

The prediction is represented as a set of probability estimates of possible continuations of the stochastic process. The prediction problem is solved in two settings: 1) given that the sequence is computable or 2) given that the sequence is stationary. In mathematical terms, the problem is related to coding theory and its solution accordingly relies on known information-theoretical results.\*

### 1. Introduction

Consider the prediction of the stochastic process  $\xi(t)$ ,  $t = 0, 1, \dots$ , where the random variables  $\xi(t)$  take values in some finite set (alphabet)  $A$ , without assuming that a metric is defined on  $A$ . One of the prediction problems of this kind goes back to Laplace [1, 2]: given that some event  $\alpha$  occurred in each of  $N$  ( $N \geq 1$ ) trials, predict the outcome of the  $(N + 1)$ -th trial. Laplace suggested the following example: it is known that the sun rose during 5000 years ( $\approx 1,826,213$  days); use this information to predict if the sun will or will not rise tomorrow [1]. Here, the set  $A$  consists of two elements, "will rise" and "will not rise" and no metric is defined on  $A$  (no statement predicting an intermediate value is admissible). Other prediction problems of this kind include forecasting droughts and similar climatic anomalies, predicting infectious epidemics, etc. Laplace represented the prediction by two numbers: the probabilities of occurrence and nonoccurrence of the event  $\alpha$  in the  $(N + 1)$ -th trial. In this study, we similarly represent the prediction as a set of probabilities of possible continuations of an observed random sequence.

The main result is the construction of an asymptotically optimal prediction method for stationary processes. Informally, the prediction is constructed in two stages: first, using the given data, we estimate the "memory" of the observed stochastic process, i.e., determine the order or the "connectedness" of the Markov chain that provides the best approximation to the observed series. Then we estimate the probability characteristics of the Markov chain and use them to construct the prediction. Let us consider two examples of predictions constructed by this method. A stationary process takes values in the set  $\{0, 1\}$  and a series of four observations is given, "0 0 0 0." It is required to predict the fifth symbol in the series. The corresponding prediction is the following: the probability of getting a zero is 0.85, the probability of getting a one is 0.15. (This example is directly related to Laplace's sunrise problem. In general, the probability of getting a zero in place  $N + 1$  in a sequence of  $N$  zeros is close to  $1 - 1/N$ .) If the observed sequence is "0 1 0 1" then the proposed prediction method will "notice" the periodicity and predict that the fifth symbol is 0 with probability  $\approx 2/3$  and 1 with probability  $\approx 1/3$ .

In addition to prediction of stationary sequences, we also consider the prediction of constructive stochastic processes, introduced in [2]. The solution of both problems relies on known information theoretical constructions: for stationary processes, it relies on the universal code of [3], and for constructive processes on the measure from [2].

Let us formalize our problem. We denote by  $\mathbb{N}$  the set of nonnegative integers, by  $A^n$ ,  $n \in \mathbb{N}$ , the set of words of length  $n$  in the alphabet  $A$ ,  $A^0 = \emptyset$ . Let  $A^* = \bigcup_{n=1}^{\infty} A^n$ ,  $\Omega(A)$  be the set of all one-sided infinite words on  $A$ . For each  $n \in \mathbb{N}$  and  $\omega \in \Omega(A)$  define  $x_n(\omega)$  as the  $n$ -th coordinate of the word  $\omega$ , i.e., if  $\omega = \omega_0 \omega_1 \dots \omega_n \dots$ , then  $x_n(\omega) = \omega_n$ . For words of the form  $x_n(\omega) \dots x_{n+k}(\omega)$  we use the abbreviation  $x_n x_{n+1} \dots x_{n+k}(\omega)$ . For  $y \in A^*$ , denote by  $\ell(y)$  the set of words from  $\Omega(A)$  starting with  $y$ , i.e., such that  $x_0 x_1 \dots x_{k-1}(\omega) = y$ , where  $k = |y|$ .

\* The main findings were reported at the 12th All-Union School on Information Theory, February, 1985.

Translated from Problemy Peredachi Informatsii, Vol. 24, No. 2, pp. 3-14, April-June, 1988. Original article submitted January 27, 1986.

(Here and below,  $|y|$  is the length of  $y$  if  $y$  is a word and the cardinality of  $y$  if  $y$  is a set.) Below we consider probability measures on  $\Omega(A)$  defined on the Borel  $\sigma$ -algebra generated by the sets  $\mathcal{L}(y)$ ,  $y \in A^*$ . (Sometimes these measures will be called stochastic processes or sources generating the elements of  $A$ .) For the function  $f$  defined on  $\Omega(A)$ , we denote by  $E_\mu(f)$  the mean over the measure  $\mu$ ; by definition,  $\mu(u) = \mu(\mathcal{L}(u))$  for  $u \in A^*$ .

Given are a family of measures  $M$  and a realization of the stochastic process  $\mu x_0 x_1 \dots x_{t-1}$  ( $\omega$ ), where  $\mu \in M$  and  $\mu$  is not known exactly. It is required to make a prediction of the process  $n \in \mathbb{N}$  steps into the future. If the measure  $\mu$  were known, then all the information about the future  $n$  steps would be contained in the set of conditional probabilities

$$\{\mu(x_t x_{t+1} \dots x_{t+n-1}(\omega) = v | x_0 x_1 \dots x_{t-1}(\omega)); v \in A^n\}. \quad (1)$$

But the measure  $\mu$  is not known, and we only know that  $\mu \in M$ , and therefore instead of the true probabilities (1) we can only obtain their estimates. The prediction problem is concerned with obtaining these probability estimates.

We denote by

$$\{\mu^*(x_t x_{t+1} \dots x_{t+n-1}(\omega) = v | x_0 x_1 \dots x_{t-1}(\omega)); v \in A^n\} \quad (2)$$

the set of estimates of the probabilities (1) and assume that fairly natural conditions are satisfied: first, all the values in (2) are nonnegative and sum to 1; second, the  $n$ - and  $(n+k)$ -step predictions should be consistent for all  $n, k, t \in \mathbb{N}$  in the following sense:

$$\sum_{u \in A^k} \mu^*(x_{t+1} \dots x_{t+n} x_{t+n+1} \dots x_{t+n+k}(\omega) = uv | x_0 \dots x_t(\omega)) = \mu^*(x_{t+1} \dots x_{t+n}(\omega) = u | x_0 x_1 \dots x_t(\omega)).$$

It is easy to see that if the estimates  $\mu^*$  are given for all  $n, k, t \in \mathbb{N}$ , then this is equivalent to specifying the probability measure  $\mu^*$  on  $\Omega(A)$ , and (2) is the ordinary conditional probability defined by the equality

$$\mu^*(x_{t+1} x_{t+2} \dots x_{t+n-1}(\omega) | x_0 \dots x_t(\omega)) = \frac{\mu^*(x_0 \dots x_{t+n-1}(\omega))}{\mu^*(x_0 \dots x_t(\omega))}$$

for  $\mu^*(x_0 \dots x_t(\omega)) > 0$ .

The accuracy of the prediction  $\mu^*$  on the measure  $\mu$  and  $\omega \in \Omega(A)$  is estimated by

$$r(\mu, \mu^*, t, n, \omega) = \log \frac{\mu(x_t x_{t+1} \dots x_{t+n-1}(\omega) | x_0 \dots x_{t-1}(\omega))}{\mu^*(x_t x_{t+1} \dots x_{t+n-1}(\omega) | x_0 \dots x_{t-1}(\omega))} \quad (3)$$

(here and below,  $\log x = \log_2 x$ ).  $r(\mu, \mu^*, t, n, \omega)$  is called the prediction error. Note that the mean error on the measure  $\mu$ , given by

$$\bar{r}(\mu, \mu^*, t, n) = E_\mu(r(\mu, \mu^*, t, n, \omega)), \quad (4)$$

is widely known in information theory and mathematical statistics: this is the entropy of a measure by a measure or a Kullback-Leibler deviation. It is always nonnegative and vanishes only when the estimate by the measure  $\mu^*$  is exact for all  $\omega \in \Omega(A)$  (for fixed  $n, t \in \mathbb{N}$ ).

The error  $\bar{r}(\mu, \mu^*, t, n)$  depends only on the size of the class of measures for which a prediction is required. The smaller  $M$  is, then in general more accurate is the prediction. In the limiting case when  $M$  consists of a single measure  $\mu$ , the prediction by the measure  $\mu^* = \mu$  is absolutely accurate. One of the largest classes is the class of computable measures introduced in [2]. An exact definition of a computable measure is given in Sec. 2, and here we only note that for a computable measure there is an algorithm that computes  $\mu(u)$  on any set  $\mathcal{L}(u)$ ,  $u \in A^*$  with any prespecified accuracy. There exists a measure  $\lambda$  which is also constructed in [2], such that for any computable measure  $\mu$  the mean prediction error  $\bar{r}(\mu, \lambda, t, n)$  goes to zero as the length of the observation series  $t$  increases.

Unfortunately, the measure  $\lambda$  is not computable. Moreover, there exists no computable measure having this property. It is thus relevant to consider the prediction problem for other classes of processes, in particular, for the family of stationary measures. For this class, we construct a measure  $\rho$  such that for any stationary process  $\mu$  for almost all  $\omega \in \Omega(A)$  the mean prediction error

$$\frac{1}{T} \sum_{t=0}^{T-1} r(\mu, \rho, t, n, \omega)$$

converges to zero for  $T \rightarrow \infty$  for any  $n \in \mathbb{N}$ . The measure  $\rho$  was previously constructed in [3]

and is based on optimal universal code constructions. (A similar measure was independently constructed in [4].)

The fact that the measures  $\lambda$  and  $\rho$  used respectively for prediction on classes of computable and stationary processes existed in "ready-made" form in information theory suggests that our prediction problem is close to coding theory. The first step toward the construction of the measures  $\lambda$  and  $\rho$  was made in the seminal work [5]. First, an algorithmic approach was proposed in [5] (and independently in [6]) to the notions of information and randomness: the development of this approach in [2] has led to the construction of the measure  $\lambda$ . Second, the first universal code was outlined in [5] and such a code was independently constructed in [7] for the class of Bernoulli sources. Subsequently, optimal universal codes were constructed for Bernoulli [8, 9] and Markov [10, 11] sources, and a code for arbitrary ergodic sources was proposed in [12]. Finally, Ryabko [3] constructed a code that is asymptotically optimal at the same time for all these classes of sources; this code defines the measure  $\rho$ .

Mathematically, the proposed approach to prediction is very close to the coding of a source from given observations [8, 13-15]. This is not surprising, since our problem is part of both universal coding theory and mathematical statistics. The close link between these two disciplines has been noted and utilized on numerous occasions. Thus, a universal code construction similar to the maximum likelihood method in statistics was proposed in [11].

Conversely, in [4, 16], universal coding methods are applied to statistical estimation and choice of a prediction model for an ARMA Gaussian stochastic process.

The approach to prediction developed in this study has not been considered before, although the conceptual link between information and prediction theories has been explicitly noted and exploited previously in [2, 4, 16].

Section 2 and 3 deal with prediction for the class of computable and stationary measures, respectively. Section 4 gives some examples of the application of prediction methods. The Appendix presents some proofs.

## 2. Prediction for the Class of Effective Stochastic Processes

The notion of computable and semicomputable measure was introduced in [2]. By definition, the measure  $\mu$  defined on  $\Omega(A)$  is computable if there exist general recursive functions  $F(y, n)$  and  $G(y, n)$ ,  $y \in A^*$ ,  $n \in \mathbb{N}$ , such that the number  $\alpha_\mu(y, n) = F(y, n)/G(y, n)$  approaches  $\mu(y)$  with accuracy  $2^{-n}$ . (Recall that the function  $f(y, n)$  is called general recursive if there exists an algorithm realized, say, by a Turing machine such that for all  $y \in A^*$ ,  $n \in \mathbb{N}$  its application to the pair  $(y, n)$  produces a number from  $\mathbb{N}$ .) The measure  $\nu$  is called semicomputable if there exist general recursive functions  $F(x, t)$  and  $G(x, t)$  such that the function

$$\beta_\nu(y, t) = F(x, t)/G(x, t)$$

is monotonically nondecreasing in  $t$  and

$$\lim_{t \rightarrow \infty} \beta_\nu(y, t) = \nu(y)$$

for all  $y \in A^*$ ; a semicomputable measure is defined on  $\Omega(A) \cup A^*$ . Any computable measure is obviously also semicomputable.

The universal semicomputable measure  $\lambda$  is constructed in [2]: by definition, this is a measure such that for any semicomputable measure  $\nu$  there is a constant  $c_\nu$  satisfying

$$\lambda(x) \geq \nu(x)/c_\nu \quad (5)$$

for all  $x \in A^*$ .

We have the following

**Proposition 1.** 1) The mean prediction error by universal semicomputable measure  $\lambda$  for any computable measure  $\mu$  goes to zero as the number of observations increases, i.e.,

$$\lim_{n \rightarrow \infty} F(\mu, \lambda, t, n) = 0$$

for any given prediction length  $n \in \mathbb{N}$ .

2) There exists a sequence of computable measures  $\{\lambda_\tau, \tau \in \mathbb{N}\}$  such that for a computable measure  $\mu$  for  $t \rightarrow \infty$

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} F(\mu, \lambda_\tau, t, n) = 0. \quad (6)$$

Remark 1. The measure  $\lambda$  is closely linked with Kolmogorov complexity. More precisely, it is shown in [2] that

$$|K(U) - \log(1/\lambda(U))| = O(\log |U|),$$

where  $K(U)$  is the Kolmogorov complexity of the word  $U$ .

Proof is shown only for one-step prediction ( $n = 1$ ); the general case is entirely similar. We know that any two probability distributions  $(p_1, p_2, \dots, p_n)$  and  $(q_1, q_2, \dots, q_n)$  satisfy the inequality [17]

$$\sum_{i=1}^n p_i \log(p_i/q_i) \geq 0.$$

Hence it follows that for any measure  $\mu$  and any  $i \in \mathbb{N}$

$$E_\mu[\log(\mu(x_i(\omega)|x_0 \dots x_{i-1}(\omega))/\lambda(x_i(\omega)|x_0 \dots x_{i-1}(\omega)))] \geq 0. \quad (7)$$

From the property (5) of the measure  $\lambda$  we obtain that there exists a constant  $c_\mu$  such that for all  $t \in \mathbb{N}$

$$E_\mu \left[ \log \frac{\mu(x_0 \dots x_{t-1}(\omega))}{\lambda(x_0 \dots x_{t-1}(\omega))} \right] \leq c_\mu.$$

This and the identity

$$E_\mu \left[ \log \frac{\mu(x_0 \dots x_{t-1}(\omega))}{\lambda(x_0 \dots x_{t-1}(\omega))} \right] = E_\mu \left[ \log \frac{\mu(x_0(\omega))}{\lambda(x_0(\omega))} \right] + \sum_{k=1}^{t-1} E_\mu \left[ \log \frac{\mu(x_k(\omega)|x_0 \dots x_{k-1}(\omega))}{\lambda(x_k(\omega)|x_0 \dots x_{k-1}(\omega))} \right]$$

combined with (7), (4) give for  $t \rightarrow \infty$

$$F(\mu, \lambda, t, 1) = o(1). \quad (8)$$

In order to prove (6), note that by definition of semicomputable measure there exists a sequence of computable measures  $\{\lambda_\tau, \tau \in \mathbb{N}\}$  such that for any word  $U \in A^*$ ,  $\lambda_\tau(U)$  is monotonically increasing and converges to  $\lambda(U)$ . This sequence of measures satisfies (6). The proof is the same as for (8).

### 3. Prediction of Stationary Sequences

In order to define the measure  $\rho$  that solves the prediction problem for stationary sequences, we introduce some auxiliary definitions. Let  $u = u_1 u_2 \dots u_n$  and  $v = v_1 \dots v_k$  be two words in  $A^*$  and  $k \leq n$ . By  $r_v(u)$  we denote the number of occurrences of the word  $v$  in the sequence  $u_1 \dots u_k, u_2 \dots u_{k+1}, \dots, u_{n-k+1} \dots u_n$ .

We define the measure  $\rho_k, k \in \mathbb{N}$  on  $\Omega(A)$  by defining it on the sets  $\Omega(y), y \in A^*$  by the equality

$$\rho_k(y) = \begin{cases} |A|^{-|y|} & \text{for } |y| \leq k, \\ \left( \frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \right)^{|A|^k} \frac{1}{|A|^k} \prod_{\alpha \in A^k} \frac{\Gamma(r_{\alpha\alpha}(y) + 1/2)}{\Gamma(\bar{r}_\alpha(y) + |A|/2)} & \text{for } |y| > k, \end{cases} \quad (9)$$

where  $\bar{r}_\alpha(y) = \sum_{\alpha \in A} r_{\alpha\alpha}(y)$ ,  $\Gamma(x)$  is the gamma-function.

The measures  $\rho_k, k \in \mathbb{N}$  were constructed in universal coding theory; for any  $n \geq 1$ , the quantities  $\lceil -\log \rho_k(u) \rceil, u \in A^n$  define an optimal universal code on the family of Markov sources with memory  $k$  ( $k = 0$  corresponds to Bernoulli sources) [18].

In order to define the measure  $\rho$  we have to determine the probability distribution on the set of nonnegative integers introduced in [19]. By definition, let

$$\log^{(0)}(x) = x, \quad \log^{(i)}(x) = \log(\log^{(i-1)}(x))$$

for  $i \geq 1$  and  $m(x) = i$  such that

$$0 \leq \log^{(i)} x < 1.$$

For  $n \in \mathbb{N}$  let

$$w(n) = \sum_{i=1}^{m(n)} \lfloor \log^{(i)}(n) \rfloor + m(n) + 1,$$

$$\lambda(n) = 2^{-w(n)}.$$

It is shown in [19] that there exists a mapping  $\varphi: N \rightarrow \{0, 1\}^*$  such that the set of words  $\{\varphi(n), n \in N\}$  is separable (decodable) and  $|\varphi(n)| = w(n)$ , so that

$$|\varphi(n)| = \log n + O(\log \log n) \quad (10)$$

for  $n \rightarrow \infty$ .

The measure  $\rho$  is defined by the equality

$$\rho(y) = \sum_{k=1}^{\infty} \lambda(k) \rho_k(y). \quad (11)$$

(This measure is well-defined, since  $\sum_{k=1}^{\infty} \lambda(k) = 1$  [19]). Now let

$$v(y) = \max_{k \in N} (\lambda(k) \rho_k(y)), \quad \bar{\rho}(y) = v(y) / \left( \sum_{u \in A^{|y|}} v(u) \right).$$

The measures  $\rho$  and  $\bar{\rho}$  have the same asymptotic properties, and therefore below we only consider the measure  $\rho$ . The values of this measure for some  $u \in \{0, 1\}^*$  are given in Table 1. We will show that this table can be used to obtain a prediction for the examples in the introduction. Let  $x_0 x_1 x_2 x_3(\omega) = 0000$ . Then the probability of occurrence of a zero after this sequence is given by the general formula

$$\rho(0/0000) = \frac{\rho(00000)}{\rho(0000)} = \frac{0.14518}{0.14518 + 0.02637} \approx 0.85.$$

The probability of occurrence of a one after 0000 is much lower,  $\approx 0.15$ . We similarly find that the probability of occurrence of a zero after 0101 is  $\approx 2/3$  and the probability of occurrence of a one is  $\approx 1/3$ . The following theorem characterizes the properties of  $\rho$ .

**THEOREM 1.** Let  $\mu$  be a stationary measure on  $\Omega(A)$ . Then the mean prediction error by the measure  $\rho$  for any  $n \in N$  for almost all (in the measure  $\mu$ )  $\omega \in \Omega(A)$  goes to zero:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} r(\mu, \rho, t, n, \omega) = 0. \quad (12)$$

Moreover,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \bar{r}(\mu, \rho, t, n) = 0. \quad (13)$$

The proof is given in the Appendix.

For Bernoulli and Markov measures, the theorem can be strengthened. Let  $M_0$  and  $M_k$  ( $k \geq 1$ ) be the set of Bernoulli measures and the set of Markov measures with memory  $k$  on  $\Omega(A)$ . We have

**Proposition 2.** For any measure  $\mu \in M_k$ ,  $k \in N$  and any  $n \geq 1$ , for almost all  $\omega \in \Omega(A)$  we have

$$\lim_{t \rightarrow \infty} r(\mu, \rho, t, n, \omega) = 0. \quad (14)$$

The proof easily follows from the central limit theorem for Markov chains.

Proposition 1 does not hold for arbitrary stationary measures. We moreover have

**Proposition 3.** For any prediction method specified by some measure  $\pi$ , there exists a stationary (and ergodic) measure  $\nu$  such that with positive probability

$$\lim_{t \rightarrow \infty} r(\mu, \pi, t, 1, \omega) \neq 0.$$

The proof is given in the Appendix.

Theorem 1 and the Shannon-McMillan-Breiman theorem [20] (see Appendix) easily lead to

**Proposition 4.** For any ergodic and stationary measure, for almost all (in the measure  $\mu$ )  $\omega \in \Omega(A)$

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \rho(x_n \dots x_{n-1}, \omega) = h(\mu), \quad (15)$$

TABLE 1. Values of the Measure  $\rho$ 

$u$	$\rho(u)$	$u$	$\rho(u)$
00000	0.14518	01000	0.02702
00001	0.02637	01001	0.01837
00010	0.02246	01010	0.03447
00011	0.01706	01011	0.01787
00100	0.02344	01100	0.01532
00101	0.01837	01101	0.01837
00110	0.01532	01110	0.01837
00111	0.01837	01111	0.04004

where  $h(\mu)$  is the entropy of  $\mu$ .

Note that previously, the universal coding theory only provided the equality

$$\lim_{n \rightarrow \infty} E_{\mu} \left( -\frac{1}{n} \log \rho(x_0 \dots x_{n-1}(\omega)) \right) = h(\mu),$$

but a result very close to (15) was obtained by analyzing the relationship of Kolmogorov complexity and Shannon entropy (see [2, Theorem 5.3 and Proposition 5.1]).

There exist many other measures which, like  $\rho$ , satisfy the Theorem 1. The choice of a suitable measure for prediction is based on

**THEOREM 2.** The mean prediction error by the measure  $\rho$  is minimal on the family of Bernoulli and Markov sources. More precisely, for any  $k \in \mathbb{N}$ ,  $\mu \in M_k$ , for  $T \rightarrow \infty$ ,

$$\frac{1}{T} \sum_{t=0}^{T-1} \bar{F}(\mu, \rho, t, 1) \leq \frac{(|A|-1)|A|^k}{2T} \log T + O\left(\frac{1}{T}\right).$$

On the other hand, for prediction by any measure  $\nu$  for all  $k \in \mathbb{N}$ ,

$$\sup_{\mu \in M_k} \frac{1}{T} \sum_{t=0}^{T-1} \bar{F}(\mu, \nu, t, 1) \geq \frac{(|A|-1)|A|^k}{2T} \log T + O\left(\frac{1}{T}\right).$$

Both inequalities are known in universal coding theory (see [3] and [18], respectively).

#### 4. Discussion and Examples

Let us return to the problem of Laplace. Let  $A = \{0, 1\}$  and suppose that it is required to predict the value in this alphabet at the moment  $t$ , given that zeros were observed at the moments  $0, 1, \dots, t-1$ . In this case the predictions by  $\rho$  and  $\lambda$  are close to each other: for both measures, the probability of occurrence of 1 at the moment  $t$  does not exceed  $c/t$ , where  $c$  is a constant (this bound is easily obtained from the definition of the measure  $\rho$  and Proposition 1 in Sec. 2).

In some cases, however, the predictions by the measures  $\lambda$  and  $\rho$  are essentially different. Consider the infinite sequence  $\gamma \in \Omega(\{0, 1\})$  obtained by writing the words in the alphabet  $\{0, 1\}^*$  in lexicographic order:

$$\gamma = 0 \ 1 \ 00 \ 01 \ 10 \ 11 \ 000 \dots$$

(blanks have been inserted only for the reader's convenience). We know that the limiting frequency of occurrence of any word  $u \in \{0, 1\}^*$  as a subword in  $\gamma$  is given by  $2^{-|u|}$  [20]. (Sequences from  $\Omega(\{0, 1\})$  having this property are called normal [20]). This combined with definition (11) easily show that for large  $t$  the prediction by the measure  $\rho$  is the following: 1 and 0 are observed at the moment  $(t+1)$  with probability  $1/2$ . The prediction by the measure  $\lambda$  is entirely different: for large  $t$ , the probability of a "correct" prediction of  $x_{t+1}(\gamma)$  tends to 1. In other words, the Turing machine realizing the measure  $\lambda$  will "notice" for large  $t$  the regularity that defines  $\gamma$  and will exploit it for prediction, whereas the machine realizing the measure  $\rho$  will "think" that  $\gamma$  is generated by tossing a symmetrical coin with 1 for heads and 0 for tails.

A prediction method based on a measure close to  $\rho$  has been "tried" for forecasting droughts in various large areas using the data of [21], where drought years are tabulated for some large zones of the planet. The longest series of observations is available for Western Europe (from 796 AD), and the period of "reliable observations" for this region starts in 1400. The series in [21] extends to 1976 (inclusive). The prediction was based

TABLE 2. Using the Measure  $\rho$  for Prediction Based on the Data of [21]

Region and observation period	Prediction for (1976) (probabilities)	
	drought	normal
Western Europe (1400-1975)	0.35	0.65
European USSR (1869-1975)	0.23	0.77
Ukraine (1861-1975)	0.41	0.59
Western Siberia and the Altai Territory (1815-1975)	0.40	0.60

on the data up to 1975 inclusive and was prepared for one year, i.e., for 1976. This technique made it possible to compare the prediction with the actual data for 1976. The results are listed in Table 2; the probability of the actual event is underscored. We see from the table that in all four cases the prediction was fairly uncertain, i.e., the probability of either continuation (drought, no drought) was fairly high. Theorem 1 suggests that, for a long series of observations, the uncertainty of the forecast is determined by the nondeterministic, stochastic nature of the process, and not by prediction accuracy.

In conclusion, note that the proposed prediction methods can be extended to two important applications more general than the problem considered above. In applications it is often required to predict one process from observations of another random sequence, which may be statistically associated with the first. (For example, it is required to predict epidemics using solar activity data, etc.) Another interesting application is prediction using observations with a variable step or, in other words, observations with missing values. (In this case, it may be necessary to predict not only the "future" but also the "past" values of the process.) In both cases, the problem is solved by computing the conditional probabilities using the relevant prediction measure ( $\rho$  or  $\lambda$ ).

#### APPENDIX

Proof of Theorem 1. We first prove the theorem for ergodic and stationary measure  $\mu$  on  $\Omega(A)$ . For  $k \in \mathbb{N}$  we define the  $k$ -th order entropy and the limiting entropy by the equalities

$$h_k(\mu) = E_\mu(-\log \mu(x_k(\omega) | x_0 \dots x_{k-1}(\omega))), \quad h(\mu) = \lim_{k \rightarrow \infty} h_k(\mu). \quad (A.1)$$

Denote by  $A_\mu$  the set of elements  $\omega = \omega_0 \omega_1 \dots \in \Omega(A)$  such that for every word  $u \in A^*$  the limiting frequency of the word  $u$  in the sequence

$$\omega_0 \omega_1 \dots \omega_{|u|-1}, \quad \omega_1 \dots \omega_{|u|}, \quad \omega_2 \omega_3 \dots \omega_{|u|+1}, \dots$$

is  $\mu(u)$  (such  $\omega$  are called  $\mu$ -normal).

Ergodicity of  $\mu$  implies that

$$\mu(A_\mu) = 1 \quad (A.2)$$

and for each  $\omega \in A_\mu$  from the definition of the measure  $\rho_k$  (8) we easily obtain that

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \rho_k(x_0 \dots x_{n-1}(\omega)) = h_k(\mu).$$

This equality and the definition of the measure  $\rho$  (11) imply that for every  $k \in \mathbb{N}$  and  $\omega \in A_\mu$

$$\overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log \rho(x_0 \dots x_{n-1}(\omega)) \leq \overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log (\rho_k(x_0 \dots x_{n-1}(\omega)) \lambda_k) = \overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log \rho_k(x_0 \dots x_{n-1}(\omega)) = h_k(\mu).$$

This combined with (A.1), (A.2) gives for almost all  $\omega \in \Omega(A)$

$$\overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log \rho(x_0 \dots x_{n-1}(\omega)) \leq h(\mu). \quad (A.3)$$

It is shown in [2, 22] that for almost all  $\omega \in \Omega(A)$

$$\lim_{n \rightarrow \infty} (K(x_0 \dots x_{n-1}(\omega))/n) = h(\mu), \quad (A.4)$$

where  $K(u)$  is the Kolmogorov complexity of the word  $u$ .

Let  $\theta(u)$  be a recursive function mapping  $\{0, 1\}^*$  to  $A^*$ ; define

$$K_\theta(u) = \begin{cases} \min\{|v| : \theta(v) = u\}, & \text{if } \{v : \theta(v) = u\} \neq \emptyset, \\ \infty, & \text{if } \{v : \theta(v) = u\} = \emptyset. \end{cases} \quad (\text{A.5})$$

We know [2] that for any general recursive function  $\theta$  there exists a constant  $c_\theta$  such that for all  $u \in A^*$

$$K(u) \leq K_\theta(u) + c_\theta. \quad (\text{A.6})$$

A code, i.e., a mapping  $\psi_n: A^n \rightarrow \{0, 1\}^*$  was constructed for every  $n \geq 1$  in [3] such that for all  $u \in A^n$

$$|\psi_n(u)| \leq \log \rho(u) + 1. \quad (\text{A.7})$$

Define the mapping  $\psi: A^* \rightarrow \{0, 1\}^*$  by the equality

$$\psi(u) = \varphi(|u|)\psi_{|u|}(u).$$

From (A.7) and (10) it follows that

$$|\psi(u)| \leq \log \rho(u) + O(\log |u|). \quad (\text{A.8})$$

Coding and decoding algorithms for the codes  $\psi_n$  and  $\varphi$  are given in [3, 19]. Using these algorithms, we can also easily construct a decoding algorithm for  $\psi$ : first in  $\varphi(|u|)\psi_{|u|}(u)$  identify the word  $\varphi(|u|)$  (this is feasible, since the code  $\varphi$  is separable [19]) and determine  $|u|$ , then use the word  $\psi_{|u|}(u)$  to find  $u$ . Since a coding and decoding algorithm exists for  $\psi$  then from (A.3), (A.6), and (A.8) we conclude that for all  $\omega \in \Omega(A)$  and  $n \in \mathbb{N}$

$$K(x_0 \dots x_{n-1}(\omega)) \leq \log \rho(x_0 \dots x_{n-1}(\omega)) + O(\log n).$$

This equality and (A.4) give

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \rho(x_0 \dots x_{n-1}(\omega)) \geq h(\mu).$$

This and (A.3) show that for almost all  $\omega \in \Omega(A)$

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \rho(x_0 \dots x_{n-1}(\omega)) = h(\mu).$$

The Shannon-McMillan-Breiman theorem [20] asserts that for almost all  $\omega \in \Omega(A)$

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mu(x_0 \dots x_{n-1}(\omega)) = h(\mu).$$

The last two equalities prove Theorem 1 for ergodic stationary measures.

Let  $\Lambda$  be the set of all stationary ergodic measures on  $\Omega(A)$ . We know that any stationary measure  $\mu$  on  $\Omega(A)$  may be represented as an integral over some probability measure on  $\Lambda$  (for an exact formulation, see, e.g., [23, Corollary 49, 7]). We will prove the theorem for the simplest case,  $\mu = \alpha_1 \nu_1 + \alpha_2 \nu_2$ ,  $\alpha_1 + \alpha_2 = 1$ ,  $\alpha_i \geq 0$ ,  $\nu_i \in \Lambda$  (the general proof is similar, although quite lengthy). Let  $A_\mu = A_{\nu_1} \cup A_{\nu_2}$ , where  $A_{\nu_i}$  is the set of  $\nu_i$ -normal  $\omega \in \Omega(A)$ ,  $i = 1, 2$ . Since  $\mu(A_\mu) = 1$ , then we will only consider  $\omega \in A_\mu$  in our proof. The sets  $A_{\nu_1}$  and  $A_{\nu_2}$  are nonintersecting, and therefore it suffices to consider the case  $\omega \in A_{\nu_1}$ . By the theorem of Shannon-McMillan-Breiman,

$$\lim_{n \rightarrow \infty} -\log(\nu_2(x_0 \dots x_{n-1}(\omega)) / \nu_1(x_0 \dots x_{n-1}(\omega))) = \infty \quad (\text{A.9})$$

for  $\omega \in A_{\nu_1}$ . The equalities

$$\begin{aligned} -\frac{1}{n} \log \frac{\rho(x_0 \dots x_{n-1}(\omega))}{\mu(x_0 \dots x_{n-1}(\omega))} &= -\frac{1}{n} \log \frac{\rho(x_0 \dots x_{n-1}(\omega))}{\alpha_1 \nu_1(x_0 \dots x_{n-1}(\omega)) + \alpha_2 \nu_2(x_0 \dots x_{n-1}(\omega))} = \\ &= -\frac{1}{n} \log \frac{\rho(x_0 \dots x_{n-1}(\omega))}{\nu_1(x_0 \dots x_{n-1}(\omega))} + \frac{1}{n} \log \left( \alpha_1 + \alpha_2 \frac{\nu_2(x_0 \dots x_{n-1}(\omega))}{\nu_1(x_0 \dots x_{n-1}(\omega))} \right) \end{aligned}$$

and (A.9) thus show that almost everywhere

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \frac{\rho(x_0 \dots x_{n-1}(\omega))}{\mu(x_0 \dots x_{n-1}(\omega))} = \lim_{n \rightarrow \infty} -\frac{1}{n} \log \frac{\rho(x_0 \dots x_{n-1}(\omega))}{\nu_1(x_0 \dots x_{n-1}(\omega))} = 0.$$

Q.E.D.

**Proof of Proposition 3.** Construct a family of ergodic stationary measures  $M$  defined on the set  $\Omega\{(a, b, c)\}$ . Each  $\mu \in M$  is defined by a Markov chain with a countable number of states  $0, 1, 2, \dots$ , describable as a random walk of a particle on the set  $N$ . From each state  $i \in N$ , there is a probability of  $1/2$  that the particle will go to state  $0$  and generate the letter  $a \in \{a, b, c\}$  and a probability of  $1/2$  that it will go to the state  $i + 1$ . In this event, there is a probability  $\Delta_i/2$  that it generates the letter  $b$  and a probability  $(1 - \Delta_i)/2$  that it generates the letter  $c$ , where  $\Delta_i$  is a parameter equal to  $1/3$  or  $2/3$ . Thus,  $\mu$  is defined by the infinite word  $\Delta = \Delta_1 \Delta_2 \Delta_3 \dots$ ,  $\Delta_i \in \{1/3, 2/3\}$ ; let  $M = \{M_\Delta\}$  contain all measures of this form. It is easy to see that  $\mu_\Delta \in M$  is an ergodic and stationary measure. Take some measure  $\tau$  for prediction.

Omitting the simple, yet lengthy proof, we will explain informally why this family may not satisfy the equality

$$\forall \mu_\Delta \in M : \lim_{T \rightarrow \infty} \log \frac{\mu_\Delta(x_T(\omega) | x_0 \dots x_{T-1}(\omega))}{\tau(x_T(\omega) | x_0 \dots x_{T-1}(\omega))} = 0$$

for almost all  $\omega$ . The point is that the random walk will take the particle to states with arbitrarily large indices (with probability 1 relative to any measure  $\mu_\Delta \in M$ ). For the first arrival in the state  $i \in N$ , exact prediction is impossible either by the measure with  $\Delta_i = 2/3$  or by the measure with  $\Delta_i = 1/3$  (by construction, the occurrence of the letter  $b$  in the next moment is equal to  $1/6$  and  $1/3$  for these measures, respectively). Since the random walk is on a countable set, there are infinitely many first arrivals for different states and wrong predictions will occur for arbitrarily large  $T$  (with nonzero probability at least for one measure  $\mu_\Delta \in M$ ).

#### LITERATURE CITED

1. W. Feller, An Introduction to Probability Theory and Its Applications [Russian translation], Vol. 1, Mir, Moscow (1984).
2. A. K. Zvonkin and L. A. Levin, "Complexity of finite objects and substantiation of the concepts of information and randomness using the theory of algorithms," *Usp. Mat. Nauk*, 25, No. 6, 85-127 (1971).
3. B. Ya. Ryabko, "Doubly universal coding," *Probl. Peredachi Inf.*, 20, No. 3, 24-28 (1984).
4. J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans., Inf. Theory*, 30, No. 4, 629-636 (1984).
5. A. N. Kolmogorov, "Three approaches to the definition of 'quantity of information'," *Probl. Peredachi Inf.*, 1, No. 1, 3-7 (1965).
6. R. J. Solomonoff, "A formal theory of inductive inference," *Inf. Control*, 7, No. 1, 1-22 (1964).
7. B. M. Fitingof, "Optimal coding with unknown and variable message statistics," *Probl. Peredachi Inf.*, 2, No. 2, 3-11 (1966).
8. R. E. Krichevskii, "Relationship between coding redundancy and source information reliability," *Probl. Peredachi Inf.*, 4, No. 3, 48-57 (1968).
9. R. E. Krichevskii, Lectures in Information Theory [in Russian], Novosibirsk State Univ. (1970).
10. V. K. Trofimov, "Redundancy of universal coding of arbitrary Markov sources," *Probl. Peredachi Inf.*, 10, No. 4, 16-24 (1974).
11. Yu. M. Shtar'kov, "Coding of finite-length messages on the output of a source of unknown statistical properties," *Proc. 5th All-Union Conf. on Coding Theory and Information Transmission, Abstracts of Papers [in Russian], Part 1, Moscow-Gor'kii (1972)*, pp. 147-152.
12. J. M. Starkov [Yu. M. Shtar'kov] and V. F. Babkin, "Combinatorial encoding for discrete stationary sources," *2nd Int. Symp. Information Theory, Tsahkadzor, Armenia, USSR, 1971, Akad. Kiado, Budapest (1973)*, pp. 249-257.
13. E. N. Gilbert, "Codes based on accurate source probabilities," *IEEE Trans., Inf. Theory*, 17, No. 3, 304-314 (1971).
14. R. E. Krichevskii, "Optimal coding of a source from observations," *Probl. Peredachi Inf.*, 11, No. 1, 37-42 (1975).
15. R. E. Krichevsky [Krichevskii] and V. K. Trofimov, "Optimal sample based coding of Markov sources," *Third Czechoslovak-Soviet-Hungarian Seminar on Information Theory, Liblice (1980)*, pp. 131-138.
16. J. Rissanen, A Predictive Inference Principle for Estimation, IBM Res. Lab. Preprint, San Jose, Calif. (1984).

17. R. Gallager, Information Theory and Reliable Communication [Russian translation], Sovet-skoe Radio, Moscow (1974).
18. R. E. Krichevsky [Krichevskii] and V. K. Trofimov, "The performance of universal coding," IEEE Trans., Inf. Theory, 27, No. 2, 199-207 (1981).
19. V. I. Levenshtein, "On redundancy and slowdown of separable coding of natural numbers," in: Probl. Kibern., No. 20, Nauka, Moscow (1968), pp. 173-179.
20. P. Billingsley, Ergodic Theory and Information [Russian translation], Mir, Moscow (1969).
21. Yu. L. Rauner, Climate and Grain Yields [in Russian], Nauka, Leningrad (1981).
22. A. A. Bruno, Entropy and Algorithmic Complexity of the Path of a Dynamic System [in Russian], Preprint, VNIИ Sistemnykh Issledovanii, Moscow (1980).
23. K. Parthasarathy, An Introduction to Probability Theory and Measure Theory [Russian translation], Mir, Moscow (1983).

## DECODING OF BERMAN MODULAR MDS CODES

I. I. Grushko

UDC 621.391.15:681.3.053

A new algorithm is described for decoding of cyclic MDS codes with the parameters  $(p, k, p - k + 1)$ , where  $p$  is an arbitrary odd prime. The decoding is based on Berman's definition [1] of modular  $p$ -codes, as the MDS codes considered in this paper are a particular subclass of Berman's modular codes.

### 1. Introduction

The modular codes described by Berman [1] in 1967 include a subclass of MDS codes with the parameters  $(p, k, d = p - k + 1)$ , where  $p$  is a prime and the codes are over the prime field  $F_p = GF(p)$ . MDS codes occupy a special place in the theory and practice of error-correcting codes because of two important properties: 1) the number of information symbols  $k$  is maximal among all codes with given block length  $p$  and distance  $d$ ; 2) any  $k$  symbols in the codeword form an information sequence, so that shortening of the code again produces a MDS code. Reed-Solomon (RS) codes constitute the best studied and most widely used class of MDS codes. Berman codes are equivalent to RS codes extended by one position. In our opinion, Berman codes are better because they do not contain a designated "parity check" position; all the codeword positions are equivalent. The introduction of a designated position, although inessential, complicates the decoding procedure. Unfortunately, unlike RS codes, Berman codes exist only for prime lengths. This is partly compensated by the fact that prime fields are naturally embedded in the field of real numbers commonly used in engineering. All calculations can be carried out in the field of reals, followed by reduction modulo a prime number.

Berman codes were rediscovered in 1973 in [2], where a realization was described which does not require multiplications for coding and syndrome computation. This was achieved by a special choice of the structure of the check matrix. In this study, we propose a different approach to the realization of Berman codes, which for small error rates is as simple as the simplest known decoding schemes.

In conclusion of the introductory section, we recall the definition of Berman codes and list their relevant properties. Let  $I_d$  be a cyclic code over the field  $F_p = GF(p)$  of residue classes modulo in the prime number  $p$  with generating polynomial  $g(x) = (x - 1)^{d-1}$ . It is well known that  $I_d$  is a MDS code with the parameters  $(p, k, d = n - k + 1)$  correcting any  $t = [(d - 1)/2]$  errors (it is a  $t$ -error correcting code). The codes  $I_d$  are nested  $I_1 \supset I_2 \supset \dots \supset I_{p-1} \supset I_p = (1)$ ; in the usual terminology,  $I_2$  is a "parity check" code and  $I_p$  is a repetition code.

For an arbitrary polynomial  $v(x) = \sum_{i=0}^{p-1} v_i x^i$  of the ring  $R_p = F_p[x]/(x^p - 1)$  denote by  $v_i(x)$  its cyclic shift  $i$  positions to the left

---

Translated from Problemy Peredachi Informatsii, Vol. 24, No. 2, pp. 15-21, 1988. Original article submitted July 16, 1985; revision submitted August 10, 1987.