] over the infinitely countable alphabet $\mathcal{A}$; the encoding we consider is lossless. We define the parameter $K_{\mathcal{C}}^*$ in Definition 3 as a metric for measuring the universality of the code.

B. Contributions

In this paper, we introduce a family of codes called generalized universal coding of integers (GUCI), which is a generalization of UCI via VV codes. The position of GUCI in VV codes is equivalent to that of UCI in FV codes. GUCI satisfies an inequality similar to (4) for any discrete memoryless source. In particular, GUCI is suitable for small entropy. For example, the frequency domain coefficients have many zeros in image compression after the quantization process [39]. The minimum expansion factor for GUCI is also studied. The major contributions of this paper are as follows.

1) The definitions of GUCI and asymptotically optimal GUCI are presented.
2) A family of GUCIs and asymptotically optimal GUICs are proposed.
3) In the proposed family of GUCIs, a class of GUCIs is proposed to achieve the expansion factor 2. Then, we show that the smallest minimum expansion factor is in the range $1 \leq K^* \leq 2$.
4) The relationship between UCI and GUCI is presented.
5) A sufficient condition for the average codeword length of GUCI to be shorter than that of UCI is obtained. In addition, when the Shannon entropy $H(P)$ is large or $P(0)$ is not large, there are some cases in which the average codeword length of GUCIs is shorter.

C. Organization of this paper

In the remainder of this paper, Section II provides some background knowledge. In Section III, we define GUCI. A family of GUCIs is provided in Section IV. In Section V, the expansion factor of GUCIs is discussed. In Section VI, we compare the average code length of this family of GUCIs and the original UCI. In Section VII, we study the definition and properties of the asymptotically optimal GUCI. Section VIII concludes this work.

II. Preliminaries

A. Notations

Let $\mathbb{N} \triangleq \{0\} \bigcup \mathbb{N}^+$ denote the set of nonnegative numbers. Let $\alpha(m)$ denote the unary representation of the positive number m. For example, $\alpha(1) = 1$, $\alpha(2) = 01$ and $\alpha(5) = 00001$. Let $\beta(m)$ denote the standard binary representation of a positive integer m and $[\beta(m)]$ denote the binary code by removing the most significant bit 1 of $\beta(m)$.[5] For example, $\beta(9) = 1001$ and $[\beta(9)] = 001$. Then, we obtain

$$\begin{aligned} |\alpha(m)| &= m, \\ |\beta(m)| &= 1 + \lfloor \log_2 m \rfloor, \\ |[\beta(m)]| &= \lfloor \log_2 m \rfloor, \end{aligned}$$

for all $m \in \mathbb{N}^+$, where $|a|$ denotes the length of string a.

B. Elias γ code and the codeword lengths of several classical UCIs

Now, we introduce the specific structure of the Elias γ code. For other classic UCIs, please refer to [9, 22, 23]. The Elias γ code was introduced by Elias [9]. It is an encoding scheme for message length. Elias γ code: $\mathbb{N}^+ \to \{0,1\}^*$ can be expressed as

$$\gamma(m) = \alpha(|\beta(m)|)[\beta(m)],$$

for all $m \in \mathbb{N}^+$. The role of the leading $0'$s is to ensure that the Elias γ code is a prefix code. The codeword length is given by

$$\begin{aligned} |\gamma(m)| &= |\alpha(1 + \lfloor \log_2 m \rfloor)| + |[\beta(m)]| \\ &= 1 + \lfloor \log_2 m \rfloor + \lfloor \log_2 m \rfloor \\ &= 1 + 2\lfloor \log_2 m \rfloor. \end{aligned}$$

For example, $\gamma(9) = 0001001$ and $|\gamma(9)| = 1+2\lfloor \log_2 9 \rfloor = 7$. The Elias γ code is universal; however, it is not asymptotically optimal. Next, we provide a lemma about the length function of other classic UCIs.

Lemma 1 ([9, 22, 23])**.** *The following classic UCIs satisfy $L_{\mathcal{C}}(1) = 1$. For all $2 \leq n \in \mathbb{N}^+$,*

1) *the length function of the δ code satisfies $L_\delta(n) = 1 + \lfloor \log_2 n \rfloor + 2\lfloor \log_2(1 + \lfloor \log_2 n \rfloor) \rfloor$;*
2) *the length function of the η code satisfies $L_\eta(n) = 3 + \lfloor \log_2(n-1) \rfloor + \lfloor \frac{\lfloor \log_2(n-1) \rfloor}{2} \rfloor$;*
3) *the length function of the θ code satisfies $L_\theta(n) = 3 + \lfloor \log_2 n \rfloor + \lfloor \log_2 \lfloor \log_2 n \rfloor \rfloor + \lfloor \frac{\lfloor \log_2 \lfloor \log_2 n \rfloor \rfloor}{2} \rfloor$;*
4) *the length function of the ι code satisfies $L_\iota(n) = 2 + \lfloor \log_2 n \rfloor + \lfloor \frac{1 + \lfloor \log_2 n \rfloor}{2} \rfloor$;*
5) *the length function of the ω code satisfies $L_\omega(n) = 1 + \sum_{m=1}^{t}(1 + \lambda^m(n))$, where $\lambda(n) \triangleq \lfloor \log_2 n \rfloor$ and λ^m denotes the m-fold composition of λ, and $t = t(n) \in \mathbb{N}^+$ is a unique integer satisfying $\lambda^t(n) = 1$. Furthermore, $L_\omega(n) \leq 3 + 2\lfloor \log_2 n \rfloor$.*

C. Run-length encoding

Run-length encoding (RLE) [40] is essentially a method of encoding run-length rather than encoding individual values. For example, a scan line consisting of black pixels B and white pixels W may read as follows:

$$WWWWWWWBBBWWWWBWWWWWW$$
$$WWWWWWWWBBWWWWWWWWWW.$$

With the RLE algorithm, it is encoded as

$$7W3B4W1B13W2B10W.$$

[4] Each probability distribution can be turned into a DPD by adjusting the order.

[5] When $m = 1$, $[\beta(m)]$ is a null string.

Moreover, the RLE can be modified to accommodate data properties. For instance, the above scan line can also be encoded as

$$(W, 7, 3, 4, 1, 13, 2, 10),$$

where a prefix code can encode the numbers.

D. Variable-to-fixed length codes

The VF codes can be divided into two parts, called the parser and the string encoder. First, the parser partitions the source sequence into a concatenation of variable-length strings. Each variable-length string belongs to a dictionary $\mathcal{D}$ containing a set of strings. Next, the string encoder maps the variable-length string $\alpha \in \mathcal{D}$ into the fixed-length string. To ensure the completeness and the uniqueness of the segmentation of the source sequence, $\mathcal{D}$ must be proper and complete.

Definition 1 ([41]).

1) *If every variable-length string $\alpha_i \in \mathcal{D}$ is not a prefix of another variable-length string $\alpha_j \in \mathcal{D}$, then $\mathcal{D}$ is termed proper.*
2) *If every infinite sequence has a prefix in $\mathcal{D}$, then $\mathcal{D}$ is termed complete.*

For example, the dictionary $\mathcal{D} = \{1, 00, 01\}$ over $\{0, 1\}$ is clearly proper. If the first element of the infinite sequence is 1, then $1 \in \mathcal{D}$ is its prefix; if the first element of the infinite sequence is 0, then $01 \in \mathcal{D}$ or $00 \in \mathcal{D}$ is its prefix. Therefore, the dictionary $\mathcal{D}$ is complete.

E. Variable-to-variable length codes

VV codes can be considered a concatenation of VF codes and FV codes [5–7]. First, the VF encoder maps the variable-length string $\alpha \in \mathcal{D}$ into the fixed-length string. Then, the FV encoder maps the fixed-length string into the variable-length string. Nishiara *et al.* [42] defined the almost surely complete (ASC) dictionary and the corresponding VV code rate.

Definition 2 ([42]).

1) *For every infinite sequence, if the probability that the dictionary $\mathcal{D}$ has a prefix of the infinite sequence is one, then $\mathcal{D}$ is termed almost surely complete.*
2) *Let $\mathcal{C}$ be a VV code with a proper and ASC dictionary $\mathcal{D}$ and a VV encoder φ. Then, the coding rate of $\mathcal{C}$ is*

$$R_{\mathcal{C}} = \frac{\sum_{\alpha \in \mathcal{D}} P(\alpha)|\varphi(\alpha)|}{\sum_{\alpha \in \mathcal{D}} P(\alpha)|\alpha|}.$$

We have an example of a dictionary $\mathcal{D} = \{1, 01, 001, \cdots\}$ over $\{0, 1\}$ that is proper and ASC. However, $\mathcal{D}$ is incomplete because the all-zero infinite sequence has no prefix in $\mathcal{D}$.

III. Generalized universal coding of integers

In this section, we first formally define our problem. We consider the stationary and memoryless source $S = (P, \mathcal{A})$, where $\mathcal{A} = \mathbb{N}$ is an infinitely countable alphabet,[6] and P is an unknown probability distribution on $\mathcal{A}$. Encoding source S with VV codes, i.e., we consider mapping sequences of input symbols to $\{0, 1\}^*$.

Second, we define GUCI and explain the rationality of the definition. Let $\mathcal{C} = (\mathcal{D}, \varphi)$ denote a VV code $\mathcal{C}$ with a proper and ASC dictionary $\mathcal{D}$ and a VV encoder φ, where the dictionary $\mathcal{D}$ is over the alphabet $\mathcal{A}$ and $\mathcal{C} = (\mathcal{D}, \varphi)$ satisfies the prefix property. A VV code $\mathcal{C} = (\mathcal{D}, \varphi)$ that satisfies the prefix property means that $\varphi(\beta)$ is not a prefix of $\varphi(\alpha)$ for any $\beta \neq \alpha \in \mathcal{D}$. By introducing the VV codes, the definition of GUCI is as follows.

Definition 3 (GUCI). *Let $\mathcal{C} = (\mathcal{D}, \varphi)$ be a VV code. The encoder $\varphi : \mathcal{D} \to \{0, 1\}^*$ is prefix free. $\mathcal{C}$ is called generalized universal if there exists a constant $K_{\mathcal{C}}$ independent of P for all DPDs P with $0 < H(P) < \infty$ such that*

$$\frac{R_{\mathcal{C}}}{H(P)} \leq K_{\mathcal{C}}, \tag{5}$$

where $R_{\mathcal{C}}$ is the coding rate of $\mathcal{C}$ and $K_{\mathcal{C}}$ denotes the expansion factor of GUCI $\mathcal{C}$. Let

$$K_{\mathcal{C}}^* \triangleq \min\{K_{\mathcal{C}} \,|\, \forall \text{ DPD } P \text{ and } 0 < H(P) < \infty\}$$

denote the minimum expansion factor of GUCI $\mathcal{C}$.

Remark 1. *Each GUCI $\mathcal{C}$ has a unique $K_{\mathcal{C}}^*$ corresponding to it. $K_{\mathcal{C}}^*$ is a measure of the compression effect of GUCI $\mathcal{C}$. The smaller $K_{\mathcal{C}}^*$ is, the better. Therefore, we define the smallest minimum expansion factor*

$$K^* \triangleq \inf\{K_{\mathcal{C}}^* \,|\, \forall \text{ GUCI } \mathcal{C}\}.$$

Let us note that there may exist an optimal GUCI $\mathcal{C}$ such that $K_{\mathcal{C}}^* = K^*$, or there may be only one family of GUCIs $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} K_{\mathcal{C}_n}^* = K^*$.

Next, we discuss the rationality of Definition 3. First, the definition of GUCI is an extension of UCI. Comparing inequality (1) with inequality (5), since the denominator of the fraction on the left-hand side of (1) removes the $\max$ function, it is extended from this perspective. The numerators of the fractions on the left-hand side of the two inequalities are essentially equivalent. Since the variable-length code is a special VV code, when the dictionary $\mathcal{D}$ of the VV code is equal to the alphabet $\mathbb{N}$, the VV code degenerates into a variable-length code. Let us note that when the VV code $\mathcal{C} = (\mathcal{D}, \varphi) = (\mathbb{N}, \varphi)$, $\mathcal{C}$ is a variable-length code with a coding rate

$$\begin{aligned} R_{\mathcal{C}} &= \frac{\sum_{\alpha \in \mathbb{N}} P(\alpha)|\varphi(\alpha)|}{\sum_{\alpha \in \mathbb{N}} P(\alpha) \times 1} \\ &= \sum_{n=0}^{\infty} P(n)|\varphi(n)| \\ &= E_P(L_{\mathcal{C}}). \end{aligned}$$

At this time, $R_{\mathcal{C}}$ denotes the expected codeword length of $\mathcal{C}$. Thus, $E_P(L_{\mathcal{C}})$ is a special $R_{\mathcal{C}}$. Essentially, both $R_{\mathcal{C}}$ and $E_P(L_{\mathcal{C}})$ represent the average codeword length required for a source symbol. Therefore, for convenience, $R_{\mathcal{C}}$ and $E_P(L_{\mathcal{C}})$ can be collectively referred to as the *average codeword length*.

[6] In the setting of the GUCI problem, we denote the alphabet $\mathcal{A}$ as $\mathbb{N}$, not $\mathbb{N}^+$. For an infinitely countable alphabet $\mathcal{A}$, this is essentially indistinguishable. In this paper, we denote the alphabet $\mathcal{A}$ as $\mathbb{N}$ for ease of expression in subsequent constructions.

The average codeword length $E_P(L_\mathcal{C})$ of the variable-length code $\mathcal{C}$ is greater than or equal to 1. Then, the ratio

$$\frac{R_\mathcal{C}}{H(P)} = \frac{E_P(L_\mathcal{C})}{H(P)}$$

tends to infinity when $H(P)$ tends to 0. Hence, there is no such thing as a variable-length code $\mathcal{C}$ that is GUCI. Therefore, a class of UCIs $\mathcal{C}$ must not be a GUCI.

Finally, we prove that the expansion factor of GUCI has the same property as that of UCI. In a groundbreaking paper [9], Elias proved that $E_P(L_\mathcal{C}) \geq \max\{1, H(P)\}$. Therefore, the expansion factor of UCI is greater than or equal to 1. Before presenting the relevant theorem, we first introduce two important lemmas.

Lemma 2. *[42] Let $S = (P, \mathcal{A})$ denote a discrete memoryless source with entropy $H(P) < \infty$ and a countable alphabet $\mathcal{A}$. Given a VV code $\mathcal{C}$ with a proper and ASC dictionary $\mathcal{D}$; then,*

$$H(\mathcal{D}) = H(P)\overline{l(\mathcal{D})},$$

where $H(\mathcal{D}) = -\sum_{\alpha\in\mathcal{D}} P(\alpha)\log_2 P(\alpha)$ denotes the entropy of $\mathcal{D}$ and $\overline{l(\mathcal{D})} = \sum_{\alpha\in\mathcal{D}} P(\alpha)|\alpha|$ denotes the average length of $\mathcal{D}$.

Lemma 2 was first described by Nishiara *et al.* [42], but they did not provide a complete proof. The complete proof of Lemma 2 can be found in [43].

Lemma 3.
1) *If n positive integers $L_1, L_2, \cdots, L_n$ satisfy $\sum_{i=1}^{n} 2^{-L_i} \leq 1$. Then, there are n positive integers $M_1, M_2, \cdots, M_n$ that satisfy $\sum_{i=1}^{n} 2^{-M_i} = 1$ and $L_i \geq M_i$ for all $i \in \{1, 2, \cdots, n\}$.*
2) *If the integer sequence $\{L_i\}_{i=1}^{\infty}$ satisfies $\sum_{i=1}^{\infty} 2^{-L_i} \leq 1$. Then, there is an integer sequence $\{M_i\}_{i=1}^{\infty}$ that satisfies $\sum_{i=1}^{\infty} 2^{-M_i} = 1$ and $L_i \geq M_i$ for all $i \in \mathbb{N}^+$.*

The proof of Lemma 3 is given in the Appendix. Next, we present a theorem similar to $E_P(L_\mathcal{C}) \geq \max\{1, H(P)\}$ for variable-length codes.

Theorem 1. *Let $S = (P, \mathcal{A})$ denote a discrete memoryless source with entropy $H(P) < \infty$ and a countable alphabet $\mathcal{A}$. Assuming that a VV code $\mathcal{C} = (\mathcal{D}, \varphi)$ satisfies the prefix property, then $R_\mathcal{C} \geq H(P)$.*

Proof. From Lemma 2, we obtain

$$\begin{aligned} R_\mathcal{C} = \frac{\sum_{\alpha\in\mathcal{D}} P(\alpha)|\varphi(\alpha)|}{\overline{l(\mathcal{D})}} &\geq H(P) = \frac{H(\mathcal{D})}{\overline{l(\mathcal{D})}} \\ \iff \sum_{\alpha\in\mathcal{D}} P(\alpha)|\varphi(\alpha)| &\geq H(\mathcal{D}) = -\sum_{\alpha\in\mathcal{D}} P(\alpha)\log_2 P(\alpha) \\ \iff \sum_{\alpha\in\mathcal{D}} P(\alpha)\Big(|\varphi(\alpha)| + \log_2 P(\alpha)\Big) &\geq 0 \\ \iff \sum_{\alpha\in\mathcal{D}} P(\alpha)\log_2 \frac{P(\alpha)}{2^{-|\varphi(\alpha)|}} &\geq 0. \end{aligned}$$

Below, we prove that the last inequality holds. As the codeword set $\{\varphi(\alpha) \mid \alpha \in \mathcal{D}\}$ satisfies the prefix property, we have

$$\sum_{\alpha\in\mathcal{D}} 2^{-|\varphi(\alpha)|} \leq 1$$

due to Kraft inequality [44]. From Lemma 3, we can find an integer set $\{M_\alpha \mid \alpha \in \mathcal{D}\}$ that satisfies

$$\sum_{\alpha\in\mathcal{D}} 2^{-M_\alpha} = 1$$

and $|\varphi(\alpha)| \geq M_\alpha$ for all $\alpha \in \mathcal{D}$. Then,

$$\begin{aligned} \sum_{\alpha\in\mathcal{D}} P(\alpha)\log_2 \frac{P(\alpha)}{2^{-|\varphi(\alpha)|}} &\geq \sum_{\alpha\in\mathcal{D}} P(\alpha)\log_2 \frac{P(\alpha)}{2^{-M_\alpha}} \\ &= D(P \parallel P_M) \\ &\geq 0, \end{aligned}$$

where the probability distribution represented by P_M satisfies $P_M(\alpha) = 2^{-M_\alpha}$ for all $\alpha \in \mathcal{D}$. □

Remark 2. *[45, P11] proves that $\sum_{\alpha\in\mathcal{D}} P(\alpha)|\varphi(\alpha)| \geq H(\mathcal{D})$. Therefore, Theorem 1 can be proved according to [45] and Lemma 2 as well. Since [45] does not use Lemma 3 in the proof, we provide an independent proof here.*

From Theorem 1, we obtain that the expansion factor of GUCI is greater than or equal to 1.

IV. Explicit construction of GUCIs

In this section, the explicit structure of a family of GUCIs is proposed. The traditional UCI cannot satisfy the inequality (4), as there is no constant $K_\mathcal{C}$ available to satisfy the inequality (4) when $H(P)$ tends to 0. Thus, we consider the case in which $H(P)$ tends to 0 in the construction of GUCIs. When $H(P)$ tends to 0, without loss of generality, we assume that $P(0)$ tends to 1. In this case, the nonnegative integer source string contains several consecutive $0'$s, which the RLE can compress. Specifically, the proposed VV code $\mathcal{C} = (\mathcal{D}, \varphi)$ is the concatenation of RLE and UCI ψ. The encoding process is as follows.

First, the dictionary $\mathcal{D}_{RLE}$ selected by the encoder is

$$\mathcal{D}_{RLE} = \{\underbrace{00\cdots0}_{i}n | i \in \mathbb{N}, n \in \mathbb{N}^+\}.$$

Next, the encoder maps the variable-length string $\underbrace{00\cdots0}_{i}n \in \mathcal{D}_{RLE}$ into the fixed-length string $(i+1, n)$. Finally, the encoder maps string $(i+1, n)$ into $\psi(i+1)\psi(n)$ via UCI ψ. That is, $\varphi_\psi(\underbrace{00\cdots0}_{i}n) = \psi(i+1)\psi(n)$.

$\mathcal{D}_{RLE}$ is proper and incomplete, as the all-zero infinite sequence has no prefix in $\mathcal{D}_{RLE}$. However, as the probability of the all-zero infinite sequence is 0 due to $H(P) > 0$, $\mathcal{D}_{RLE}$ is the ASC. We prove that the constructed VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ is a GUCI when UCI ψ satisfies the easily reachable condition below. First, we present two auxiliary lemmas.

Lemma 4. *The following inequality holds.*

$$-\log_2\Big(P(0)^i P(n)\Big) \geq 1 + \log_2 n + \log_2(i+1),$$

for all DPDs P and every $i \in \mathbb{N}^+$ and $n \in \mathbb{N}^+$.

Proof. Since P is DPD, we obtain

$$P(0)^i P(n) \leq \frac{P(0)^i\Big(1-P(0)\Big)}{n}.$$

Let $g(x) = x^i(1-x)$, for $0 < x < 1$. We know that $g(x)$ is strictly increasing when $x \in (0, \frac{i}{i+1})$ and $g(x)$ is strictly decreasing when $x \in (\frac{i}{i+1}, 1)$ via its derivative. Thus,

$$P(0)^i P(n) \leq \frac{1}{n} \cdot g\left(\frac{i}{i+1}\right) = \frac{1}{n} \cdot \frac{1}{i+1} \cdot \left(\frac{i}{i+1}\right)^i.$$

We prove that the sequence $\{a_i = (\frac{i}{i+1})^i\}_{i=1}^{\infty}$ is strictly monotonically decreasing below. Let

$$b_i = \frac{1}{a_i} = \left(\frac{i+1}{i}\right)^i = \left(1+\frac{1}{i}\right)^i,$$

Then, $\{a_i\}_{i=1}^{\infty}$ strictly monotonically decreasing is equivalent to $\{b_i\}_{i=1}^{\infty}$ strictly monotonically increasing. Due to the inequality of arithmetic and geometric means, we obtain

$$\begin{aligned} b_i &= 1 \cdot \underbrace{\left(1+\frac{1}{i}\right) \cdots \left(1+\frac{1}{i}\right)}_{i} \\ &< \left[\frac{1+i(1+\frac{1}{i})}{i+1}\right]^{i+1} \\ &= \left(1+\frac{1}{i+1}\right)^{i+1} \\ &= b_{i+1}. \end{aligned}$$

Thus,

$$P(0)^i P(n) \leq \frac{1}{n} \cdot \frac{1}{i+1} \cdot \left(\frac{i}{i+1}\right)^i \leq \frac{1}{n} \cdot \frac{1}{i+1} \cdot \frac{1}{2},$$

and hence,

$$-\log_2\Big(P(0)^i P(n)\Big) \geq 1 + \log_2 n + \log_2(i+1).$$

□

Lemma 5. *Given two positive numbers a and b, then*

$$-(2a+b)\log_2\Big(P(0)^i P(n)\Big) \geq 2a + b\log_2 n + b\log_2(i+1), \tag{6}$$

for all DPDs P and every $i \in \mathbb{N}$ and $n \in \mathbb{N}^+$.

Proof. We first consider $i = 0$. In this case, (6) can be rewritten as

$$2a + b\log_2 n \leq -(2a+b)\log_2 P(n).$$

As $P(0) \geq P(1) \geq \cdots \geq P(n) \geq \cdots$, then

$$1 = \sum_{m=0}^{\infty} P(m) \geq \sum_{m=0}^{n} P(m) \geq (n+1)P(n),$$

and hence, $-\log_2 P(n) \geq \log_2(n+1)$, for $n \in \mathbb{N}^+$. Thus,

$$\begin{aligned} -(2a+b)\log_2 P(n) &\geq (2a+b)\log_2(n+1) \\ &= 2a\log_2(n+1) + b\log_2(n+1) \\ &> 2a + b\log_2 n. \end{aligned}$$

Then, we consider $i \geq 1$. Due to Lemma 4, we have

$$\begin{aligned} &-(2a+b)\log_2\Big(P(0)^i P(n)\Big) \\ &\geq (2a+b)\Big(1+\log_2 n + \log_2(i+1)\Big) \\ &> 2a + b\log_2 n + b\log_2(i+1). \end{aligned}$$

□

Now, we present the main theorem in this section.

Theorem 2. *Let $S = (P, \mathcal{A})$ denote a discrete memoryless source with entropy $0 < H(P) < \infty$ and a countable alphabet $\mathcal{A}$. Given that the VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ satisfies*

$$L_\psi(n) \leq a + b\log_2 n, \text{ for } n \in \mathbb{N}^+, \tag{7}$$

where a and b are two positive constants. Then we have

$$\frac{R_{\mathcal{C}}}{H(P)} \leq 2a + b,$$

for all DPDs P.

Proof. From Lemma 5 and (7), we obtain

$$\begin{aligned} |\varphi_\psi(\underbrace{00\cdots0}_{i}n)| &= L_\psi(n) + L_\psi(i+1) \\ &\leq 2a + b\log_2 n + b\log_2(i+1) \\ &\leq -(2a+b)\log_2\Big(P(0)^i P(n)\Big), \end{aligned} \tag{8}$$

for all $i \in \mathbb{N}$ and $n \in \mathbb{N}^+$. From Lemma 2, we have

$$\begin{aligned} \frac{R_{\mathcal{C}}}{H(P)} &= \frac{\sum_{\alpha \in \mathcal{D}_{RLE}} P(\alpha)|\varphi_\psi(\alpha)|}{H(\mathcal{D}_{RLE})} \\ &= \frac{\sum_{i,n} P(0)^i P(n)\Big(L_\psi(n) + L_\psi(i+1)\Big)}{-\sum_{i,n} P(0)^i P(n)\log_2\Big(P(0)^i P(n)\Big)} \\ &\overset{(a)}{\leq} \frac{\sum_{i,n} P(0)^i P(n)\Big[-(2a+b)\log_2\Big(P(0)^i P(n)\Big)\Big]}{-\sum_{i,n} P(0)^i P(n)\log_2\Big(P(0)^i P(n)\Big)} \\ &= 2a + b, \end{aligned}$$

where (a) is due to (8). □

Remark 3. *A variable-length code ψ satisfying* (7) *is a sufficient condition for ψ to be a UCI [13]. To the best of our knowledge, all UCI codes currently proposed satisfy* (7). *Therefore, when we construct a GUCI $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$, we can choose any known UCI code.*

V. The tighter upper bound of K^*

Based on UCIs, we have provided a family of GUCIs. In this section, we explore the expansion factors of some specific GUCIs, and we obtain a tighter upper bound of K^*. For any UCI $\mathcal{C}$, its expansion factor $K_{\mathcal{C}}$ is greater than or equal to 2 [22]. The best-known UCI to date is the ι code [23] with $K_\iota = 2.5$. Therefore, the optimal UCI is in the range $2 \leq K_{\mathcal{C}}^* \leq 2.5$. Theorem 1 shows that K^* is greater than or equal to 1. In this section, we investigate the tighter upper bounds of K^*.

When constructing a VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$, we select the Elias γ code as the UCI ψ. From Theorem 2 and $L_\gamma(n) =$

$1 + 2\lfloor \log_2 n \rfloor \le 1 + 2\log_2 n$, we obtain $K_{\mathcal{C}} = 4$. It has been shown [1] that $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\gamma)$ can reach $K_{\mathcal{C}} = \frac{6}{\log_2 5} \approx 2.584$. However, this result is not tight, and we show that the VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\gamma)$ can achieve $K_{\mathcal{C}} = 2$. First, we present two auxiliary lemmas.

Lemma 6. *For all DPDs P defined on $\mathbb{N}$ and all $m \in \mathbb{N}^+$, we find*

1) $\sum_{j=1}^{m} P(j) \le \frac{m}{m+1}$*;*
2) $\prod_{j=1}^{m} P(j) \le \left(\frac{1}{m+1}\right)^m$*;*
3) *Let* $A_m \triangleq 2^m \times m! \times \left(\frac{1}{m+1}\right)^m$*; then,* $A_m \le 1$*;*
4) *Let* $B_m \triangleq \sum_{j=1}^{m} \Big(1 + \log_2 j + \log_2 P(j)\Big)$*; then,* $B_m \le 0$*.*

Proof.

1) We prove that $\sum_{j=1}^{m} P(j) \le \frac{m}{m+1}$ by contradiction. Let us suppose that there exists a DPD P_0 defined on $\mathbb{N}$ such that $\sum_{j=1}^{m} P_0(j) > \frac{m}{m+1}$. Thus,
$$P_0(0) \le 1 - \sum_{j=1}^{m} P_0(j) < \frac{1}{m+1},$$
and hence,
$$\frac{m}{m+1} > mP_0(0) \ge \sum_{j=1}^{m} P_0(j) > \frac{m}{m+1},$$
which is a contradiction. Thus, the assumption is not true.
2) Due to the inequality of arithmetic and geometric means, we obtain
$$\prod_{j=1}^{m} P(j) \le \left(\frac{\sum_{j=1}^{m} P(j)}{m}\right)^m \le \left(\frac{1}{m+1}\right)^m.$$
3) We prove that $A_m \le 1$ by mathematical induction. When $m = 1$, then $A_1 = 2 \times 1 \times \frac{1}{2} \le 1$. Let us suppose that $A_m \le 1$ holds when $m = n$. When $m = n + 1$, we have
$$\begin{aligned} A_{n+1} &= A_n \times 2(n+1)\left(\frac{1}{n+2}\right)^{n+1} \div \left(\frac{1}{n+1}\right)^n \\ &= 2A_n \times \left(\frac{n+1}{n+2}\right)^{n+1} \\ &\overset{(a)}{\le} 2A_n \times \left(\frac{2}{3}\right)^2 \\ &= \frac{8}{9}A_n \\ &< 1, \end{aligned}$$
where (a) is calculated from the fact that the sequence $\{a_i = (\frac{i}{i+1})^i\}_{i=1}^{\infty}$ is strictly monotonically decreasing.
4) From the above results, we obtain
$$\begin{aligned} B_m &= \sum_{j=1}^{m} \log_2 \Big(2 \times j \times P(j)\Big) \\ &= \log_2 \left(2^m \times m! \times \prod_{j=1}^{m} P(j)\right) \\ &\le \log_2 \left(2^m \times m! \times \left(\frac{1}{m+1}\right)^m\right) \\ &= \log_2 A_m \\ &\le 0. \end{aligned}$$

□

Lemma 7. *For all DPDs P defined on $\mathbb{N}$ and all $m \in \mathbb{N}^+$, we define*
$$S_m \triangleq \sum_{j=1}^{m} P(j)\Big(1 + \log_2 j + \log_2 P(j)\Big).$$
Then, $S_m \le 0$ for all $m \in \mathbb{N}^+$. Furthermore, we have
$$\sum_{j=1}^{\infty} P(j)\Big(1 + \log_2 j + \log_2 P(j)\Big) \le 0.$$

Proof. When $m = 1$, we have $S_1 = P(1)B_1 \le 0$. When $m \ge 2$, we obtain
$$\begin{aligned} S_m &= P(1)B_1 + \sum_{j=2}^{m} P(j)(B_j - B_{j-1}) \\ &= \Big(P(1) - P(2)\Big)B_1 + P(2)B_2 + \sum_{j=3}^{m} P(j)(B_j - B_{j-1}) \\ &\le P(2)B_2 + \sum_{j=3}^{m} P(j)(B_j - B_{j-1}) \\ &\vdots \\ &\le P(m-1)B_{m-1} + P(m)(B_m - B_{m-1}) \\ &\le P(m)B_m \\ &\le 0. \end{aligned}$$
Thus, we have
$$\sum_{j=1}^{\infty} P(j)\Big(1 + \log_2 j + \log_2 P(j)\Big) = \lim_{m\to+\infty} S_m \le \lim_{m\to+\infty} 0 = 0.$$

□

Now, we present the main results of this section.

Theorem 3. *Let $S = (P, \mathcal{A})$ denote a discrete memoryless source with entropy $0 < H(P) < \infty$ and a countable alphabet $\mathcal{A}$. Given a VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ satisfying*
$$L_\psi(n) \le a + 2a \log_2 n, \ \textit{for}\ n \in \mathbb{N}^+, \tag{9}$$
where a is a positive constant. Then,
$$\frac{R_{\mathcal{C}}}{H(P)} \le 2a,$$
for all DPDs P.

Proof. From Lemma 2 and (9), we obtain

$$\begin{aligned}\frac{R_{\mathcal{C}}}{H(P)} &= \frac{\sum_{\alpha\in\mathcal{D}_{RLE}} P(\alpha)|\varphi_\psi(\alpha)|}{H(P)\overline{l(\mathcal{D}_{RLE})}}\\ &= \frac{\sum_{i=0}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)\Big(L_\psi(i+1)+L_\psi(n)\Big)}{H(\mathcal{D}_{RLE})}\\ &\le \frac{2a\sum_{i=0}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)\Big(1+\log_2 n+\log_2(i+1)\Big)}{H(\mathcal{D}_{RLE})}.\end{aligned}$$

Therefore, proving $\frac{R_{\mathcal{C}}}{H(P)} \le 2a$ is equivalent to showing that

$$\sum_{i=0}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)\Big(1+\log_2 n+\log_2(i+1)\Big) \le H(\mathcal{D}_{RLE}). \tag{10}$$

When $i \ge 1$, from Lemma 4, we have

$$1+\log_2 n+\log_2(i+1) \le -\log_2\Big(P(0)^i P(n)\Big).$$

Thus, we obtain

$$\begin{aligned}&\sum_{i=1}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)\Big(1+\log_2 n+\log_2(i+1)\Big)\\ \le &-\sum_{i=1}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)\log_2\Big(P(0)^i P(n)\Big).\end{aligned} \tag{11}$$

When $i = 0$, from Lemma 7, we have

$$\sum_{n=1}^{\infty} P(n)\left(1+\log_2 n\right) \le -\sum_{n=1}^{\infty} P(n)\log_2 P(n). \tag{12}$$

From (11) and (12), (10) holds. □

Remark 4. *As $1 \le L_\psi(1) \le a$, the minimum of a in Theorem 3 is 1. From Theorem 3 and $L_\gamma(n) \le 1+2\log_2 n$, we know that $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\gamma)$ can reach $K_{\mathcal{C}} = 2$. Thus, the Elias γ code achieves the best case of Theorem 3.*

Next, we discuss K_Cs for GUCIs constructed by using other classical UCIs. Before presenting the results, we derive an upper bound on the length function in the form $a + 2a\log_2 n$ for each code listed in Lemma 1.

Lemma 8. *For all $n \in \mathbb{N}^+$,*

1) *the length function of the δ code satisfies $L_\delta(n) \le \frac{4}{3} + \frac{8}{3}\log_2 n$;*
2) *the length function of the η code satisfies $L_\eta(n) \le \frac{6}{1+2\log_2 5} + \frac{12}{1+2\log_2 5}\log_2 n$;*
3) *the length function of the θ code satisfies $L_\theta(n) \le \frac{4}{3} + \frac{8}{3}\log_2 n$;*
4) *the length function of the ι code satisfies $L_\iota(n) \le \frac{4}{3} + \frac{8}{3}\log_2 n$;*
5) *the length function of the ω code satisfies $L_\omega(n) \le \frac{11}{9} + \frac{22}{9}\log_2 n$.*

Proof. 1) Obviously, the inequality $\lfloor\log_2(1+x)\rfloor \le \frac{1}{6}+\frac{5}{6}x$ holds, for all $x \in \mathbb{N}$. Thus, we obtain

$$\begin{aligned}L_\delta(n) &= 1+\lfloor\log_2 n\rfloor + 2\lfloor\log_2(1+\lfloor\log_2 n\rfloor)\rfloor\\ &\le 1+\lfloor\log_2 n\rfloor + 2\left(\frac{1}{6}+\frac{5}{6}\lfloor\log_2 n\rfloor\right)\\ &\le \frac{4}{3}+\frac{8}{3}\log_2 n.\end{aligned}$$

TABLE I: The expansion factors that can be achieved for VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$

UCI ψ	expansion factor $K_{\mathcal{C}}$
γ code	2
η code	$\frac{12}{1+2\log_2 5} \approx 2.13$
ω code	$\frac{22}{9} \approx 2.44$
δ code, θ code and ι code	$\frac{8}{3} \approx 2.67$

2) Let $f(n) \triangleq \frac{6}{1+2\log_2 5} + \frac{12}{1+2\log_2 5}\log_2 n$. We directly verify $L_\eta(n) \le f(n)$ for $n < 16$. When $n \ge 16$, we have

$$\begin{aligned}L_\eta(n) &= 3+\lfloor\log_2(n-1)\rfloor + \lfloor\frac{\lfloor\log_2(n-1)\rfloor}{2}\rfloor\\ &\le 3+\frac{3}{2}\lfloor\log_2 n\rfloor\\ &\le 1+2\lfloor\log_2 n\rfloor\\ &< f(n).\end{aligned}$$

3) Obviously, the inequality $\frac{5}{3}+\frac{3}{2}\lfloor\log_2 x\rfloor \le \frac{5}{3}x$ holds, for all $x \in \mathbb{N}^+$. Thus, we obtain $L_\theta(1) = 1 < \frac{4}{3}$ and

$$\begin{aligned}L_\theta(n) &= 3+\lfloor\log_2 n\rfloor+\lfloor\log_2\lfloor\log_2 n\rfloor\rfloor+\lfloor\frac{\lfloor\log_2\lfloor\log_2 n\rfloor\rfloor}{2}\rfloor\\ &\le 3+\lfloor\log_2 n\rfloor+\frac{3}{2}\lfloor\log_2\lfloor\log_2 n\rfloor\rfloor\\ &= \frac{4}{3}+\lfloor\log_2 n\rfloor+\left(\frac{5}{3}+\frac{3}{2}\lfloor\log_2\lfloor\log_2 n\rfloor\rfloor\right)\\ &\le \frac{4}{3}+\frac{8}{3}\log_2 n,\end{aligned}$$

for $n \ge 2$.

4) We obtain $L_\iota(1) = 1 < \frac{4}{3}$ and

$$\begin{aligned}L_\theta(n) &= 2+\lfloor\log_2 n\rfloor+\lfloor\frac{1+\lfloor\log_2 n\rfloor}{2}\rfloor\\ &\le \frac{5}{2}+\frac{3}{2}\lfloor\log_2 n\rfloor\\ &= \frac{4}{3}+\frac{8}{3}\lfloor\log_2 n\rfloor+\frac{7}{6}\left(1-\lfloor\log_2 n\rfloor\right)\\ &\le \frac{4}{3}+\frac{8}{3}\log_2 n,\end{aligned}$$

for $n \ge 2$.

5) We directly verify $L_\omega(n) \le \frac{11}{9}+\frac{22}{9}\log_2 n$ for $n<16$. When $n \ge 16$, we have

$$\begin{aligned}L_\omega(n) &\le 3+2\lfloor\log_2 n\rfloor\\ &= \frac{11}{9}+\frac{22}{9}\lfloor\log_2 n\rfloor+\frac{4}{9}\left(4-\lfloor\log_2 n\rfloor\right)\\ &\le \frac{11}{9}+\frac{22}{9}\log_2 n.\end{aligned}$$

□

From Theorem 3 and Lemma 8, Table I lists the expansion factors of GUCIs when choosing various UCIs. Let us note that, based on previous proofs, we find that $1 \le K^* \le 2$.

VI. Comparison of the average codeword lengths of UCI ψ and GUCI $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$

In this section, we compare the expected codeword length $E_P(L_\psi)$ of UCI ψ and the coding rate $R_{\mathcal{C}}$ of the GUCI

$\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$. Intuitively, when the Shannon entropy $H(P)$ is small or $P(0)$ is large, $E_P(L_\psi) > R_\mathcal{C}$; When the Shannon entropy $H(P)$ is large or $P(0)$ is small, $E_P(L_\psi) < R_\mathcal{C}$. The following two conclusions can be drawn from the investigation in this section. First, when $P(0)$ is relatively large, $E_P(L_\psi) > R_\mathcal{C}$. That is, a sufficient condition for the average codeword length of the GUCI to be shorter than that of the UCI is obtained. Second, when the Shannon entropy $H(P)$ is very large or $P(0)$ is not large, there are still cases where $E_P(L_\psi) > R_\mathcal{C}$. A detailed discussion is given below.

First, we recall a definition. If a class of UCIs ψ satisfies

$$L_\psi(m) \leq L_\psi(m+1), \text{ for } m \in \mathbb{N}. \tag{13}$$

Then, ψ is termed *minimal* [9]. For all DPDs P, $E_P(L_\mathcal{C})$ can be minimized when (13) is satisfied. Hence, the definition is natural.

$E_P(L_\psi)$ and $R_\mathcal{C}$ are defined as

$$E_P(L_\psi) = \sum_{n=0}^{\infty} P(n) L_\psi(n+1),$$

$$R_\mathcal{C} = \frac{\sum_{\alpha \in \mathcal{D}_{RLE}} P(\alpha)|\varphi_\psi(\alpha)|}{\overline{l(\mathcal{D}_{RLE})}}.$$

Let us note that the definition of $E_P(L_\psi)$, where n starts from 0, is slightly different from that defined in the Introduction section, where n starts from 1. Furthermore, we obtain

$$\begin{aligned}
\overline{l(\mathcal{D}_{RLE})} &= \sum_{i=0}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n)(i+1) \\
&= \sum_{n=1}^{\infty} P(n) \left(\sum_{i=0}^{\infty} P(0)^i + \sum_{i=0}^{\infty} i P(0)^i \right) \\
&= \Big(1 - P(0)\Big) \left(\frac{1}{1-P(0)} + \frac{P(0)}{\Big(1-P(0)\Big)^2} \right) \\
&= \frac{1}{1-P(0)},
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{\alpha \in \mathcal{D}_{RLE}} P(\alpha)|\varphi_\psi(\alpha)| \\
=& \sum_{i=0}^{\infty}\sum_{n=1}^{\infty} P(0)^i P(n) \Big(L_\psi(i+1) + L_\psi(n) \Big) \\
=& \sum_{n=1}^{\infty} P(n) \sum_{i=0}^{\infty} P(0)^i L_\psi(i{+}1) {+} \sum_{i=0}^{\infty} P(0)^i \sum_{n=1}^{\infty} P(n) L_\psi(n) \\
=& \Big(1-P(0)\Big) \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) + \frac{\sum_{n=1}^{\infty} P(n) L_\psi(n)}{1-P(0)}.
\end{aligned}$$

Thus, we have

$$R_\mathcal{C} = \Big(1-P(0)\Big)^2 \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) + \sum_{n=1}^{\infty} P(n) L_\psi(n).$$

Let

$$\begin{aligned}
\Delta \triangleq& R_\mathcal{C} - E_P(L_\psi) \\
=& \Big(1-P(0)\Big)^2 \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) - P(0) L_\psi(1) \\
&- \sum_{n=1}^{\infty} P(n) \Big(L_\psi(n+1) - L_\psi(n) \Big) \\
=& \Big(1-P(0)\Big)^2 \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) - P(0) L_\psi(1) \\
&- \sum_{n=1}^{\infty} P(n) \Delta_\psi(n),
\end{aligned} \tag{14}$$

where $\Delta_\psi(n) \triangleq L_\psi(n+1) - L_\psi(n)$ is the jump value of ψ at n. We note that Δ is a function of the probability distribution P and the length function $L_\psi(\cdot)$.

When analyzing Δ without imposing restrictions on ψ, it is difficult to determine which is larger between Δ and 0. Then, we restrict ψ with reasonable conditions, and we conclude that $\Delta < 0$ when $P(0)$ is relatively large. The main theorem is proposed.

Theorem 4. *When constructing a VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$, the UCI ψ is minimal, and its length function satisfies*

$$L_\psi(n) \leq a + b\lfloor \log_2 n \rfloor, \textit{ for } 2 \leq n \in \mathbb{N}^+,$$

where a and b are two positive constants. If there exists $t \in (0,1)$ such that

$$L_\psi(1)\left(t{+}\frac{1}{t}{-}3\right){+}a\,(1{-}t){+}b\,(1{-}t)\left(1{+}t^2{+}\frac{t^6}{1{-}t^8}\right) \leq 0,$$

then $\Delta < 0$ when $P(0) \geq t$, i.e., $R_\mathcal{C} < E_P(L_\psi)$ when $P(0) \geq t$.

Proof. First, we perform calculations. We know that

$$\begin{aligned}
&\sum_{i=1}^{\infty} P(0)^i \lfloor \log_2(i+1) \rfloor \\
=& \sum_{n=1}^{\infty} n \left(\sum_{j=2^n-1}^{2^{n+1}-2} P(0)^j \right) \\
=& \sum_{n=1}^{\infty} n \cdot \frac{P(0)^{2^n-1} - P(0)^{2^{n+1}-1}}{1-P(0)} \\
=& \frac{1}{1-P(0)} \lim_{m \to +\infty} \sum_{n=1}^{m} n \Big(P(0)^{2^n-1} - P(0)^{2^{n+1}-1} \Big) \\
=& \frac{1}{1-P(0)} \lim_{m \to +\infty} \Big[P(0) {-} m P(0)^{2^{m+1}-1} {+} \sum_{n=2}^{m} P(0)^{2^n-1} \Big] \\
=& \frac{1}{1-P(0)} \sum_{n=1}^{\infty} P(0)^{2^n-1}
\end{aligned} \tag{15}$$

and that

$$
\begin{aligned}
&\sum_{i=0}^{\infty} P(0)^i L_\psi(i+1)\\
=&L_\psi(1) + \sum_{i=1}^{\infty} P(0)^i L_\psi(i+1)\\
\le&L_\psi(1) + \frac{aP(0)}{1-P(0)} + b\sum_{i=1}^{\infty} P(0)^i \lfloor \log_2(i+1)\rfloor\\
=&L_\psi(1) + \frac{aP(0)}{1-P(0)} + \frac{b}{1-P(0)}\sum_{n=1}^{\infty} P(0)^{2^n-1}\\
<&L_\psi(1)+\frac{aP(0)}{1-P(0)}+\frac{bP(0)}{1-P(0)}\Big(1+P(0)^2+\sum_{n=0}^{\infty} P(0)^{6+8n}\Big)\\
=&L_\psi(1)+\frac{aP(0)}{1-P(0)}+\frac{bP(0)}{1-P(0)}\left(1+P(0)^2+\frac{P(0)^6}{1-P(0)^8}\right).
\end{aligned}
$$

We then have

$$
\begin{aligned}
\Delta &\overset{(c)}{\le}\Big(1-P(0)\Big)^2\sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) - P(0)L_\psi(1)\\
&<\Big(1-P(0)\Big)^2\Big[L_\psi(1)+\frac{aP(0)}{1-P(0)}\\
&\quad+\frac{bP(0)}{1-P(0)}\left(1+P(0)^2+\frac{P(0)^6}{1-P(0)^8}\right)\Big]-P(0)L_\psi(1)\\
&= P(0)\Big[L_\psi(1)\left(P(0)+\frac{1}{P(0)}-3\right)+a\Big(1-P(0)\Big)\\
&\quad+b\Big(1-P(0)\Big)\left(1+P(0)^2+\frac{P(0)^6}{1-P(0)^8}\right)\Big]\\
&\overset{(d)}{\le} P(0)\Big[L_\psi(1)\left(t+\frac{1}{t}-3\right)+a\,(1-t)\\
&\quad+b\,(1-t)\left(1+t^2+\frac{t^6}{1-t^8}\right)\Big]\\
&\le 0,
\end{aligned}
$$

where (c) is because ψ is minimal, (d) follows the monotonic decrease in $g_1(x) = x + \frac{1}{x} - 3$, $g_2(x) = 1 - x$ and $g_3 = (1-x)(1+x^2+\frac{x^6}{1-x^8})$ over the interval $(0, 1)$. □

We apply several UCIs to Theorem 4 to yield several examples. In the first example, the corresponding parameters of the Elias γ code are $L_\gamma(1) = 1$, $a = 1$ and $b = 2$. Let $h(x) \triangleq (x+\frac{1}{x}-3)+(1-x)+2(1-x)(1+x^2+\frac{x^6}{1-x^8})$. We know that $h(0.81) < 0$ by calculation. Thus, when $P(0) \ge 0.81$, the coding rate $R_{\mathcal{C}}$ of $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\gamma)$ is less than the expected codeword length $E_P(L_\gamma)$ of the Elias γ code. In another example, the corresponding parameters of the ι code are $L_\gamma(1) = 1$, $a = 2.5$ and $b = 1.5$. When $P(0) \ge 0.83$, $R_{\mathcal{C}}$ of $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\iota)$ is less than $E_P(L_\iota)$ of the ι code.

Let us note that $R_{\mathcal{C}}$ is less than $E_P(L_\psi)$ not only when the entropy is small. In other words, $P(0)$ is relatively large, which does not mean that the entropy is small. For example, we consider the probability distribution

$$P_1 = \left(P_1(0) = 0.9, P_1(1) = P_1(2) = \cdots P_1(n) = \frac{1}{10n}\right).$$

Due to Theorem 4, we know that $R_{\mathcal{C}} < E_{P_1}(L_\gamma)$. However, taking the limit $n \to +\infty$, the entropy $H(P_1) = 0.1\log_2(10n) - 0.9\log_2 0.9$ tends to infinity. This tells us that when the entropy is large, $R_{\mathcal{C}}$ is still less than $E_P(L_\gamma)$. However, if $P(0)$ is relatively large, a long string of zeros is prone to appear. Considering the structure of $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$, it is reasonable that $R_{\mathcal{C}}$ is less than $E_P(L_\psi)$ at this time.

Finally, we explore the situation in which $P(0)$ is not large. This situation must be analyzed with a specific UCI. Since the Elias γ code performs best regarding the expansion factor, we use the Elias γ code for analysis. Due to $L_\gamma(n) = 1 + 2\lfloor\log_2 n\rfloor$ and (15), (14) can be rewritten as

$$
\begin{aligned}
\Delta =&\Big(1-P(0)\Big)^2\sum_{i=0}^{\infty} P(0)^i\Big(1+2\lfloor\log_2(i+1)\rfloor\Big)-P(0)\\
&-\sum_{n=1}^{\infty} P(n)\Delta_\gamma(n)\\
\overset{(a)}{=}&\ 2\Big(1-P(0)\Big)^2\sum_{i=1}^{\infty} P(0)^i\lfloor\log_2(i+1)\rfloor\\
&+\Big(1-P(0)\Big)^2\sum_{i=0}^{\infty} P(0)^i - P(0) - 2\sum_{t=1}^{\infty} P(2^t-1)\\
=&\ 2\Big(1-P(0)\Big)\sum_{n=1}^{\infty} P(0)^{2^n-1}+1-2\sum_{t=0}^{\infty} P(2^t-1),
\end{aligned}
$$

where (a) is because when $n \in \{2^t - 1 \,|\, t \in \mathbb{N}^+\}$,

$$\Delta_\gamma(n) = (1+2\lfloor\log_2 2^t\rfloor) - (1+2\lfloor\log_2(2^t-1)\rfloor) = 2\,;$$

when the positive integer is $n \notin \{2^t - 1 \,|\, t \in \mathbb{N}^+\}$, $\Delta_\gamma(n) = 0$. Considering the probability distribution

$$
\begin{aligned}
P_2 = \Big(&P_2(0) = P_2(1) = P_2(2) = P_2(3) = 0.24,\\
&P_2(4) = P_2(5) = \cdots P_2(n+3) = \frac{1}{25n}\Big),
\end{aligned}
$$

we obtain

$$
\begin{aligned}
\Delta <&\ 2P_2(0)\Big(1-P_2(0)\Big)\Big(1+P_2(0)^2+\frac{P_2(0)^6}{1-P_2(0)^8}\Big)\\
&+1-2P_2(0)-2P_2(1)-2P_2(3)\\
=&\ 0.3648\times\left(1+0.0576+\frac{0.24^6}{1-0.24^8}\right)-0.44\\
\approx&-0.054.
\end{aligned}
$$

Taking the limit $n \to +\infty$, the entropy $H(P_2)$ tends to infinity. Therefore, when $P(0)$ is not large, $R_{\mathcal{C}}$ may be less than $E_P(L_\psi)$. Let us note that from the calculation, we know that the main reason for $\Delta < 0$ in this example is the displacement term $-\sum_{n=1}^{\infty} P(n)\Delta_\psi(n)$.

In summary, Theorem 4 shows that when $P(0)$ is relatively large, $R_{\mathcal{C}}$ is less than $E_P(L_\psi)$. When $P(0)$ is not large, it is difficult to judge whether Δ is positive or negative. When the entropy is very large or $P(0)$ is not large, it is still possible that Δ is less than 0.

VII. Asymptotically optimal GUCI

In this section, the asymptotically optimal GUCI is discussed. First, the formal definition of the asymptotically optimal GUCI is given as follows.

Definition 4. *(asymptotically optimal GUCI) $\mathcal{C}$ is said to be an asymptotically optimal GUCI if $\mathcal{C}$ is a class of GUCIs and there exists a function $T_{\mathcal{C}}(\cdot)$ such that*

$$\frac{R_{\mathcal{C}}}{H(P)} \le T_{\mathcal{C}}(H(P)), \tag{16}$$

for all DPD P with $0 < H(P) < \infty$ and

$$\lim_{H(P)\to+\infty} T_{\mathcal{C}}(H(P)) = 1.$$

Then, we present an important property of the asymptotically optimal GUCI.

Theorem 5. *Let $S = (P, \mathcal{A})$ denote a discrete memoryless source with entropy $0 < H(P) < \infty$ and a countable alphabet $\mathcal{A}$. Let $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ denote the VV code satisfying* (7) *and let UCI ψ be minimal. If there exists a function $R_\psi(\cdot)$ satisfying* (2) *and*

$$\lim_{H(P)\to+\infty} R_\psi(H(P)) = c,$$

where c is constant, then there exists a function $T_{\mathcal{C}}(\cdot)$ satisfying (16) *and*

$$\lim_{H(P)\to+\infty} T_{\mathcal{C}}(H(P)) = c.$$

Proof. From (14), we have

$$\frac{R_{\mathcal{C}}}{H(P)} = \frac{\Delta + E_P(L_\psi)}{H(P)},$$

where $\Delta = \Big(1-P(0)\Big)^2 \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1) - P(0)L_\psi(1) - \sum_{n=1}^{\infty} P(n)\Delta_\psi(n)$. (7) indicates that there exists an integer n_0 such that

$$L_\psi(n) \le n, \text{ for } n_0 \le n \in \mathbb{N}^+. \tag{17}$$

From (17) and $\Delta_\psi(n) \ge 0$, for all $n \in \mathbb{N}$, we obtain

$$\begin{aligned}\Delta &< \Big(1-P(0)\Big)^2 \sum_{i=0}^{\infty} P(0)^i L_\psi(i+1)\\ &< \sum_{i=0}^{n_0-1} L_\psi(i+1) + \Big(1-P(0)\Big)^2 \sum_{i=n_0}^{\infty} P(0)^i (i+1)\\ &= \sum_{i=0}^{n_0-1} L_\psi(i+1) + \Big(n_0 + 1 - n_0 P(0)\Big) P(0)^{n_0}\\ &\overset{(a)}{\le} \sum_{i=0}^{n_0-1} L_\psi(i+1) + 1,\end{aligned}$$

where (a) is because $f(x) = (n_0 + 1 - n_0 x)x^{n_0}$ is strictly monotonically increasing over the interval $(0, 1)$ by calculating the derivative of $f(x)$. Furthermore, when $H(P) \ge 1$, we find that

$$\begin{aligned}\frac{R_{\mathcal{C}}}{H(P)} &= \frac{\Delta + E_P(L_\psi)}{H(P)}\\ &< \frac{\sum_{i=0}^{n_0-1} L_\psi(i+1) + 1}{H(P)} + R_\psi(H(P)).\end{aligned}$$

When $H(P) < 1$, we have $\frac{R_{\mathcal{C}}}{H(P)} \le 2a + b$ due to Theorem 2. We define

$$T_{\mathcal{C}}(H(P)) \triangleq \begin{cases} 2a + b, & \text{if } H(P) < 1,\\ V(H(P)), & \text{if } H(P) \ge 1,\end{cases}$$

where $V(H(P)) \triangleq \frac{\sum_{i=0}^{n_0-1} L_\psi(i+1)+1}{H(P)} + R_\psi(H(P))$. Hence, we obtain $\frac{R_{\mathcal{C}}}{H(P)} \le T_{\mathcal{C}}(H(P))$ and

$$\begin{aligned}&\lim_{H(P)\to+\infty} T_{\mathcal{C}}(H(P))\\ &= \lim_{H(P)\to+\infty} \frac{\sum_{i=0}^{n_0-1} L_\psi(i+1)+1}{H(P)} + \lim_{H(P)\to+\infty} R_\psi(H(P))\\ &= c.\end{aligned}$$

□

Finally, we present the theorem describing the relationship between the asymptotically optimal UCI and the asymptotically optimal GUCI.

Theorem 6. *For any discrete memoryless source $S = (P, \mathcal{A})$ with entropy $0 < H(P) < \infty$ and a countable alphabet $\mathcal{A}$, the VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ satisfies* (7)*, and UCI ψ is minimal and asymptotically optimal. Then, $\mathcal{C}$ is an asymptotically optimal GUCI.*

Proof. From Theorem 2, we know that $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ is a GUCI. Due to Theorem 5 and

$$\lim_{H(P)\to+\infty} R_\psi(H(P)) = 1,$$

we obtain

$$\lim_{H(P)\to+\infty} T_{\mathcal{C}}(H(P)) = 1.$$

Therefore, $\mathcal{C}$ is an asymptotically optimal GUCI. □

VIII. Conclusions

In this paper, GUCIs are proposed to resolve the UCI issue in which the expected codeword length ratio to $H(P)$ is not bounded by a constant factor K when $H(P)$ is extremely small.

We construct a VV code $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\psi)$ through RLE and UCI ψ. We prove that $\mathcal{C}$ is a GUCI or an asymptotically optimal GUCI when UCI ψ satisfies certain conditions. We propose a class of GUCIs $\mathcal{C} = (\mathcal{D}_{RLE}, \varphi_\gamma)$ to reach the expansion factor $K_{\mathcal{C}} = 2$, and we show that the smallest minimum expansion factor is in the range $1 \le K^* \le 2$. Furthermore, when the entropy is very large or $P(0)$ is not large, it is still possible that the coding rate $R_{\mathcal{C}}$ is less than the expected codeword length $E_P(L_\psi)$. Future work is as follows.

1) Is it possible to achieve $K^*_{\mathcal{C}} < 2$ when VV code $\mathcal{C}$ is obtained by concatenating other FV codes and UCI? For example, one can use Lempel-Zip coding, but the extra storage space should be considered when using dictionary coding.
2) The exact value of K^* is still unknown.

Appendix

Lemma 9 (Lemma 3 Restated)**.**

1) *If n positive integers $L_1, L_2, \cdots, L_n$ satisfy $\sum_{i=1}^{n} 2^{-L_i} \leq 1$. Then, there are n positive integers $M_1, M_2, \cdots, M_n$ that satisfy $\sum_{i=1}^{n} 2^{-M_i} = 1$ and $L_i \geq M_i$ for all $i \in \{1, 2, \cdots, n\}$.*
2) *If the integer sequence $\{L_i\}_{i=1}^{\infty}$ satisfies $\sum_{i=1}^{\infty} 2^{-L_i} \leq 1$. Then, there is an integer sequence $\{M_i\}_{i=1}^{\infty}$ that satisfies $\sum_{i=1}^{\infty} 2^{-M_i} = 1$ and $L_i \geq M_i$ for all $i \in \mathbb{N}^+$.*

Proof. If the Kraft inequality does not become an equality, i.e.,

$$\sum_i 2^{-L_i} < 1,$$

Then, $\{L_i\}$ is called *redundant*. If the Kraft inequality is equal, i.e.,

$$\sum_i 2^{-L_i} = 1,$$

Then, $\{L_i\}$ is called *complete*. Let $[n] \triangleq \{1, 2, \cdots, n\}$.

1) If $\{L_i\}_{i=1}^{n}$ is complete, then let $M_i = L_i$ for all $i \in [n]$. If $\{L_i\}_{i=1}^{n}$ is redundant, we suppose the largest value among n integers $\{L_i\}_{i=1}^{n}$ is L_t. Then, 2^{-L_i} is a multiple of 2^{-L_t} for all $i \in [n]$. Furthermore, we obtain

$$\sum_{i=1}^{n} 2^{-L_i} = N \cdot 2^{-L_t},$$

where $N \in \mathbb{N}^+$. We consider n integers $\{\widetilde{L}_i\}_{i=1}^{n}$, where $\widetilde{L}_t = L_t - 1$ and $\widetilde{L}_i = L_i$ for all $i \in [n] \setminus \{t\}$. Since

$$N \cdot 2^{-L_t} = \sum_{i=1}^{n} 2^{-L_i} < 1 = 2^{L_t} \cdot 2^{-L_t}$$

and both N and 2^{L_t} are integers, we have $N + 1 \leq 2^{L_t}$. Thus,

$$\begin{aligned}\sum_{i=1}^{n} 2^{-\widetilde{L}_i} &= 2^{-L_t} + \sum_{i=1}^{n} 2^{-L_i} \\ &= (N+1) \cdot 2^{-L_t} \\ &\leq 2^{L_t} \cdot 2^{-L_t} \\ &= 1.\end{aligned}$$

If $\{\widetilde{L}_i\}_{i=1}^{n}$ is complete, then let $M_i = \widetilde{L}_i$ for all $i \in [n]$. If $\{\widetilde{L}_i\}_{i=1}^{n}$ is redundant, then we repeat the above process until the Kraft inequality is an equality. Let us note that the above process is performed at most

$$\frac{1 - \sum_{i=1}^{n} 2^{-L_i}}{2^{-L_t}} = 2^{L_t}\Big(1 - \sum_{i=1}^{n} 2^{-L_i}\Big)$$

times since the sum of the n terms increases by at least 2^{-L_t} each time. Therefore, after a finite number of the above procedures, the Kraft inequality is equal.

2) Without loss of generality, we can assume that $L_n \leq L_{n+1}$ for all $n \in \mathbb{N}^+$. There is no maximum value in the integer sequence $\{L_i\}_{i=1}^{\infty}$. Otherwise, assuming L_t is the maximum value, then

$$\infty = \sum_{i=1}^{\infty} 2^{-L_t} \leq \sum_{i=1}^{\infty} 2^{-L_i} \leq 1.$$

If $\{L_i\}_{i=1}^{\infty}$ is complete, then let $M_i = L_i$ for all $i \in \mathbb{N}^+$. If $\{L_i\}_{i=1}^{\infty}$ is redundant, let

$$a_1 \triangleq \sum_{i=1}^{\infty} 2^{-L_i} \leq 1.$$

Let n_1 be the only positive integer satisfying the following inequalities:

$$\frac{1 - a_1}{2} < 2^{-x} \leq 1 - a_1.$$

Then, we obtain

$$1 - a_1 - 2^{-n_1} < \frac{1 - a_1}{2}.$$

Because there is no maximum value in the integer sequence $\{L_i\}_{i=1}^{\infty}$, there exists a sufficiently large $t \in \mathbb{N}^+$ such that

$$1 - a_1 - 2^{-n_1} + 2^{-L_t} < \frac{1 - a_1}{2}.$$

Thus, we have

$$2^{-n_1} - 2^{-L_t} > \frac{1 - a_1}{2} > 0.$$

We assume that the value of n_1 is between L_{k-1} and L_k, i.e., $L_{k-1} \leq n_1 \leq L_k$. We consider an integer sequence $\{\widetilde{L(1)}_i\}_{i=1}^{\infty}$, where

$$\widetilde{L(1)}_i = \begin{cases} n_1, & \text{if } i = k\ , \\ L_{i-1}, & \text{if } i = k+1, k+2, \cdots, t\ , \\ L_i, & \text{otherwise.} \end{cases}$$

According to the definitions of $\{\widetilde{L(1)}_i\}_{i=1}^{\infty}$, we know that $L_i \geq \widetilde{L(1)}_i$ for all $i \in \mathbb{N}^+$. Let $a_2 \triangleq \sum_{i=1}^{\infty} 2^{-\widetilde{L(1)}_i}$; then,

$$a_1 < a_1 + 2^{-n_1} - 2^{-L_t} = a_2 \leq 1 - 2^{-L_t} < 1.$$

We continue the above process indefinitely. Then, we obtain the sequence $\{a_i\}_{i=1}^{\infty}$ and the sequence $\{n_i\}_{i=1}^{\infty}$. The following two facts are proven below.

a) $n_m < n_{m+1}$ for all $m \in \mathbb{N}^+$.
b) $\lim\limits_{m \to +\infty} a_m = 1$.

Since

$$1 - a_{m+1} = 1 - a_m - 2^{-n_m} + 2^{-L_t} < \frac{1 - a_m}{2}, \tag{18}$$

we have

$$2^{-n_{m+1}} \leq 1 - a_{m+1} < \frac{1 - a_m}{2} < 2^{-n_m}.$$

Furthermore, we obtain $n_m < n_{m+1}$ for all $m \in \mathbb{N}^+$. The first fact is proven. Since $\{a_i\}_{i=1}^{\infty}$ is a strictly increasing sequence with an upper bound, the limit of sequence $\{a_i\}_{i=1}^{\infty}$ exists. Therefore, we assume that

$$l \triangleq \lim_{m \to +\infty} a_m \leq 1.$$

Taking the limit on both sides of (18), we find that

$$1 - l = \lim_{m \to +\infty} (1 - a_{m+1}) \leq \lim_{m \to +\infty} \frac{1 - a_m}{2} = \frac{1 - l}{2}.$$

Furthermore, we have $l \geq 1$. Therefore, $l = \lim_{m \to +\infty} a_m = 1$. The second fact is proven.

Finally, we construct the integer sequence $\{M_i\}_{i=1}^{\infty}$. For any given $i \in \mathbb{N}^+$, from the first fact, there exists an integer n_m such that $\widetilde{L(m-1)}_{k-1} \leq n_m \leq \widetilde{L(m-1)}_k$ and $k-1 \geq i$. Let $M_i \triangleq \widetilde{L(m)}_i$. We note that $n_m < n_{m+1} < n_{m+2} < \cdots$. Thus, $\widetilde{L(m)}_i = \widetilde{L(m+1)}_i = \widetilde{L(m+2)}_i = \cdots$. Therefore, M_i is well defined. Due to the definitions of M_i and $\{\widetilde{L(m)}_i\}_{i=1}^{\infty}$, we know that $L_i \geq M_i$ for all $i \in \mathbb{N}^+$. Furthermore, we find that

$$\sum_{i=1}^{\infty} 2^{-M_i} = \lim_{m \to +\infty} a_m = 1.$$

The proof is complete. □

References

[1] W. Yan and S.-J. Lin, "Generalized universal coding of integers," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Oct. 2021, pp. 1–6.

[2] B. P. Tunstall, *Synthesis of noiseless compression codes*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, 1967.

[3] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[4] G. L. Khodak, "Bounds of redundancy estimates for word-based encoding of sequences produced by a bernoulli source (Russian)," *Problemy Peredachi Informatsii*, vol. 8(2), pp. 21–32, 1972.

[5] Y. Bugeaud, M. Drmota, and W. Szpankowski, "On the construction of (explicit) Khodak's code and its analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5073–5086, Nov. 2008.

[6] S. A. Savari and W. Szpankowski, "On the analysis of variable-to-variable length codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2002, p. 176.

[7] M. Drmota and W. Szpankowski, "Variable-to-variable codes with small redundancy rates," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2004, p. 91.

[8] V. I. Levenshtein, "On the redundancy and delay of decodable coding of natural numbers (in Russian)," *Problems of Cybernetics*, vol. 20, pp. 173–179, 1968.

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

[10] R. M. Capocelli, "Flag encodings related to the zeckendorf representation of integers," in *Sequences, Combinatorics, Compression, Security, and Transmission*. New York, NY, USA: Springer-Verlag, 1990, pp. 449–466.

[11] Q. F. Stout, "Improved prefix encodings of the natural numbers (corresp.)," *IEEE Trans. Inf. Theory*, vol. 26, no. 5, pp. 607–609, Sep. 1980.

[12] H. Yamamoto, "A new recursive universal code of the positive integers," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 717–723, Mar. 2000.

[13] K. Lakshmanan, "On universal codeword sets," *IEEE Trans. Inf. Theory*, vol. 27, no. 5, pp. 659–662, Sep. 1981.

[14] A. Apostolico and A. S. Fraenkel, "Robust transmission of unbounded strings using Fibonacci representations," *IEEE Trans. Inf. Theory*, vol. 33, no. 2, pp. 238–245, Mar. 1987.

[15] M. Wang, "Almost asymptotically optimal flag encoding of the integers," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 324–326, Mar. 1988.

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

[17] T. Amemiya and H. Yamamoto, "A new class of the universal representation for the positive integers," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. E76A, no. 3, pp. 447–452, Mar. 1993.

[18] B. T. Ávila and R. M. C. de Souza, "Meta-Fibonacci codes: Efficient universal coding of natural numbers," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2357–2375, Apr. 2017.

[19] R. Ahlswede, T. S. Han, and K. Kobayashi, "Universal coding of integers and unbounded search trees," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 669–682, Mar. 1997.

[20] S. Leung-Yan-Cheong and T. Cover, "Some equivalences between shannon entropy and kolmogorov complexity," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 331–338, May 1978.

[21] L. Allison, A. S. Konagurthu, and D. F. Schmidt, "On universal codes for integers: Wallace tree, Elias omega and beyond," in *Proc. 2021 Data Compression Conference (DCC)*, Mar. 2021, pp. 313–322.

[22] W. Yan and S.-J. Lin, "On the minimum of the expansion factor for universal coding of integers," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7309–7319, Nov. 2021.

[23] W. Yan and S.-J. Lin, "A tighter upper bound of the expansion factor for universal coding of integers and its code constructions," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4429–4438, Jul. 2022.

[24] K. Daily, P. Rigor, S. Christley, X. Xie, and P. Baldi, "Data structures and compression algorithms for high-throughput sequencing technologies," *BMC Bioinform.*, vol. 11, p. 514, Oct. 2010.

[25] J. J. Selva and X. Chen, "SRComp: Short read sequence compression using burstsort and Elias omega coding," *PLOS ONE*, vol. 8, no. 12, pp. 1–7, 12 Dec. 2013.

[26] J. Zobel and A. Moffat, "Inverted files for text search engines," *ACM Comput. Surv.*, vol. 38, no. 2, pp. 1–56, Jul. 2006.

[27] I. E. Richardson, *The H.264 Advanced Video Compression Standard*. John Wiley & Sons, Ltd., 2010.

[28] V. Sze, M. Budagavi, and G. J. Sullivan, *High Efficiency Video Coding (HEVC), Algorithms and Architectures*. Springer, 2014.

[29] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM Journal of Research and Development*, vol. 23, no. 2, pp. 149–162, Mar. 1979.

[30] Y.-H. Liu and D.-C. Wu, "A high-capacity performance-

preserving blind technique for reversible information hiding via MIDI files using delta times," *Multimedia Tools and Appl.*, vol. 79, no. 25–26, pp. 17 281–17 302, Jul. 2020.

[31] W. Yan, S.-J. Lin, and Y. S. Han, "A new metric and the construction for evolving 2-threshold secret sharing schemes based on prefix coding of integers," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2906–2915, May 2023.

[32] J. L. Bentley and A. C.-C. Yao, "An almost optimal algorithm for unbounded searching," *Inf. Process. Lett.*, vol. 5, no. 3, pp. 82–87, Aug. 1976.

[33] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

[34] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 6, pp. 674–682, Nov. 1978.

[35] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd ed.* NY, USA: Wiley, 2006.

[36] L. Gyorfi, I. Pali, and E. Van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 267–271, Jan. 1994.

[37] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 358–373, Jan. 2009.

[38] J. F. Silva and P. Piantanida, "Universal weak variable-length source coding on countably infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 649–668, Jan. 2020.

[39] A. N. Azman and E. Ferda, "A novel psychovisual threshold on large dct for image compression," *The Scientific World Journal*, vol. 2015, p. 821497, Mar. 2015.

[40] J. Capon, "A probabilistic model for run-length coding of pictures," *IRE Trans. Inf. Theory*, vol. 5, no. 4, pp. 157–163, Dec. 1959.

[41] F. Jelinek and K. Schneider, "On variable-length-to-block coding," *IEEE Trans. Inf. Theory*, vol. 18, no. 6, pp. 765–774, Nov. 1972.

[42] M. Nishiara and H. Morita, "Almost surely complete parsing and variable-to-variable length coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2000, p. 347.

[43] W. Yan and Y. S. Han, "A complete proof of an important theorem for variable-to-variable length codes," *arXiv: 2309.06267*, 2023. [Online]. Available: https://arxiv.org/abs/2309.06267

[44] L. G. Kraft, "A device for quantizing, grouping, and coding amplitude-modulated pulses," Master's thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1949.

[45] E. Gassiat, *Universal Coding and Order Identification by Model Selection Methods*. Springer Cham, 2018.

Wei Yan (Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics and the Ph.D. degree in Cyberspace Security from the University of Science and Technology of China (USTC), Hefei, China, in 2017 and 2022, respectively. From 2022 to 2023, he was a Researcher with the Theory Laboratory, 2012 Labs, Huawei Technologies Company Ltd. He is currently a Lecturer with the National University of Defense Technology (NUDT), Hefei, China. His research interest includes coding theory, data compression, and secret sharing.

Yunghsiang S. Han (Fellow, IEEE) was born in Taipei, Taiwan, in 1962. He received B.Sc. and M.Sc. degrees in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1984 and 1986, respectively, and a Ph.D. from the School of Computer and Information Science, Syracuse University, Syracuse, NY, in 1993. From 1986 to 1988, he was a lecturer at Ming-Hsin Engineering College, Hsinchu, Taiwan. He was a teaching assistant from 1989 to 1992 and a research associate in the School of Computer and Information Science at Syracuse University from 1992 to 1993. From 1993 to 1997, he was an Associate Professor in the Department of Electronic Engineering at Hua Fan College of Humanities and Technology, Taipei Hsien, Taiwan. He was with the Department of Computer Science and Information Engineering at National Chi Nan University, Nantou, Taiwan from 1997 to 2004. He was promoted to Professor in 1998. He was a visiting scholar in the Department of Electrical Engineering at the University of Hawaii at Manoa, HI from June to October 2001, the SUPRIA visiting research scholar in the Department of Electrical Engineering and Computer Science and CASE center at Syracuse University, NY from September 2002 to January 2004 and July 2012 to June 2013, and the visiting scholar in the Department of Electrical and Computer Engineering at the University of Texas at Austin, TX from August 2008 to June 2009. He was with the Graduate Institute of Communication Engineering at National Taipei University, Taipei, Taiwan from August 2004 to July 2010. From August 2010 to January 2017, he was Chair Professor with the Department of Electrical Engineering at the National Taiwan University of Science and Technology. From February 2017 to February 2021, he was with the School of Electrical Engineering & Intelligentization at Dongguan University of Technology, China. Now he is with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China. He is also a Chair Professor at National Taipei University since February 2015. His research interests are in error-control coding, wireless networks, and security.

Dr. Han was a winner of the 1994 Syracuse University Doctoral Prize and a Fellow of IEEE. One of his papers won the prestigious 2013 ACM CCS Test-of-Time Award in cybersecurity.