from Lemma 3 that they have the same mutual information. Lemma 3 also implies that, besides the dimensions, the mutual information can only depend on one of the distribution parameters, namely α.

- For the two-dimensional **ordered Weinman exponential distribution** the mutual information is

$$I(X_1, X_2) = \begin{cases} \ln\left(\frac{1}{\theta_1}\left(\frac{\theta_0}{2} - \theta_1\right)\right) + \Psi\left(\frac{\theta_0}{\theta_0 - 2\theta_1}\right) + \Psi(1), & \text{if } \theta_1 < \frac{\theta_0}{2} \\ \Psi(1), & \text{if } \theta_0 = \frac{\theta_0}{2} \\ \ln\left(\frac{1}{\theta_1}\left(\theta_1 - \frac{\theta_0}{2}\right)\right) + \Psi\left(\frac{2\theta_1}{2\theta_1 - \theta_0}\right) + \Psi(1), & \text{if } \theta_1 > \frac{\theta_0}{2}. \end{cases} \tag{25}$$

  It is obtained through direct integration.

- The mutual information of the **Gamma-exponential distribution** is

$$I(X_1, X_2) = \Psi(\theta_2) - \ln\theta_2 + \frac{1}{\theta_2}. \tag{26}$$

  It is derived through direct integration. Note that it depends only on the parameter θ_2. This is a consequence of Lemma 3.

References

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, V. H. Deutsch, Ed. Frankfurt/Main, Germany: abridged version, 1984, .

[2] N. A. Ahmed and D. V. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Trans. Inform. Theory*, vol. 35, pp. 688–692, May 1989.

[3] B. C. Arnold, *Pareto Distributions*. Burtonsville, MD: Int. Coop. Publishing House, 1983.

[4] J. Beirlant, E. J. Dudewicz, L. Györfi, and E .C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, pp. 17–39, June 1997.

[5] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Inst. Math. Statist., 1986, vol. 9. Lecture Notes.

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

[7] G. A. Darbellay, "Predictability: An information-theoretic perspective," in *Signal Analysis and Prediction*, A. Procházka, J. Uhlíř, P. J. W. Rayner, and N. G. Kingsbury, Eds. Boston, MA: Birkhäuser-Verlag, 1998, pp. 249–262.

[8] G. A. Darbellay and I. Vajda. (1998, Feb.) Entropy Expressions for Multivariate Continuous Distributions. UTIA, Academy of Sciences, Prague, Czech Republic. [Online]. Available: http://siprint.utia.cas.cz/darbellay

[9] ——, "Estimation of the mutual information with data-dependent partitions," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1315–1320, May 1999.

[10] S. Ihara, *Information Theory for Continuous Systems*. Singapore, Singapore: World Scientific, 1993.

[11] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley, 1972.

[12] A. C. G. V. Lazo and P. N. Rathie, "On the entropy of continuous probability distributions," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 120–122, 1978.

Source Code with Cost as a Nonuniform Random Number Generator

Te Sun Han, *Fellow, IEEE*, and Osamu Uchida, *Student Member, IEEE*

***Abstract*—We show that an optimal source code with a cost function for code symbols can be regarded as a random number generator generating a random sequence (not necessarily a sequence of *fair coin* bits) as the target distribution in the sense that the normalized conditional divergence between the distribution of the generated codeword distribution and the target distribution vanishes as the block length tends to infinity.**

***Index Terms*—Cost function, general source, normalized conditional divergence, random number generation, source code with cost.**

I. Introduction

In the problem of random number generation, the purpose is in general to simulate the source $\boldsymbol{Y}$ with a prescribed distribution $\boldsymbol{q}$ (called the *target* distribution) by using the source $\boldsymbol{X}$ with a given probability $\boldsymbol{p}$ (called the *coin* distribution). von Neumann [1] has initially addressed this problem. He has considered the problem of simulating a fair random bit by repeatedly using a biased coin with an unknown distribution. Elias [2] has clarified that the optimal expected number of generated fair random bits per coin toss is equal asymptotically to the entropy rate of the source $\boldsymbol{X}$. Moreover, Vembu and Verdú [3] have shown that the optimal rate at which we can generate fair random bits from a general source $\boldsymbol{X}$ with arbitrary accuracy in the sense of some vanishing distance (e.g., the variational distance, the d-bar distance, and the normalized divergence) between the distribution of the generated codeword process and the uniform distribution is equal to $\liminf_{n\to\infty}(1/n)H(X^n)$. On the other hand, it was conjectured for a long time on the basis of the folklore that an output sequence from an optimal source code is a *uniform* random sequence, because any incompressible sequence seemingly looks like a uniform random sequence. Visweswariah *et al.* [4] and Han [5] have independently made clear that this folklore is in fact true, that is, they have shown that an optimal variable-length source code can be regarded as a variable-length random number generator in the sense that the normalized divergence distance between the distribution of the generated codeword process and the *uniform* distribution actually vanishes as the block length tends to infinity.

On the other hand, as is well known, if we impose *unequal costs* on code symbols, it is no longer optimal to use the code which minimizes the average codeword length. It is instead required to use the codes which minimize the average codeword cost. Several studies have been made on the source coding problem in this interesting setting. Karp [6] has given an algorithm for constructing minimum-redundancy prefix codes with unequal cost symbols. Iwata *et al.* [7] have proposed a universal lossless coding algorithm for minimizing the average codeword cost for stationary sources based on the Lempel-Ziv (LZ78) code. Hereafter, we shall call the code constructed in the case with unequal cost symbols the *source code with cost*. Naturally, there would exist a bias in the frequency of code symbols generated by an optimal source code

Manuscript received July 23, 1999; revised October 18, 1999.

The authors are with the Graduate School of Information Systems, University of Electro-Communications, Chofugaoka 1-5-1, Chofu, Tokyo 182-8585, Japan.

Communicated by N. Merhav, Associate Editor for Source Coding.
Publisher Item Identifier S 0018-9448(00)01673-4.

S0018–9448/00$10.00 © 2000 IEEE

with cost. Can we then consider the optimal variable-length source code with cost as a variable-length *nonuniform* random number generator? The purpose of this correspondence is to demonstrate that the answer to this question is “yes.”

II. Variable-Length Source Coding with Cost

In order to state our problem in a more formal manner, let $\mathcal{X}$ be a *countably infinite* source alphabet and $\mathcal{Y}$ be a *finite* code alphabet, respectively. In the sequel all the logarithms are taken to the base $K \equiv |\mathcal{Y}|$, where $|\mathcal{Y}|$ denotes the cardinality of $\mathcal{Y}$. We denote the set of all nonnull finite-length sequences taken from $\mathcal{Y}$ by $\mathcal{Y}^*$. In this correspondence, we consider quite general sources as follows. Let us define a *general source* as an infinite sequence

$$\boldsymbol{X} = \{X^n = (X_1^{(n)}, \cdots X_n^{(n)})\}_{n=1}^{\infty}$$

of n-dimensional random variables X^n where each component random variable $X_i^{(n)} (1 \le i \le n)$ takes values in $\mathcal{X}$. It should be noted here that each component of X^n may change depending on block length n. This implies that the sequence $\boldsymbol{X}$ is quite general in the sense that it may not satisfy even the consistency condition as usual processes. The class of sources thus defined covers a very wide range of source including all nonstationary and/or nonergodic sources.

We define the *cost function* $c : \mathcal{Y}^* \to \mathbf{R}^+ \equiv (0, +\infty]$ as follows: First, each symbol $y \in \mathcal{Y}$ is assigned the corresponding cost $c(y)$ such that $0 < c(y) \le +\infty$ $(\forall y \in \mathcal{Y})$, and then the *additive* cost $c(\boldsymbol{y})$ of $\boldsymbol{y} = (y_1, y_2, \cdots, y_k) \in \mathcal{Y}^k$ is defined by

$$c(\boldsymbol{y}) \equiv \sum_{i=1}^{k} c(y_i). \tag{1}$$

Definition 1: R is called an *achievable variable-length source coding cost-rate* for the source $\boldsymbol{X}$ if there exists a variable-length prefix encoder $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ given the cost function $c : \mathcal{Y}^* \to \mathbf{R}^+$ such that

$$\limsup_{n \to \infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} \le R$$

and the infimum of R that are achievable variable-length source coding cost-rates is denoted by $R_v^{\mathrm{c}}(\boldsymbol{X})$, which we call the *infimum achievable variable-length source coding cost-rate*. □

Then, we have the following variable-length source coding theorem with cost[1] for the general source $\boldsymbol{X}$.

Theorem 1:

$$R_v^{\mathrm{c}}(\boldsymbol{X}) = \frac{1}{\alpha_c} \limsup_{n \to \infty} \frac{1}{n} H(X^n) \tag{2}$$

where the *cost capacity* α_c is the positive unique root α of the equation

$$\sum_{y \in \mathcal{Y}} K^{-\alpha c(y)} = 1$$

and

$$H(X^n) \equiv - \sum_{\boldsymbol{x} \in \mathcal{X}^n} P_{X^n}(\boldsymbol{x}) \log P_{X^n}(\boldsymbol{x}).$$

Proof: See the Appendix. □

III. Source Code with Cost as a Nonuniform Random Number Generator

In this section we address the relationship between source codes with cost and nonuniform independent and identically distributed (i.i.d.) random number generators. Given a variable-length prefix encoder $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$, we define for any positive integer m as

$$\mathcal{D}_m \equiv \{\boldsymbol{x} \in \mathcal{X}^n \mid l(\varphi_n(\boldsymbol{x})) = m\}$$

where $l(\cdot)$ denotes the length of a string and we put

$$\mathcal{J}(\varphi_n) \equiv \{m | \Pr\{X^n \in \mathcal{D}_m\} > 0\}.$$

For any $m \in \mathcal{J}(\varphi_n)$, we define X_m^n as the random variable taking values in $\mathcal{D}_m$ with the distribution given by

$$P_{X_m^n}(\boldsymbol{x}) \equiv \frac{P_{X^n}(\boldsymbol{x})}{\Pr\{X^n \in \mathcal{D}_m\}} \qquad (\boldsymbol{x} \in \mathcal{D}_m).$$

For any positive integer m, $V^{(m)}$ indicates an i.i.d. sequence of length m. Let us now define the conditional divergence $D(\varphi_n(X^n) \| V^{(I_n)} | I_n)$ by

$$D(\varphi_n(X^n) \| V^{(I_n)} | I_n) \equiv \sum_{m \in \mathcal{J}(\varphi_n)} \Pr\{I_n = m\} D(\varphi_n(X_m^n) \| V^{(m)})$$

where I_n is the random variable such that $I_n = m$ for $X^n \in \mathcal{D}_m$.

Then, the following theorem shows that, with the cost function $c : \mathcal{Y}^* \to \mathbf{R}^+$, the optimal variable-length source code with cost can be considered as a variable-length random number generator generating the variable-length i.i.d. random sequence subject to the distribution $\boldsymbol{q}_c$ corresponding to the cost function $c : \mathcal{Y}^* \to \mathbf{R}^+$, in the sense that the normalized conditional divergence between the distribution of the generated codeword process and the i.i.d. target distribution vanishes as block length n tends to infinity.

Theorem 2: We assume that the entropy rate of the general source $\boldsymbol{X}$ has the limit $\lim_{n \to \infty} (1/n) H(X^n)$.[2] Let $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ be any *optimal* variable-length prefix encoder in the sense that

$$\lim_{n \to \infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} = R_v^{\mathrm{c}}(\boldsymbol{X}). \tag{3}$$

If we define the probability distribution $\boldsymbol{q}_c = \{q_c(y)\}_{y \in \mathcal{Y}}$ corresponding to the cost function c by

$$q_c(y) = K^{-\alpha_c c(y)} \qquad (y \in \mathcal{Y}) \tag{4}$$

then we have

$$\lim_{n \to \infty} \frac{1}{n} D(\varphi_n(X^n) \| V^{(I_n)} | I_n) = 0 \tag{5}$$

where $V^{(m)}$ stands for the i.i.d. sequence of length m subject to the distribution $\boldsymbol{q}_c$.

Proof: Let $\boldsymbol{y} \equiv (y_1, y_2, \cdots, y_m) \in \mathcal{Y}^m$. From (4) we have

$$\begin{aligned} \Pr\{V^{(m)} = \boldsymbol{y}\} &= \prod_{i=1}^{m} q_c(y_i) \\ &= \prod_{i=1}^{m} K^{-\alpha_c c(y_i)} \\ &= K^{-\alpha_c c(\boldsymbol{y})} \end{aligned}$$

[1] This kind of theorem has first been shown by Krause [8] for independent and identically distributed (i.i.d.) finite alphabet sources.

[2] The sources satisfying this assumption are not limited only to stationary sources.

for all $m \in \mathcal{J}(\varphi_n)$. Then

$$\begin{aligned} D(\varphi_n(X_m^n)\|V^{(m)}) &= \sum_{\boldsymbol{y}\in\mathcal{Y}^m} \Pr\{\varphi_n(X_m^n)=\boldsymbol{y}\}\log\frac{\Pr\{\varphi_n(X_m^n)=\boldsymbol{y}\}}{\Pr\{V^{(m)}=\boldsymbol{y}\}} \\ &= \alpha_c \sum_{\boldsymbol{y}\in\mathcal{Y}^m} \Pr\{\varphi_n(X_m^n)=\boldsymbol{y}\}c(\boldsymbol{y}) - H(\varphi_n(X_m^n)) \\ &= \alpha_c \sum_{\boldsymbol{y}\in\mathcal{Y}^m} \Pr\{\varphi_n(X_m^n)=\boldsymbol{y}\}c(\boldsymbol{y}) - H(X_m^n) \end{aligned}$$

where the last equality follows from the fact that φ_n is the one-to-one mapping. Thus we have

$$\begin{aligned} \rho_n &\equiv \frac{1}{n}\sum_{m\in\mathcal{J}(\varphi_n)} \Pr\{I_n=m\} D(\varphi_n(X_m^n)\|V^{(m)}) \\ &= \frac{\alpha_c}{n}\sum_{m\in\mathcal{J}(\varphi_n)}\sum_{\boldsymbol{y}\in\mathcal{Y}^m} \Pr\{I_n=m\}\Pr\{\varphi_n(X_m^n)=\boldsymbol{y}\}c(\boldsymbol{y}) \\ &\quad - \frac{1}{n}\sum_{m\in\mathcal{J}(\varphi_n)} \Pr\{I_n=m\}H(X_m^n) \\ &= \frac{\alpha_c}{n}E\{c(\varphi_n(X^n))\} - \frac{1}{n}H(X^n|I_n) \\ &= \frac{\alpha_c}{n}E\{c(\varphi_n(X^n))\} - \frac{1}{n}H(X^n) + \frac{1}{n}H(I_n). \end{aligned} \tag{6}$$

Let

$$c_{\min} \equiv \min_{y\in\mathcal{Y}} c(y) > 0$$

then it follows from $c_{\min}I_n \le c(\varphi_n(X^n))$ that

$$E(I_n) \le \frac{E\{c(\varphi_n(X^n))\}}{c_{\min}}$$

which, together with (6) and the inequality (cf. [9])

$$H(I_n) \le \log[e(E(I_n))]$$

yields

$$\begin{aligned} \rho_n \le{} & \frac{\alpha_c}{n}E\{c(\varphi_n(X^n))\} - \frac{1}{n}H(X^n) \\ & + \frac{1}{n}\log\left[e\left(\frac{E\{c(\varphi_n(X^n))\}}{c_{\min}}\right)\right]. \end{aligned} \tag{7}$$

We see from (3) that

$$\lim_{n\to\infty}\frac{1}{n}\log\left[e\left(\frac{E\{c(\varphi_n(X^n))\}}{c_{\min}}\right)\right] = 0.$$

On the other hand, a consequence of Theorem 1 is

$$R_v^{\mathrm{c}}(\boldsymbol{X}) = \frac{1}{\alpha_c}\lim_{n\to\infty}\frac{1}{n}H(X^n). \tag{8}$$

Thus by (3), (7), and (8) we conclude that

$$\limsup_{n\to\infty}\rho_n \le \alpha_c R_v^{\mathrm{c}}(\boldsymbol{X}) - \alpha_c R_v^{\mathrm{c}}(\boldsymbol{X}) = 0$$

which proves (5). □

Remark 1: We point out that Iwata *et al.*'s universal code [7] satisfies the condition (3) for any stationary source $\boldsymbol{X}$, and, therefore, their code can be regarded as providing a *universal* algorithm for nonuniform i.i.d. random number generation in the sense of (5), although it works only when the source alphabet $\mathcal{X}$ is *finite*.

IV. Comparison with Previous Results

Han [5] has earlier established the following result on the *optimal* variable-length prefix code with *equal cost* $c(y) = 1$ $(\forall y \in \mathcal{Y})$, i.e., $c(\boldsymbol{y}) = l(\boldsymbol{y})$ $(\forall \boldsymbol{y} \in \mathcal{Y}^*)$.

Theorem 3 [5]: We assume that the entropy rate of the general source $\boldsymbol{X}$ has the limit $\lim_{n\to\infty}(1/n)H(X^n)$. Let $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ be any *optimal* variable-length prefix encoder satisfying

$$\lim_{n\to\infty}\frac{1}{n}E\{l(\varphi_n(X^n))\} = R_v(\boldsymbol{X}) \tag{9}$$

where $R_v(\boldsymbol{X})$ is *the infimum of achievable variable-length source coding rates*. Then, we have

$$\lim_{n\to\infty}\frac{1}{n}D(\varphi_n(X^n)\|U^{(I_n)}|I_n) = 0 \tag{10}$$

where $U^{(m)}$ is the i.i.d. sequence subject to *uniform* distribution on $\mathcal{Y}^m$. □

We notice here (cf. [5]) that, under the assumption of Theorem 3, $R_v(\boldsymbol{X})$ is given by

$$R_v(\boldsymbol{X}) = \lim_{n\to\infty}\frac{1}{n}H(X^n).$$

It is easy to check that Theorem 3 is a special case of Theorem 2, because, in the case where all code symbols have equal cost $c(y) = 1$, the cost capacity $\alpha_c = 1$ and hence $R_v^{\mathrm{c}}(\boldsymbol{X}) = R_v(\boldsymbol{X})$. Our proof of Theorem 2 is just paralleling the original proof of Theorem 3, and hence Theorem 2 is a straightforward generalization of Theorem 3. On the other hand, Visweswariah *et al.* [4] have also shown a variant of Theorem 3, i.e., they have shown that the *optimal* variable-length source code with equal cost $c(y) = 1$ can be considered as a random number generator in the following sense.

Theorem 4: Let $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ be any variable-length prefix encoder satisfying the condition (9), where the source alphabet $\mathcal{X}$ is *finite*, unlike in Theorems 2 and 3. Then, there exists a sequence of sets G_n of positive integers such that

$$\lim_{n\to\infty}\Pr\{I_n \in G_n\} = 1$$

$$\lim_{n\to\infty}\max_{m\in G_n}\frac{1}{m}D(\varphi_n(X_m^n)\|U^{(m)}) = 0.$$ □

However, it does not seem to be easy to generalize Theorem 4 to be valid also in the case with unequal costs $c(y)$. One reason is that the rather intractable set G_n intervenes in Theorem 4 but not in Theorem 3. It should be noted that the proof demonstrated in this correspondence does not need the assumption that the source alphabet $\mathcal{X}$ is *finite* and also that either of Theorems 3 or 4 does not imply one another because $G_n \ne \mathcal{J}(\varphi_n)$ in general.

Remark 2: The existence of the limit $\lim_{n\to\infty}(1/n)H(X^n)$ for the source $\boldsymbol{X}$ is the necessary and sufficient condition for (10) to hold under the condition (9). To see this, we need the following theorem on the variable-length random number generation. First, we call R an *achievable variable-length intrinsic randomness rate* for the source $\boldsymbol{X}$ if there exists a variable-length mapping $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ such that

$$\liminf_{n\to\infty}\frac{1}{n}E\{l(\varphi_n(X^n))\} \ge R$$

and

$$\lim_{n\to\infty}\frac{1}{n}D(\varphi_n(X^n)\|U^{(I_n)}|I_n) = 0.$$

Moreover, the supremum of R that are achievable variable-length intrinsic randomness rates is denoted by $S_v^*(\boldsymbol{X})$, which we call the *supremum achievable variable-length intrinsic randomness rate*. Then, we have

Theorem 5 (Han [5], [10]): For any general source $\boldsymbol{X}$ with a *countably infinite* source alphabet $\mathcal{X}$

$$S_v^*(\boldsymbol{X}) = \liminf_{n\to\infty} \frac{1}{n} H(X^n). \qquad \square$$

Since the sufficiency is implied by Theorem 3, it suffices to show the necessity. Suppose that (10) holds. Then, from (10) and Theorem 5, we have

$$\liminf_{n\to\infty} \frac{1}{n} E\{l(\varphi_n(X^n))\} \le S_v^*(\boldsymbol{X}) = \liminf_{n\to\infty} \frac{1}{n} H(X^n).$$

Moreover, by means of Theorem 1 with $c(y) = 1$ $(\forall y \in \mathcal{Y})$

$$\limsup_{n\to\infty} \frac{1}{n} E\{l(\varphi_n(X^n))\} \ge R_v(\boldsymbol{X}) = \limsup_{n\to\infty} \frac{1}{n} H(X^n).$$

As a consequence, since (9) implies

$$\limsup_{n\to\infty} \frac{1}{n} E\{l(\varphi_n(X^n))\} = \liminf_{n\to\infty} \frac{1}{n} E\{l(\varphi_n(X^n))\}$$

it follows that

$$\liminf_{n\to\infty} \frac{1}{n} H(X^n) \ge \limsup_{n\to\infty} \frac{1}{n} H(X^n)$$

which claims that the source $\boldsymbol{X}$ must have the limit

$$\lim_{n\to\infty} (1/n) H(X^n) \qquad \square$$

V. Variable-Length Coding with General Cost Function

In Section III, we have shown that the optimal variable-length source code with the *additive* cost function$c : \mathcal{Y}^* \to \mathbf{R}^+$ defined by (1) can be regarded as a variable-length random number generator generating the variable-length *i.i.d.* random sequence $V^{(I_n)}$ subject to the distribution $\boldsymbol{q}_c$ depending on the cost function c. In the same spirit, we may consider the problem of generating a more general stochastic process instead of $V^{(I_n)}$. To do so, what kind of cost function should we introduce? In the following, we consider the generation of an arbitrarily prescribed general stochastic process (which may be nonstationary or nonergodic) satisfying the consistency condition

$$q(\boldsymbol{y}) = \sum_{y \in \mathcal{Y}} q(\boldsymbol{y}y) \qquad (\boldsymbol{y} \in \mathcal{Y}^*) \tag{11}$$

where q denotes the probability measure. We denote the conditional probability of $y_i \in \mathcal{Y}$ given the sequence $y_1^{i-1} \equiv (y_1, y_2, \cdots, y_{i-1}) \in \mathcal{Y}^{i-1}$ by $q(y_i|y_1^{i-1})$ and we assume that there exist some constants $q_{\min}, q_{\max}$ such that

$$0 \le q(y_i|y_1^{i-1}) \le q_{\max} < 1 \qquad (\forall i, \forall y_i \in \mathcal{Y}, \forall y_1^{i-1} \in \mathcal{Y}^{i-1}) \tag{12}$$

$$0 < q_{\min} \le \inf_{i,\, y_i,\, y_1^{i-1} :\, q(y_i|y_1^{i-1}) > 0} q(y_i|y_1^{i-1}). \tag{13}$$

Using this conditional probability, the probability $q(\boldsymbol{y})$ of $\boldsymbol{y} \in \mathcal{Y}^l$ is written as

$$q(\boldsymbol{y}) = \prod_{i=1}^{l} q(y_i|y_1^{i-1}) \qquad (\boldsymbol{y} \in \mathcal{Y}^l).$$

Let us now define the *general cost function* $c : \mathcal{Y}^* \to \mathbf{R}^+$ as

$$c(\boldsymbol{y}) \equiv -\log q(\boldsymbol{y}) = -\sum_{i=1}^{l} \log q(y_i|y_1^{i-1}) \qquad (\boldsymbol{y} \in \mathcal{Y}^l). \tag{14}$$

Define the *conditional cost* $c(y_i|y_1^{i-1})$ of $y_i \in \mathcal{Y}$ given the sequence $y_1^{i-1} \in \mathcal{Y}^{i-1}$ by

$$c(y_i|y_1^{i-1}) \equiv -\log q(y_i|y_1^{i-1})$$

and call the root $\alpha = \alpha_c$ of the equation

$$\sum_{y_i \in \mathcal{Y}} K^{-\alpha c(y_i|y_1^{i-1})} = 1$$

the *cost capacity* α_c of the general cost function c. It is then obvious that $\alpha_c = 1$ for all $y_1^{i-1} \in \mathcal{Y}^{i-1}$. Then, as a general version of Theorem 1, we have the following variable-length source coding theorem with the general cost function (14) for the general source $\mathbf{X}$. First, let us call R an *achievable variable-length source coding cost-rate* for the source $\boldsymbol{X}$ if there exists a variable-length prefix encoder $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ with the general cost function (14) such that

$$\limsup_{n\to\infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} \le R$$

and the infimum of R that are achievable variable-length source coding cost-rates is denoted by $R_v^{\mathrm{c}}(\boldsymbol{X})$, which we call the *infimum achievable variable-length source coding cost-rate.*

Theorem 6:

$$R_v^{\mathrm{c}}(\boldsymbol{X}) = \frac{1}{\alpha_c} \limsup_{n\to\infty} \frac{1}{n} H(X^n) \qquad (\alpha_c = 1).$$

Proof: On the basis of the assumption (13), we see that there exists a constant $c_{\max}$ such that

$$\sup_{i,\, y_i,\, y_1^{i-1} :\, c(y_i|y_1^{i-1}) < \infty} c(y_i|y_1^{i-1}) \le c_{\max} < \infty$$

Then, Theorem 6 follows in entirely the same manner as in the proof of Theorem 1, provided that the additive cost $c(y_i)$ is replaced by the conditional cost $c(y_i|y_1^{i-1})$, and accordingly $q(y_i) = K^{-\alpha_c c(y_i)}$ by $q(y_i|y_1^{i-1}) = K^{-\alpha_c c(y_i|y_1^{i-1})}$. □

Finally, we have the following main theorem of this section which says that the optimal variable-length source code with the general cost function c defined by (14) can be considered as a variable-length random number generator generating the random sequence subject to the given probability measure q.

Theorem 7: We assume that the entropy rate of the general source $\boldsymbol{X}$ has the limit $\lim_{n\to\infty}(1/n)H(X^n)$. Given an arbitrary probability measure q satisfying (11)–(13), we define the cost function $c : \mathcal{Y}^* \to \mathbf{R}^+$ by (14) and let $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ be any *optimal* variable-length prefix encoder such that

$$\lim_{n\to\infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} = R_v^{\mathrm{c}}(\boldsymbol{X}).$$

Then, we have

$$\lim_{n\to\infty} \frac{1}{n} D(\varphi_n(X^n) \| V_q^{(I_n)} | I_n) = 0$$

where $V_q^{(m)}$ is the random variable subject to the marginal distribution on $\mathcal{Y}^m$ of the probability measure q.

Proof: From the assumption (12) we see that there exists a constant $c_{\min}$ such that

$$0 < c_{\min} \le c(y_i|y_1^{i-1}) \quad (\forall i, \forall y_i \in \mathcal{Y}, \forall y_1^{i-1} \in \mathcal{Y}^{i-1}).$$

Using this property, we can show Theorem 7 in entirely the same manner as in the proof of Theorem 2, provided that $c(y_i)$, $q(y_i)$ are replaced by $c(y_i|y_1^{i-1})$, $q(y_i|y_1^{i-1})$, respectively. □

Example: With a *finite* code alphabet $\mathcal{Y}$ let us consider a Markov process subject to transition probabilities $q(y|y')$ such that $q(y|y') < 1$ $(\forall y, y' \in \mathcal{Y})$. Denoting the initial distribution by $q(y)$, set

$$c(y) = -\log q(y) \qquad c(y|y') = -\log q(y|y')$$

and define the cost $c(\boldsymbol{y})$ of a sequence $\boldsymbol{y} = (y_1, y_2, \cdots, y_n) \in \mathcal{Y}^*$ by

$$c(\boldsymbol{y}) = c(y_1) + c(y_2|y_1) + c(y_3|y_2) + \cdots + c(y_n|y_{n-1}). \tag{15}$$

Then, Theorem 7 tells us that the *optimal* variable-length prefix coding for any general source $\boldsymbol{X}$ with the cost function (15) asymptotically generates the Markov process subject to the transition probabilities $q(y|y')$. □

APPENDIX

Proof of Theorem 1

1) Direct Part: Without loss of generality we may assume that $0 < c(y) < \infty$ $(\forall y \in \mathcal{Y})$. Let $\mathcal{Y} \equiv \{1, 2, \cdots, K\}$ and set $q(i) \equiv K^{-\alpha_c c(i)}$ $(i = 1, 2, \cdots, K)$. For any $\boldsymbol{y} = (y_1, y_2, \cdots, y_l) \in \mathcal{Y}^*$, we define

$$\alpha(\boldsymbol{y}) = \sum_{\boldsymbol{y}' : \, \boldsymbol{y}' \prec \boldsymbol{y}} q(\boldsymbol{y}')$$
$$\beta(\boldsymbol{y}) = \sum_{\boldsymbol{y}' : \, \boldsymbol{y}' \preceq \boldsymbol{y}} q(\boldsymbol{y}') \equiv \alpha(\boldsymbol{y}) + q(\boldsymbol{y})$$

where $\prec$, $\preceq$ indicate the lexicographic order on the set $\mathcal{Y}^l$ and we have put for $\boldsymbol{z} = (z_1, z_2, \cdots, z_l) \in \mathcal{Y}^l$

$$q(\boldsymbol{z}) = q(z_1)q(z_2)\cdots q(z_l).$$

Let the interval $[\alpha(\boldsymbol{y}), \beta(\boldsymbol{y}))$ be denoted by $I(\boldsymbol{y})$. Obviously, $I(\boldsymbol{y}) \subset [0, 1)$ and $|I(\boldsymbol{y})| = K^{-\alpha_c c(\boldsymbol{y})}$ (the width of $I(\boldsymbol{y})$). Then, we first have the following trivial lemma.

Lemma 1: A code $\mathcal{C} \equiv \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots\}$ $(\boldsymbol{y}_i \in \mathcal{Y}^*)$ is prefix if and only if all intervals $I(\boldsymbol{y}_1), I(\boldsymbol{y}_2), \cdots \subset [0, 1)$ are mutually disjoint. □

Let all the elements of $\mathcal{X}^n$ be ordered as $\mathcal{X}^n \equiv \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ and define

$$P_i \equiv \sum_{j=1}^{i-1} P_{X^n}(\boldsymbol{x}_j) \qquad (i = 1, 2, \cdots)$$
$$Q_i \equiv P_i + \tfrac{1}{2}\, P_{X^n}(\boldsymbol{x}_i) \qquad (i = 1, 2, \cdots)$$

where $P_1 \equiv 0$. Now, to each $\boldsymbol{x}_i$ we uniquely assign $\boldsymbol{y}_i$ as

$$\boldsymbol{y}_i \equiv \arg \min_{\boldsymbol{y} \in \mathcal{K}(\boldsymbol{y})} |\boldsymbol{y}|$$

where $\mathcal{K}(\boldsymbol{y})$ is the set of $\boldsymbol{y} \in \mathcal{Y}^*$ such that $I(\boldsymbol{y})$ includes Q_i but does not include either P_i or P_{i+1}. It then follows from $I(\boldsymbol{y}_i) \subset [P_i, P_{i+1})$ that each interval $I(\boldsymbol{y}_1), I(\boldsymbol{y}_2), \cdots$ is disjoint. Then, from Lemma 1, the code $\mathcal{C} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots\}$ is prefix. Therefore, we can define the encoder $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ by

$$\varphi_n(\boldsymbol{x}_i) \equiv \boldsymbol{y}_i.$$

Now, set $\overline{\boldsymbol{y}}_i \equiv (y_1, y_2, \cdots, y_{l-1})$ for each sequence $\boldsymbol{y}_i = (y_1, y_2, \cdots, y_{l-1}, y_l)$. Since $I(\boldsymbol{y}_i) \subset I(\overline{\boldsymbol{y}}_i)$ we have $Q_i \in I(\overline{\boldsymbol{y}}_i)$. Moreover, we see from the definition of $I(\boldsymbol{y}_i)$ that $P_i \in I(\overline{\boldsymbol{y}}_i)$ or $P_{i+1} \in I(\overline{\boldsymbol{y}}_i)$. Then, the width $|I(\overline{\boldsymbol{y}}_i)|$ of the interval $I(\overline{\boldsymbol{y}}_i)$ must be larger than $P_{X^n}(\boldsymbol{x}_i)/2$, so that

$$|I(\overline{\boldsymbol{y}}_i)| = K^{-\alpha_c c(\overline{\boldsymbol{y}}_i)} > \frac{P_{X^n}(\boldsymbol{x}_i)}{2}$$

from which it follows that

$$\begin{aligned} c(\boldsymbol{y}_i) &\le c(\overline{\boldsymbol{y}}_i) + c_{\max} \\ &< \frac{-\log P_{X^n}(\boldsymbol{x}_i)}{\alpha_c} + \frac{\log 2}{\alpha_c} + c_{\max} \end{aligned}$$

where $c_{\max} \equiv \max_{y \in \mathcal{Y}} c(y) < \infty$. Then, we have

$$E\{c(\varphi_n(X^n))\} < -\sum_{\boldsymbol{x} \in \mathcal{X}^n} P_{X^n}(\boldsymbol{x}) \, \frac{\log P_{X^n}(\boldsymbol{x})}{\alpha_c} + \frac{\log 2}{\alpha_c} + c_{\max}$$

which concludes that

$$\limsup_{n \to \infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} \le \frac{1}{\alpha_c} \limsup_{n \to \infty} H(X^n). \qquad \square$$

2) Converse Part:

Let $\varphi_n : \mathcal{X}^n \to \mathcal{Y}^*$ be any variable-length prefix encoder and put

$$c_i \equiv c(\varphi_n(\boldsymbol{x}_i)) \qquad (i = 1, 2, \cdots)$$

and define $q_i \equiv K^{-\alpha_c c_i}$. Then, from Lemma 1, we have

$$q \equiv \sum_{i=1}^{\infty} q_i \le 1.$$

Set $p_i \equiv P_{X^n}(\boldsymbol{x}_i)$ and $p \equiv \sum_{i=1}^{\infty} p_i = 1$. From the log-sum inequality [9], we have

$$\begin{aligned} \sum_{i=1}^{\infty} p_i \log \frac{p_i}{q_i} &\ge p \log \frac{p}{q} \\ &= \log \frac{1}{q} \\ &\ge 0. \end{aligned} \tag{16}$$

On the other hand,

$$\begin{aligned} \sum_{i=1}^{\infty} p_i \log \frac{p_i}{q_i} &= -\sum_{i=1}^{\infty} p_i \log q_i + \sum_{i=1}^{\infty} p_i \log p_i \\ &= \alpha_c \sum_{i=1}^{\infty} p_i c_i + \sum_{i=1}^{\infty} p_i \log p_i \\ &= \alpha_c E\{c(\varphi_n(X^n))\} - H(X^n) \end{aligned}$$

which together with (16) implies that

$$E\{c(\varphi_n(X^n))\} \ge \frac{1}{\alpha_c} H(X^n).$$

Then, we conclude that

$$\limsup_{n \to \infty} \frac{1}{n} E\{c(\varphi_n(X^n))\} \ge \frac{1}{\alpha_c} \limsup_{n \to \infty} H(X^n). \qquad \square$$

REFERENCES

[1] J. von Neumann, “Various techniques used in connection with random digits,” in *Applied Mathematics Series*. Notes by G. E. Forstyle, Wshington, DC: Nat. Bur. Stand., 1951, vol. 12, pp. 36–38.

[2] P. Elias, “The efficient construction of an unbiased random sequences,” in *Ann. Math. Statist.*, 1972, vol. 43, pp. 865–870.

[3] S. Vembu and S. Verdú, “Generating random bits from an arbitrary source: Fundamental limits,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 1322–1332, Sept. 1995.

[4] K. Visweswariah, S. R. Kulkarni, and S. Verdú, “Source codes as random number generators,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 462–471, Mar. 1998.

[5] T. S. Han, *Information-Spectrum Methods in Information Theory* (in Japanese). Tokyo: Baifukan, 1998.

[6] R. M. Karp, “Minimum redundancy coding for the discrete noiseless channel,” *IRE Trans. Inform. Theory*, vol. IT-7, pp. 27–38, Jan. 1961.

[7] K. Iwata, M. Morii, and T. Uyematsu, “An efficient universal coding algorithm for noiseless channel with symbols of unequal cost,” *IEICE Trans. Fundamentals,* vol. E80-A, no. 11, pp. 2232–2237, Nov. 1997.

[8] R. M. Krause, “Channels which transmit letters of unequal duration,” *Inform. Control*, vol. 5, pp. 13–24, 1962.

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

[10] T. S. Han, Theorems on the variable-length intrinsic randomness, to be published.

A New Recursive Universal Code of the Positive Integers

Hirosuke Yamamoto, *Member, IEEE*

***Abstract*—A new recursive universal code of the positive integers is proposed, in which any given sequence can be used as a delimiter of codeword while bit “0” is used as a delimiter in known universal codes, e.g., Levenshtein code, Elias ω code, Even–Rodeh code, Stout code, Bentley–Yao code, etc. The codeword length of the proposed code is shorter than $\log_2^* n$ in almost all of sufficiently large positive integers although the known codes are longer than $\log_2^* n$ for any positive integer n.**

***Index Terms*—Elias ω code, log-star function, universal code of positive integers, universal coding.**

I. Introduction

Many researchers have treated the universal coding of the positive integers that satisfy

$$P(n) \geq P(n+1), \qquad \text{for any } n \in \mathcal{N}, \tag{1}$$

where $P(n)$ is a probability distribution on the set of positive integers $\mathcal{N} = \{1, 2, 3, \cdots\}$ [1]–[7]. These codes can be used practically in various adaptive dictionary codes [8]. Besides the practical uses, it is an interesting coding problem to consider how efficiently we can encode the positive integers under the prefix condition.

Let $\log_2^k n$ be the k-fold composition of the function $\log_2 n$ and let $\log_2^* n$ be

$$\log_2^* n = \log_2 n + \log_2^2 n + \cdots + \log_2^{w^*(n)} n \tag{2}$$

where $w^*(n)$ is the largest integer w which satisfies $\log_2^w n \geq 0$. Then, it is shown theoretically that any positive integer n can be represented with $\log_2^* n - \alpha w^*(n)$ bits if $\alpha < \log_2 \log_2 e$ [2], [3].

On the other hand, many researchers, e.g., Levenshtein [2],[1] Elias [4], Bentley–Yao [5], Even–Rodeh [6], Stout [7], etc., have proposed $\log^* n$-type codes with a recursive structure to attain high performance in large n. But, in their codes, codeword length $l(n)$ cannot become shorter than $\log_2^* n$ although it satisfies $l(n) \leq \log_2^* n + w^*(n) + c$ where c is a constant.

In this correspondence, we propose a new $\log^* n$-type code with a recursive structure, which satisfies that

$$l(n) \leq \log_2^* n - \log_2 (1 - 2^{-f}) w_f^*(n) + c_f$$

even in the worst cases and

$$l(n) \leq \log_2^* n - (1 + \log_2 (1 - 2^{-f})) w_f^*(n) + c_f$$

in the best cases. Here, f is a parameter of the code and c_f is a constant which depends on f. $w_f^*(n)$ is a similar function to $w^*(n)$, which satisfies $w_f^*(n) \leq w^*(n)$.

Since the best and worst cases occur at infinitely many n’s, and, roughly speaking, $l(n)$ is distributed uniformly between two extreme cases, $l(n)$ can become shorter than $\log_2^* n$ in large parts of integers.

In Section II, we review Elias ω code, which is a typical one of the known $\log^* n$-type codes, and we show the reason why the codeword length cannot become shorter than $\log_2^* n$ in the known codes. To overcome this defect, we devise a new representation of binary numbers that never has a given sequence as a prefix. In Section III, we propose a new recursive universal code of the positive integers based on the new binary number representation and we evaluate the performance of the proposed code theoretically. It is shown that the codeword length of the proposed code is shorter than $\log_2^* n$ in almost all of sufficiently large positive integers. The case of r-ary universal codes are treated in Section IV.

We use the following notation in this correspondence.

- $[n]_r$ is the ordinary r-ary number of positive integer n such that the most significant digit of $[n]_r$ is nonzero.
- $[n]_r^i$ is the ordinary r-ary number of n with i digits.
- $\lfloor t \rfloor$ is the largest integer not exceeding t.

Examples: $[14]_2 = 1110$, $[14]_2^5 = 01110$, $[14]_3 = 112$, $[14]_3^5 = 00112$, $\lfloor \log_2 14 \rfloor = 3$.

II. New Binary Number Representation Excluding a Forbidden Prefix

Elias ω code $C_E(n)$ has the following recursive structure [4]:

$$C_E(n_0) = [n_K]_2 [n_{K-1}]_2 \cdots [n_1]_2 [n_0]_2 0 \tag{3}$$

where $[n]_2$ is the ordinary binary number of n, the most significant bit (MSB) of which is always one. Each n_k in (3) is determined recursively by $n_k = \lfloor \log_2 n_{k-1} \rfloor$. In other words, $n_k + 1$ represents the bit length of $[n_{k-1}]_2$. The recursion in (3) stops when the length of $[n_K]_2$ is two. Finally, bit “0” is attached as a delimiter to indicate the end of $C_E(n_0)$.[2] In the decoding, n_K is obtained from the first two bits of $C_E(n_0)$, and the length of $[n_{k-1}]_2$ is recursively obtained from n_k. Since the MSB of every $[n_k]_2$ is “1,” delimiter “0” can stop the recursion and $[n_0]_2$ can easily be found.

Levenshtein W_2 code [2], Even–Rodeh code [6], and Stout code [7] have similar structures and their codes also use bit “0” as a delimiter in the same way as Elias ω code. Levenshtein W_2' code [2] and Bentley–Yao search-tree code [5] have a little different structure. However, it is known that their code can be derived from Elias ω like code by gathering the MSB’s of all $[n_k]_2$ and delimiter “0” as a prefix.

Manuscript received June 4, 1998; revised July 23, 1999. The material in this paper was presented in part at the 1998 IEEE International Symposium on Information Theory, MIT, Cambridge, MA, August 16–21, 1998.

The author is with the Department of Mathematical Engineering and Information Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-8656, Japan (e-mail: yamamoto@hy.t.u-tokyo.ac.jp).

Communicated by N. Merhav, Associate Editor for Source Coding.

Publisher Item Identifier S 0018-9448(00)01667-9.

[1] Levenshtein code is the first $\log^* n$-type code although Elias ω code is famous.

[2] “$n_0 = 1$” is the exception case, for which the codeword is defined as “$C_E(1) = 0$.”