Open access • Journal Article • DOI:10.1109/18.272495

There is no universal source code for an infinite source alphabet — Source link

László Györfi, I. Pali, E.C. van der Meulen

Published on: 01 Jan 1994 - IEEE Transactions on Information Theory (IEEE)

Topics: Universal code, Entropy (information theory), Kullback–Leibler divergence, Source code and Information theory

There is no Universal Source Code for an Infinite Source Alphabet

László Györfi, *Member, IEEE,* István Páli, and
Edward C. van der Meulen, *Fellow, IEEE*

***Abstract*—We show that a discrete infinite distribution with finite entropy cannot be estimated consistently in information divergence. As a corollary we get that there is no universal source code for an infinite source alphabet over the class of all discrete memoryless sources with finite entropy.**

***Index Terms*—Universal source coding, discrete infinite alphabet, distribution estimation, consistency in information divergence.**

I. Introduction

The vast majority of results in information theory pertain to situations where the actual probability law is known. Applying information theory in real life problems there is an obvious question of whether the probability law can be learned from data as far as information theory is concerned. In noiseless source coding, for example, if the source alphabet is finite, then the answer to this question is yes, since there are good universal source coding procedures (see, e.g., [1], [2]). This paper is on discrete infinite source alphabets showing that there is no universal source code over the class of discrete memoryless sources with finite entropy.

Let X be a random variable taking values in $\mathcal{X} = \{1, 2, 3, \cdots\}$ with probability distribution μ and entropy $H(\mu) < \infty$. For a sample $X_1, \cdots, X_n$ from this distribution an estimate of μ is denoted by $\hat{\mu}_n$.

For a discrete memoryless source let f_n be a variable length uniquely decodable code with source block length n. Let $l_n(x_1, x_2, \cdots, x_n)$ denote the length of the codeword $f_n(x_1, x_2, \cdots, x_n)$ and let the average codeword length of f_n be denoted by $\bar{l}_n$. The redundancy per letter of f_n is given by

$$R_n = \frac{1}{n}(\bar{l}_n - H(X_1, \cdots, X_n)).$$

For a uniquely decodable f_n we have by the noiseless source coding theorem (which is also valid for a source with countably infinite source alphabet and finite entropy, see [3, Problem 3.7, p. 514] and [4]) that $R_n \geq 0$.

For two probability distributions $P = \{p_i\}$ and $Q = \{q_i\}$ over $\mathcal{X}$ the information divergence is defined as

$$I(P, Q) = \sum_{i=1}^{\infty} p_i \log \frac{p_i}{q_i}.$$

There is a well-known duality between universal coding and distribution estimation consistent in information divergence: there is a universal source code over a subset of the set of all discrete memoryless sources with finite entropy if and only if there is a distribution estimate consistent in expected information divergence for all sources within this subset. Concerning the aim of this paper the important direction of this equivalence is as follows:

Theorem 1: Given a discrete memoryless source with source alphabet $\mathcal{X} = \{1, 2, 3, \cdots\}$ and source distribution μ, we can construct for any uniquely decodable code f_n a distribution estimate $\hat{\mu}_n$ of μ such that its redundancy satisfies

$$R_n \geq E\{I(\mu, \hat{\mu}_n)\}. \tag{1}$$

Proof: Let

$$A_n = \sum_{\boldsymbol{x} \in \mathcal{X}^n} 2^{-l_n(\boldsymbol{x})}.$$

Then, since a uniquely decodable code for an infinite source alphabet satisfies the Kraft inequality (see [5, Corollary to Theorem 5.5.1, p. 92]), we have $0 < A_n \leq 1$. For $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ let

$$\eta_{n,n}(x_1, x_2, \cdots, x_n) = \eta_{n,n}(\boldsymbol{x}) = \frac{2^{-l_n(\boldsymbol{x})}}{A_n}.$$

Then $\eta_{n,n}$ is a probability distribution on $\mathcal{X}^n$. For $i = n-1, n-2, \cdots, 2, 1$ define recursively

$$\eta_{n,i}(x_1, \cdots, x_i) = \sum_x \eta_{n,i+1}(x_1, \cdots, x_i, x),$$

so that $\eta_{n,i}$ is a probability distribution on $\mathcal{X}^i$. Define

$$g_{n,0}(x) = \eta_{n,1}(x)$$

and for $1 \leq i \leq n-1$ let

$$g_{n,i}(x; x_1, x_2, \cdots, x_i) = \frac{\eta_{n,i+1}(x_1, x_2, \cdots, x_i, x)}{\eta_{n,i}(x_1, x_2, \cdots, x_i)}.$$

We have

$$\prod_{k=0}^{n-1} g_{n,k}(x_{k+1}; x_1, x_2, \cdots, x_k) = \eta_{n,n}(x_1, x_2, \cdots, x_n).$$

For $0 \leq i \leq n-1$ and $x \in \mathcal{X}$ define

$$\hat{\mu}_{n,i}(\{x\}) = g_{n,i}(x; X_1, X_2, \cdots, X_i),$$

and also define

$$\hat{\mu}_n(\{x\}) = \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}_{n,k-1}(\{x\}).$$

Then $\hat{\mu}_{n,i}$ and $\hat{\mu}_n$ are random distributions on $\mathcal{X}$. The redundancy of f_n is bounded below by

$$\begin{aligned}
R_n &= \frac{1}{n}\sum_{\boldsymbol{x} \in \mathcal{X}^n} \mu^n(\boldsymbol{x}) l_n(\boldsymbol{x}) + \frac{1}{n}\sum_{\boldsymbol{x} \in \mathcal{X}^n} \mu^n(\boldsymbol{x}) \log \mu^n(\boldsymbol{x}) \\
&= \frac{1}{n}\sum_{\boldsymbol{x} \in \mathcal{X}^n} \mu^n(\boldsymbol{x}) \left(-\log(A_n) + \log \frac{1}{\eta_{n,n}(\boldsymbol{x})}\right) \\
&\quad + \frac{1}{n}\sum_{\boldsymbol{x} \in \mathcal{X}^n} \mu^n(\boldsymbol{x}) \log \mu^n(\boldsymbol{x}) \\
&\geq \frac{1}{n}\sum_{\boldsymbol{x} \in \mathcal{X}^n} \mu^n(\boldsymbol{x}) \log \frac{\mu^n(\boldsymbol{x})}{\eta_{n,n}(\boldsymbol{x})} \\
&= \frac{1}{n} I(\mu^n, \eta_{n,n}) \\
&= \frac{1}{n}\sum_{x_1, \cdots, x_n} \mu^n(x_1, \cdots, x_n) \log\left(\frac{\mu^n(x_1, \cdots, x_n)}{\eta_{n,n}(x_1, \cdots, x_n)}\right)
\end{aligned}$$

Manuscript received January 22, 1993; revised May 24, 1993. L. Györfi and E. C. van der Meulen were supported by the Scientific Exchange Program between the Hungarian Academy of Sciences and the Royal Belgian Academy of Sciences. L. Györfi was supported by OTKA under Grant T 4360. This paper was presented in part at the IEEE ISIT'93, San Antonio, TX, January 17–22, 1993.

L. Györfi and I. Páli are with the Department of Mathematics, Technical University of Budapest, Stoczek u. 2, H-1521 Budapest, Hungary.

E. C. van der Meulen is with the Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200 B, B-3001 Heverlee, Belgium.

IEEE Log Number 9215188.

$$\begin{aligned}
&= \frac{1}{n}\boldsymbol{E}\left\{\log\left(\frac{\mu^n(X_1, X_2, \cdots, X_n)}{\eta_{n,n}(X_1, X_2, \cdots, X_n)}\right)\right\}\\
&= \frac{1}{n}\boldsymbol{E}\left\{\log\left(\prod_{k=1}^{n}\frac{\mu(X_k)}{g_{n,k-1}(X_k; X_1, \cdots, X_{k-1})}\right)\right\}\\
&= \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{E}\left\{\log\left(\frac{\mu(X_k)}{\hat{\mu}_{n,k-1}(X_k)}\right)\right\}\\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{x_1,\cdots,x_{k-1}}\mu^{k-1}(x_1, \cdots, x_{k-1}).\\
&\quad\cdot\sum_{x_k}\mu(x_k)\log\frac{\mu(x_k)}{g_{n,k-1}(x_k; x_1, \cdots, x_{k-1})}\\
&= \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{E}\{I(\mu, \hat{\mu}_{n,k-1})\}\\
&= \boldsymbol{E}\left\{\sum_{k=1}^{n}\frac{1}{n}I(\mu, \hat{\mu}_{n,k-1})\right\}\\
&\geq \boldsymbol{E}\{I(\mu, \hat{\mu}_n)\},
\end{aligned}$$

where the last inequality follows from the convexity of I in the pair $(\mu, \hat{\mu}_{n,k-1})$ (cf. [5, Theorem 2.7.2, p. 30]). □

Remark 1: It follows from Theorem 1 that, if there is a universal uniquely decodable source code over a subset of the set of all discrete memoryless sources with finite entropy, then there is a distribution estimate consistent in expected information divergence for all sources in this subset. Namely, given the source probability distribution μ and a uniquely decodable code f_n, let $\hat{\mu}_n$ be as constructed in Theorem 1. Since $R_n \to 0$ for a weakly universal code for all sources in this subset (cf. (19) below), it follows from Theorem 1 that

$$\lim_{n\to\infty}\boldsymbol{E}\{I(\mu, \hat{\mu}_n)\} = 0$$

for all μ in this class.

II. One Cannot Estimate a Discrete Infinite Distribution Consistently in Information Divergence

Theorem 2: If, for a discrete memoryless source with source alphabet $\mathcal{X} = \{1, 2, 3, \cdots\}$, $\{\hat{\mu}_n\}$ is an arbitrary sequence of estimates of the unknown source distribution μ, then there is a μ with $H(\mu) < \infty$ such that for all $n \geq 1$ we have

$$I(\mu, \hat{\mu}_n) = \infty \quad \text{a.s.} \tag{2}$$

Proof: The proof consists of three major parts. First we describe a fairly small parametric family of distributions within which one can get (2) (Step 1). Next we investigate the properties of an arbitrary estimator for a certain member of this family (Step 2). The proof is completed by a randomization of the underlying parameter (Step 3).

Step 1) The parametric class $\mathcal{C} = \{\mu_{\boldsymbol{c}}\colon \boldsymbol{c} \in P\}$ of discrete distributions is defined as follows. Put

$$D_k = \{m \in \mathcal{X}\colon 2^k \leq m < 2^{k+1}\}, \quad k = 1, 2, \cdots$$

Then the parameter set P is the set of all sequences $\boldsymbol{c}$ such that the k − th element c_k of $\boldsymbol{c}$ belongs to D_k, i.e.,

$$P = \{\boldsymbol{c}\colon \boldsymbol{c} = c_1, c_2, \cdots; c_k \in D_k, k = 1, 2, \cdots\}.$$

Now choose the fixed and known probability distribution $\{q_k\}$, given by

$$q_k = \begin{cases} C^*\frac{\ln k}{k^2} & \text{if } k = 2, 3, \cdots\\ 1/2 & \text{if } k = 1,\end{cases}$$

where

$$C^* = \frac{1}{2\sum_{i=2}^{\infty}\frac{\ln i}{i^2}}.$$

Then

$$-\sum_k q_k\log q_k < \infty,$$

and

$$\sum_{k=1}^{\infty}kq_k = \infty. \tag{3}$$

Let Y be a random variable such that

$$\Pr\{Y = 2^k\} = q_k \tag{4}$$

and let, for $\boldsymbol{c} \in P$, X be defined by

$$X = F(Y, \boldsymbol{c}) = \sum_k c_k I_{\{Y=2^k\}}$$

where I_A denotes the indicator random variable of an event A. Then, for $\boldsymbol{c} \in P$ the distribution $\mu_{\boldsymbol{c}}$ of X is as follows:

$$\mu(\{j\}) = \mu_{\boldsymbol{c}}(\{j\}) = \begin{cases} q_k & \text{if } j = c_k\\ 0 & \text{otherwise.}\end{cases}$$

Hence, for all $\mu_{\boldsymbol{c}}$

$$H(\mu_{\boldsymbol{c}}) = -\sum_{k=1}^{\infty}q_k\log q_k < \infty. \tag{5}$$

(But note that, under $\mu_{\boldsymbol{c}}$,

$$\boldsymbol{E}\{\log X\} \geq \sum_{k=1}^{\infty}kq_k = \infty.) \tag{6}$$

Step 2) For all $\boldsymbol{c}$ the sets $\{\mu_{\boldsymbol{c}}(\{j\})\}$ and $\{q_k\}$ are identical, the only unknowns are the locations c_k of positive mass q_k within D_k ($k = 1, 2, \cdots$). Now let $Y_1, \cdots, Y_n$ be i.i.d. according to (4) and let $X_1, \cdots, X_n$ be defined by

$$X_i = F(Y_i, \boldsymbol{c}),$$

$\boldsymbol{c} \in P$, $\boldsymbol{c}$ fixed. If there is at least one X_i from $X_1, \cdots, X_n$ falling into D_k then we know the location c_k. We assume without loss of generality that for $\hat{\mu}_n$ applied to $X_1, \cdots, X_n$

$$\hat{\mu}_n(D_k) = q_k = \mu_{\boldsymbol{c}}(D_k), \qquad n = 1, 2, \cdots, \quad \text{and} \quad k = 1, 2, \cdots. \tag{7}$$

Otherwise if there is a k with $\hat{\mu}_n(D_k) = 0$ then $I(\mu, \hat{\mu}_n) = \infty$ and any other estimator can be considered as an improvement of $\hat{\mu}_n$; and if for all k we have $\hat{\mu}_n(D_k) > 0$ then we make a partial normalization

$$\tilde{\mu}_n(\{j\}) = \hat{\mu}_n(\{j\})\frac{q_k}{\hat{\mu}_n(D_k)} \quad \text{if} \quad j \in D_k$$

and would get by Lemma 1 (see Appendix) the improvement

$$I(\mu, \hat{\mu}_n) \geq I(\mu, \tilde{\mu}_n).$$

Moreover, if μ_n denotes the standard empirical measure based on $X_1, \cdots, X_n$, then, if $\mu_n(D_k) > 0$ (which means that there is an X_i falling into D_k), we have

$$\hat{\mu}_n(\{i\}) = \mu_{\boldsymbol{c}}(\{i\}) = q_k I_{\{i=c_k\}} \quad \text{if} \quad i \in D_k. \tag{8}$$

Next, if we introduce

$$\lambda_n(m) = \frac{\hat{\mu}_n(\{m\})}{q_k} \quad \text{if} \quad m \in D_k, \tag{9}$$

and

$$k_n = \log\left(\max_{1\leq i\leq n}Y_i\right) + 1,$$

then for $k \geq k_n$ we have $\mu_n(D_k) = 0$. Assumption (7) implies that

$$\lambda_n(D_k) = 1, \tag{10}$$

i.e., λ_n is a discrete probability measure on D_k. From (8) and (9) it follows that

$$\begin{aligned} I(\mu_{\boldsymbol{c}}, \hat{\mu}_n) &= \sum_{k:\mu_n(D_k)=0} \sum_{m\in D_k} \mu_{\boldsymbol{c}}(\{m\}) \log \frac{\mu_{\boldsymbol{c}}(\{m\})}{\hat{\mu}_n(\{m\})} \\ &= \sum_{k:\mu_n(D_k)=0} \sum_{m\in D_k} \mu_{\boldsymbol{c}}(\{m\}) \log \frac{1}{\lambda_n(m)} \\ &\geq \sum_{k=k_n}^{\infty} \sum_{m\in D_k} \mu_{\boldsymbol{c}}(\{m\}) \log \frac{1}{\lambda_n(m)} \triangleq G_n(\boldsymbol{c}, \lambda_n, k_n) \end{aligned} \tag{11}$$

where in the last step we used that because of (10)

$$\sum_{m\in D_k} \mu_{\boldsymbol{c}}(\{m\}) \log \frac{1}{\lambda_n(m)} \geq 0.$$

Step 3) In order to prove (2) we shall show that there is a $\boldsymbol{c}$ such that

$$G_n(\boldsymbol{c}, \lambda_n, k_n) = \infty \qquad \text{for all} n \qquad \text{a.s.} \tag{12}$$

For this purpose we use a randomization: let $\boldsymbol{C}$ be uniformly distributed on P, or equivalently, let $C_1, C_2, \cdots$ be independent random variables such that C_k is uniformly distributed on D_k. Also let $Y_1, Y_2, \cdots$ be i.i.d. according to (4), independent of $\boldsymbol{C}$, and put

$$\tilde{X}_i = F(Y_i, \boldsymbol{C}) \quad i = 1, 2, \cdots.$$

Note that the conditional distribution of $\tilde{X}_i$ given $\boldsymbol{C} = \boldsymbol{c}$ is $\mu_{\boldsymbol{c}}$. Let $\mathcal{F}_k$ denote the σ-algebra generated by $C_1, \cdots, C_k, Y_1, Y_2, \cdots$. The following facts can be easily verified.

i) For $m \in D_k$, $\mu_{\boldsymbol{C}}(\{m\})$ is a function of C_k.

ii) On the event $\{k_n = K\}$, $\lambda_n(m)$ is measurable on $\mathcal{F}_K$ for $m \in D_k$ and $k \geq K$.

iii) The random variables $\{\boldsymbol{I}_{\{(k_n=K)\}}, \mu_{\boldsymbol{C}}(\{m\});\ m \in D_k\}, k = K, K+1, \cdots$ are conditionally independent given $\mathcal{F}_K$. Moreover

$$\begin{aligned} \sum_{m\in D_k} \mu_{\boldsymbol{C}}(\{m\}) \log \frac{1}{\lambda_n(m)} &= \sum_{m\in D_k} q_k \boldsymbol{I}_{\{m=C_k\}} \log \frac{1}{\lambda_n(m)} \\ &= q_k \log \frac{1}{\lambda_n(C_k)}. \end{aligned} \tag{13}$$

In order to prove (12) we shall show that for all n

$$G_n(\boldsymbol{C}, \lambda_n, k_n) = \infty \quad \text{a.s.}$$

or, equivalently, that for all n and all $K \geq 1$,

$$G_n(\boldsymbol{C}, \lambda_n, K) = \infty \quad \text{a.s.}$$

Let

$$M_{n,k} = \left\{ m \colon m \in D_k \text{ and } \lambda_n(m) < 2\frac{1}{2^k} \right\},$$

then we have by (10)

$$|\, M_{n,k} \,| \geq 2^{k-1},$$

and since $\{k_n = K\}$ is $\mathcal{F}_k$-measurable for all k, we have for any value of K that

$$\Pr\{k_n = K, C_k \in M_{n,k} \mid \mathcal{F}_K\} \geq \frac{1}{2}\boldsymbol{I}_{\{k_n=K\}} \qquad \text{for} \quad k \geq K. \tag{14}$$

From (11) and (13) we have

$$\begin{aligned} G_n(\boldsymbol{C}, \lambda_n, K) &= \boldsymbol{I}_{\{k_n=K\}} \sum_{k=k_n}^{\infty} q_k \log \frac{1}{\lambda_n(C_k)} \\ &\geq \sum_{k=K}^{\infty} \boldsymbol{I}_{\{k_n=K\}} \boldsymbol{I}_{\{C_k \in M_{n,k}\}} q_k \log \frac{1}{\lambda_n(C_k)} \\ &> \sum_{k=K}^{\infty} \boldsymbol{I}_{\{k_n=K\}} \boldsymbol{I}_{\{C_k \in M_{n,k}\}} q_k \log 2^{k-1} \\ &\geq \sum_{k=K}^{\infty} \boldsymbol{I}_{\{k_n=K\}} \boldsymbol{I}_{\{C_k \in M_{n,k}\}} k q_k - 1 \\ &= \sum_{k=K}^{\infty} Z(n, k, K) k q_k - 1 \end{aligned} \tag{15}$$

where

$$Z(n, k, K) = \boldsymbol{I}_{\{k_n=K\}} \boldsymbol{I}_{\{C_k \in M_{n,k}\}}.$$

By iii), for fixed n and on the event $\{k_n = K\}$, the random variables $Z(n, k, K)$, $k = K, K+1, \cdots$ are conditionally independent given $\mathcal{F}_K$. By (14)

$$\Pr\{Z(n, k, K) = 1 \mid \mathcal{F}_K\} \geq \frac{1}{2}\boldsymbol{I}_{\{k_n=K\}} \qquad \text{for} \quad k \geq K. \tag{16}$$

Let

$$Z = \sum_{k=K}^{\infty} Z(n, k, K) k q_k.$$

We will prove that on the event $\{k_n = K\}$, $Z = \infty$ a.s. given $\mathcal{F}_K$, which with (15) will complete the proof. This will be done by proving that on the event $\{k_n = K\}$ we have $\boldsymbol{E}\{Z \mid \mathcal{F}_K\} = \infty$ and also $|\, Z - \boldsymbol{E}\{Z \mid \mathcal{F}_K\} \,| < \infty$ a.s. given $\mathcal{F}_K$. By (16) we have

$$\boldsymbol{E}\{Z \cdot \boldsymbol{I}_{\{k_n=K\}} \mid \mathcal{F}_K\} \geq \boldsymbol{I}_{\{k_n=K\}} \sum_{k=K}^{\infty} \frac{k q_k}{2}.$$

Hence, by (3), on the event $\{k_n = K\}$

$$\boldsymbol{E}\{Z \mid \mathcal{F}_K\} = \infty. \tag{17}$$

Let $W_k = (Z(n, k, K) - \boldsymbol{E}\{Z(n, k, K) \mid \mathcal{F}_K\}) k q_k$, $k \geq K$. Then by iii) the W_k's, $k = K, K+1, \cdots$, are conditionally independent given $\mathcal{F}_K$ and we have

$$\boldsymbol{E}\{W_k \mid \mathcal{F}_K\} = 0 \qquad \text{for} \quad k \geq K$$

and by the definition of $Z(n, k, K)$

$$\sum_{k=K}^{\infty} \boldsymbol{E}\{W_k^2 \mid \mathcal{F}_K\} \leq \boldsymbol{I}_{\{k_n=K\}} \sum_{k=K}^{\infty} k^2 q_k^2 < \infty. \tag{18}$$

Now, since the sequence $\{W_k;\ k \geq K\}$ forms a martingale difference sequence given $\mathcal{F}_K$, Corollary 2.8.5 of [6, p. 55] together with

(18) implies that on the event $\{k_n = K\}$

$$|Z - \boldsymbol{E}\{Z \mid \mathcal{F}_K\}| = \left|\sum_{k=K}^{\infty} W_k\right| < \infty \quad \text{a.s. given} \quad \mathcal{F}_K.$$

This with (17) proves that, given $\mathcal{F}_K$, $Z = \infty$ a.s. on the event $\{k_n = K\}$. Since

$$G_n(\boldsymbol{C}, \lambda_n, k_n) = \sum_{k=1}^{\infty} \boldsymbol{I}_{\{k_n = k\}} G_n(\boldsymbol{C}, \lambda_n, k)$$

we have thus shown that $G_n(\boldsymbol{C}, \lambda_n, k_n) = \infty$ a.s. for all n. Hence, by (11) there exists a $\boldsymbol{c}$ such that $I(\mu_{\boldsymbol{c}}, \hat{\mu}_n) \geq G_n(\boldsymbol{c}, \lambda_n, k_n) = \infty$ for any given n. Together with (5) this shows the existence of a μ with finite entropy such that (2) holds for all $n \geq 1$. □

III. There is no Universal Source Code for an Infinitely Discrete Source Alphabet

As in Davisson [7], a sequence of uniquely decodable codes f_1, $f_2, \cdots$ is called weakly universal for a class of sources if

$$\lim_{n \to \infty} R_n = 0 \tag{19}$$

for all sources in this class. The following theorem implies that there is no universal code for the class of all discrete memoryless sources with infinite source alphabet and finite entropy.

Theorem 3: For any sequence of source codes $\{f_n\}$ for an infinitely discrete source alphabet there is a memoryless source with finite entropy such that

$$R_n = \bar{l}_n = \infty \qquad \text{for all} \quad n.$$

Proof: For a fixed n construct $\hat{\mu}_n$ as in Theorem 1. Theorem 2 shows, that for the sequence $\{\hat{\mu}_n\}$ there is a μ with finite entropy such that $I(\mu, \hat{\mu}_n) = \infty$ a.s. for all n, hence $\boldsymbol{E}\{I(\mu, \hat{\mu}_n)\} = \infty$ for all $n \geq 1$. (1) implies that $R_n = \infty$ for all n, and, because $H(\mu) < \infty$, we have $\bar{l}_n = \infty$ for all n. □

Remark 2: In a private communication, J. C. Kieffer suggested that the nonexistence at a weakly universal code over the class mentioned in Theorem 3 was known due to results in [8]. There he proved that a necessary and sufficient condition for the existence of a weakly universal code for a subset $\mathcal{M}$ of the set of all stationary and ergodic sources over a certain countable alphabet is the following:

Condition 1: There exists a countable set $\mathcal{N}$ of distributions on the source alphabet such that for all sources from $\mathcal{M}$ with marginal distribution μ there is a $\nu_\mu \in \mathcal{N}$ such that

$$I(\mu, \nu_\mu) < \infty.$$

He showed that Condition 1 doesn't hold for the family of all stationary and ergodic sources over a countably infinite alphabet with finite entropy. His proof can be carried out also for the class of discrete memoryless sources with countably infinite alphabet and finite entropy. We stated a bit more in two respects. We constructed a smaller set of sources (the set of i.i.d. sources with marginals in $\mathcal{C}$) for which a weakly universal code doesn't exist and we proved not only the nonexistence but the stronger property, that there is a member of this set with $R_n = \infty$ for all n and for any code sequence.

We point out that Condition 1 is equivalent to the following:

Condition 2: There exists a distribution ν such that for all sources from $\mathcal{M}$ with marginal distribution μ

$$I(\mu, \nu) < \infty.$$

(Condition 1 implies Condition 2 since denoting $\mathcal{N} = \{\nu_1, \nu_2, \cdots\}$ and defining $\nu = \sum_{i=1}^{\infty} 2^{-i}\nu_i$ one can prove that $I(\mu, \nu) < \infty$ for all μ in $\mathcal{M}$. The other direction of the equivalence is trivial.)

Remark 3: For all distributions μ of X from the class $\mathcal{C}$ defined in the proof of Theorem 2 we have (cf. (6))

$$\boldsymbol{E}\{\log X\} = \infty.$$

We recall (see [9]) that in the case of a countably discrete random variable

$$\boldsymbol{E}\{\log X\} < \infty \Rightarrow H(\mu) < \infty,$$

so that it may happen that $H(\mu) < \infty$ and $\boldsymbol{E}\{\log X\} = \infty$. Davisson [7, Theorem 8] has shown that if the source alphabet is infinitely discrete, then there is a universal code over a subset of the set of all stationary and ergodic sources with finite entropy and with marginal distribution μ if $H(\mu) < \infty$ and $I(\mu, \nu) < \infty$ for some probability distribution ν, i.e., he proved the sufficiency of Condition 2.

In [10] we generalized this result to show that there exists a weakly universal code over the class of all infinitely discrete memoryless sources with respect to which the expected codeword length of a given uniquely decodable code is finite. This latter condition implies that $H(\mu) < \infty$ and the existence of a ν such that $I(\mu, \nu) < \infty$.

In [10], it is shown how to construct a ν with $I(\mu, \nu) < \infty$ if $\boldsymbol{E}\{\log X\} < \infty$, so that in this case there is no negative result as far as universal source coding is concerned.

Here we have given an example of a class of source distributions for which $H(\mu) < \infty$ but for which there is no universal source code. Since for a source distribution we typically have $\boldsymbol{E}\{\log X\} < \infty$, our counterexample is mainly of theoretical interest.

Appendix

Lemma 1: Let $\{p_{ik}; i \in D_k, k = 1, 2, \cdots\}$ and $\{q_{ik}; i \in D_k, k = 1, 2, \cdots\}$ be probability distributions, and introduce the notations

$$\bar{p}_k = \sum_{j \in D_k} p_{jk} > 0,$$
$$\bar{q}_k = \sum_{j \in D_k} q_{jk} > 0,$$

and

$$\tilde{q}_{ik} = q_{ik}\frac{\bar{p}_k}{\bar{q}_k}.$$

Then

$$I(p, q) \geq I(p, \tilde{q}).$$

Proof:

$$\begin{aligned} I(p, q) &= \sum_k \sum_{j \in D_k} p_{jk} \log\left(\frac{p_{jk}}{q_{jk}}\right) \\ &= \sum_k \sum_{j \in D_k} p_{jk} \log\left(\frac{\bar{p}_k}{\bar{q}_k}\frac{p_{jk}}{\tilde{q}_{jk}}\right) \\ &= \sum_k \bar{p}_k \log\left(\frac{\bar{p}_k}{\bar{q}_k}\right) + \sum_k \sum_{j \in D_k} p_{jk} \log\left(\frac{p_{jk}}{\tilde{q}_{jk}}\right) \\ &\geq I(p, \tilde{q}). \end{aligned}$$

□

Acknowledgment

The authors wish to thank T. Linder for helpful suggestions which led to an improvement of the paper.

References

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

[2] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–664, Sept. 1983.

[3] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[4] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, Mar. 1975.

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

[6] W. F. Stout, *Almost Sure Convergence.* New York: Academic, 1974.

[7] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[8] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.

[9] A. D. Wyner, "An upper bound on the entropy series," *Inform. Contr.*, vol. 20, pp. 176–181, 1972.

[10] L. Györfi, I. Páli, and E. C. van der Meulen, "On universal noiseless source coding for infinite source alphabets," in *European Trans. Telecommun. Related Technol.*, vol. 4, pp. 125–132, Mar.–Apr. 1993.

On the Equivalence of McEliece's and Niederreiter's Public-Key Cryptosystems

Yuan Xing Li, *Member, IEEE,* Robert H. Deng, and Xin Mei Wang, *Member, IEEE*

***Abstract*—It is shown that McEliece's and Niederreiter's public-key cryptosystems are equivalent when set up for corresponding choices of parameters. A security analysis for the two systems, based on this equivalence observation, is presented.**

***Index Terms*—Cryptosystems, McEliece's cryptosystem, Niederreiter's cryptosystem, security, algebraic codes.**

I. Introduction

Since the concept of public-key cryptosystems appeared in the fundamental paper of Diffie and Hellman [1] in 1977, the field of cryptology has undergone a dramatic development. The last decade has seen explosive growth in unclassified research in all aspects of cryptology. Public-key cryptosystem and cryptanalysis have been two of the most active areas. So far many kinds of public-key cryptosystems have been proposed, and many of them that had been thought to be secure have been broken.

A special class of public-key cryptosystems were constructed based on algebraic error-correcting codes. In the present paper, we focus on two such systems, McEliece's [2] and Niederreiter's [3] cryptosystems, examine the relationship between the two, and derive the interesting result that the two systems are equivalent and have the same security when set up for corresponding choices of parameters. This result allows us to clarify the security evaluations of Niederreiter's cryptosystem, by Niederreiter [3] and by Brickell and Odlyzko [4]. Furthermore, we employ the best known attack, the Lee–Brickell attack [5], to cryptanalyze the two systems. Some new optimal parameter values and work factors are obtained.

We briefly review McEliece's and Niederreiter's cryptosystems in Section II. The equivalence of the two systems is derived in Section III. Finally, we cryptanalyze the two systems and comment on the selection of optimal system parameters in Section IV.

II. McEliece's and Niederreiter's Cryptosystems

In this section, brief descriptions of McEliece's and Niederreiter's cryptosystems are presented in order to facilitate discussions in later sections. Both cryptosystems are algebraic-coded two-key systems. The basic idea behind them was to construct a linear error-correcting code for which a fast decoding algorithm is known, and then to disguise it as a general linear code whose decoding problem is NP-complete.

A. McEliece's Cryptosystem [2]

This system uses a binary $(n,\ k,\ 2t+1)$ Goppa code C where n is the code length, k is the code dimension, and t is the error-correcting capability of C. C is constructed by randomly selecting an irreducible polynomial of degree t over GF(2^l) as the Goppa polynomial (note that $n = 2^l$). Let $\boldsymbol{G}$ be a $k \times n$ generator matrix of C [6], $\boldsymbol{S}$ any $k \times k$ nonsingular matrix, and $\boldsymbol{P}$ any $n \times n$ permutation matrix.

- Private Key: $\boldsymbol{G}$, $\boldsymbol{S}$, $\boldsymbol{P}$.
- Public Key: $\boldsymbol{G}' = \boldsymbol{SGP}$ and t.
- Messages: k bit vectors $\boldsymbol{m}$ over GF(2).
- Encryption: $\boldsymbol{c} = \boldsymbol{mG}' + \boldsymbol{e}$. $\boldsymbol{e}$, an n bit error vector with (Hamming) weight t. $\boldsymbol{c}$, the n bit ciphertext.
- Decryption: Since $\boldsymbol{c} = \boldsymbol{mSGP} + \boldsymbol{e}, \boldsymbol{cP}^{-1} = (\boldsymbol{mS})\boldsymbol{G} + \boldsymbol{eP}^{-1}$. Use a fast decoding algorithm for C to correct the error "$\boldsymbol{eP}^{-1}$, find $\boldsymbol{mS}$ and thus $\boldsymbol{m}$.

McEliece investigated several attacks against his system. One of those was to factor the public key to obtain the private key, but this approach was thought to be hopeless. Another attack, considered as the most promising, was to pick k "error-free" elements of the ciphertext $\boldsymbol{c}$, and then to solve a set of $\boldsymbol{k}$ linear equations to recover the message $\boldsymbol{m}$. Using this attack, McEliece suggested using $n = 1024$ and $t = 50$, i.e., (1024, 524, 101) Goppa code in his system. The corresponding work factor of the system is approximately $2^{80.7}$ [7].

B. Nederreiter's Cryptosystem [3]

This is a knapsack-type cryptosystem which employs an $(n,\ k,\ 2t+1)$ linear code C over GF(q). Let $\boldsymbol{H}$ be an $(n-k) \times n$ parity check matrix of C, $\boldsymbol{M}$ any $(n-k) \times (n-k)$ nonsingular matrix, and $\boldsymbol{P}$ any $n \times n$ permutation matrix, all over GF(q).

- Private Key: $\boldsymbol{H}$, $\boldsymbol{M}$, and $\boldsymbol{P}$.
- Public Key: $\boldsymbol{H}' = \boldsymbol{MHP}$ and t.
- Messages: n dimensional vectors $\boldsymbol{y}$ over GF(q) with weight t.
- Encryption: $\boldsymbol{z} = \boldsymbol{yH}'^{\boldsymbol{T}}$. $\boldsymbol{z}$, the ciphertext of dimension $n-k$.
- Decryption: Since $\boldsymbol{z} = \boldsymbol{y}(\boldsymbol{MHP})^{\boldsymbol{T}}, \boldsymbol{z}(\boldsymbol{M}^{\boldsymbol{T}})^{-1} = (\boldsymbol{yP}^{T})\boldsymbol{H}^{\boldsymbol{T}}$. Use a fast decoding algorithm for C to find $\boldsymbol{yP}^{\boldsymbol{T}}$ and thus $\boldsymbol{y}$.

Niederreiter [3] cryptanalyzed his system and mentioned two example systems, one using a binary concatenated (104, 24, 31) code and the other using a (30, 12, 19) Reed–Solomon code over GF(31). The examples were later verified as insecure by Brickell and Odlyzko

Manuscript received May 4, 1993; revised November 17, 1993. This work was supported in part by the National Natural Science Foundation of China under Grant 69072910.

Y. X. Li is with the Magnetic Technology Centre, National University of Singapore, Kent Ridge, Singapore 0511, Republic of Singapore.

R. H. Deng is with the Department of Electrical Engineering, National University of Singapore, Kent Ridge, Singapore 0511, Republic of Singapore.

X. M. Wang is with the Department of Information Engineering, Xidian University, People's Republic of China.

IEEE Log Number 9215123.