JOURNAL OF COMPLEXITY **10**, 281–295 (1994)

The Complexity and Effectiveness of Prediction Algorithms

BORIS YA. RYABKO*

Department of Applied Mathematics and Cybernetics, Novosibirsk Telecommunication Inst., Kirov Str.-86, Novosibirsk - 102, 630102, Russia

Received May 8, 1992

The problem of predicting an arbitrary sequence $x_1x_2x_3 \cdots$ is considered with x_{t+1} being predicted from an analysis of the word $x_1x_2 \cdots x_t$. There are no presumptions concerning the probability structure on this process. The relation between prediction effectiveness and Kolmogorov complexity is established. Then the Hausdorff dimension of sets of sequences for which effective methods of prediction exist is estimated. The prediction method which is asymptotically superior to other arbitrary ones realized by finite automata is constructed.

1. INTRODUCTION

Methods of prediction of the time series based on the "classic" approach are applicable only when the probability structure of the process is known precisely enough (e.g., stationarity or a certain trend of the process are required). Quite often, we have no information about the probability structure of the actual process to be predicted. Therefore, we need an alternate approach to the problem of prediction. We propose the algorithmic approach. Using this approach, we look for a method that is optimal, not for the class of random processes, but for the class of prediction methods. The class of prediction methods is provided by means of descriptions of algorithms realizing these methods (e.g., prediction methods that are found by finite automata, arbitrary Turing machines, and so on). Previously, the algorithmic approach to the prediction problem was discussed in [22, 12, 19, 7].

* Email address: ryabko@CAD-LAB.neic.nsk.su.

0885-064X/94 $6.00

The main differences of the present work from [22, 12] are that, first, the prediction methods proposed in [22, 12] based on the universal measure and the Occam measure are not realizable algorithmically, and, second, and present work offers a quantitative criterion for the prediction method efficiency that coincides with an asymptotic winning value in a game.

This quantitative criterion was proposed in [9]. Then it was reused in many papers concerning an optimal investigation [1, 3, 4]. Some interesting and important results concerning the game theory approach and complexity are found in [6, 20, 22].

In this paper the problem of the construction of the optimal prediction method is investigated in two statements: first, the optimal method is found within the class of all realizable ones; second, a prediction method that is superior to an arbitrary one, realized by finite automata, is discovered. In the first case the relation between an optimal prediction and Kolmogorov complexity is established. Then the set of sequences being effectively predictable is considered (see Section 3) and its Hausdorff dimension is estimated. In Section 4 the prediction method ρ, which is effective then any realized by finite automata, is constructed. It is shown in Section 5 that the method ρ is also asymptotically optimal for the "classic" classes of random processes.

We use the game-theoretic interpretation of J. L. Kelly [9]. Let us consider that gambler I has capital V_t at moment $t = 0, 1, 2, \ldots$, $V_0 = 1$. Each moment t that the gambler divides the capital V_t into $|A|$ parts equals $V_t \cdot P(Q/x_1 \cdots x_t)$, $Q \in A$. (The Stakes on the $x_{t+1} = a \in A$.) $P(a/x_1 \cdots x_t)$ reflect the I-gambler's confidence in the appearance of $a \in A$ at the moment $(t + 1)$.

The gambler I learns X_{t+1} at the moment $(t + 1)$ and his capital becomes equal to $|A| \cdot V_t \cdot P(x_{t+1}/x_1 \cdots x_t)$ (i.e., the stakes on the "right" letter are increasing $|A|$ times).

Denote win by the method J with $x_1x_2 \cdots x_t$ as $W_J(x_1x_2 \cdots x_t)$:

$$W_J(x_1x_2 \cdots x_t) = \prod_{i=1}^{t} (|A| \cdot P(x_i/x_1x_2 \cdots x_{i-1}). \tag{1}$$

In this case the gambler divides this capital into $|A|$ parts each moment (stakes on the letters of A), and at the moment $(t + 1)$ his capital should be equal to $W_J(x_1x_2 \cdots x_{t-1})$. $(|A| \cdot P(x_t/x_1x_2 \cdots x_{t-1}))$, i.e., the stake on the letter that actually appears, increases $|A|$ times.

Our goal is to find the algorithms maximizing (1). From the mathematical point of view it is convenient to study the logarithm of $W_J(x_1x_2 \cdots x_t)$ divided by t. Let

$$L_J(x_1x_2 \cdots x_t) = \frac{1}{t} \log W_J(x_1x_2 \cdots x_t).$$

(Here and below $\log x = \log_2 x$). Then

$$L_J(x_1x_2 \cdots x_t) = \log |A| + \frac{1}{t} \sum_{i=1}^{t} \log P_J(x_i/x_1x_2 \cdots x_{i-1}). \qquad (2)$$

We note that, if at the word $x_1x_2 \cdots x_t$ strategy J recommends staking all the capital V_0 ($V_0 = 1$) on x_1 (that is unknown), then the capital $V_1 = |A|$ on x_2 and so on. This results in a maximal win equal to $|A|^t$ after t games.

If the gambler breaks up all his capital into $|A|$ equal parts each time and stakes those parts on the letters of A, then the win is equal to 0 and the capital remains the same: $V_0 = V_1 = \cdots = V_t = 1$. In general, $L_J(x_1x_2 \cdots x_t)$ is the mean value of the exponent of the capital increase over the first t games: $V_t = |A| \cdot 2^{tL}J(x_1x_2 \cdots x_t)$. This value characterizes quantitatively the efficiency of prediction J.

As an illustration, we consider the following example. In Table I there are the results of the last 11 World Football Championships in the upper line. These championships were held once every four years; the last was in Spain in 1990. The winners were either European teams or teams from South America. If we denote the win of a European team by 0 and the win of a non-European team by 1, we should then have a sequence of numbers in the second line of the table. For instance, in 1950 the Uruguay team won (1), in 1954 the West German team won (0), etc.

TABLE I

THE PREDICTION OF THE RESULT OF THE MATCH FOR THE WORLD FOOTBALL CHAMPIONSHIP BY THE METHOD p_ν OPTIMAL IN THE CLASS OF FINITE-AUTOMATIC METHODS

N	1	2	3	4	5	6	7	8	9	10	11
The data of football championship	1950	1954	1958	1962	1966	1970	1974	1978	1982	1986	1990
The code of the winning team (0 = the European team, 1 = non-European)	1	0	1	1	0	1	0	1	0	1	0
The prediction: (the stake on the winning team is underlined) the stake on "0"	0.5	<u>0.375</u>	0.5	0.542	<u>0.409</u>	0.389	<u>0.431</u>	0.326	<u>0.573</u>	0.227	<u>0.678</u>
the stake on "1"	<u>0.5</u>	0.625	<u>0.5</u>	<u>0.458</u>	0.591	<u>0.611</u>	0.569	<u>0.674</u>	0.427	<u>0.773</u>	0.322
The gambler's capital after the match	1	0.75	0.75	0.687	0.563	0.687	0.594	0.801	0.918	1.416	1.920

Note. The information of the result of the preceding match is used for predicting the next match (beginning in 1950).

We suppose that there was a gambler who had the capital $V_0 = 1$ before the 1954 Championship. Prior to each championship he makes a prediction; part of his capital is allotted to 0 (the win by a European team), and other part to 1 (the win by a non-European team). The method used for this prediction is optimal for the class of methods realizable by finite automata (this method is described in the third part of this paper).

To predict the result of the next championship the information about previous cases from Table I is used. For prediction of the 1950 championship result, for instance, we had no information, but for prediction of the 1960 championship result we could use the sequence 1011, etc. Thus, prior to each next championship we had more and more information.

As may be seen from the last line of the table, the gambler's capital has increased up to 1.92 after 11 championships, and during the last four, the capital was only increasing. Thus, as the number of cases rises, the prediction becomes more and more precise because the gambler uses the available information about previous results for his prediction.

2. Turing Prediction

Let us give a definition of the prediction method J that is realizable on Turing machine M. For simplicity sake, we consider only the case of $A = \{0,1\}$ in this section. It is assumed that machine M has a single working tape. The word $y \in \{0,1\}^*$ is written on this tape beforehand. Then the machine begins to operate and stops after a finite period of time having two nonnegative numbers $P(0/y)$ and $P(1/y)$ typed on the tape. We may assume that each of those numbers is rational and is represented as a rational fraction. Denote the class of such prediction methods as T. The important concept of Kolmogorov complexity was defined in [10, 19]; see also [11]. The connection between the best prediction method and Kolmogorov complexity is established as follows.

Theorem 1. *Let x be an arbitrary word in $\{0,1\}^\infty$ and let $J \in T$ be a prediction method that is realizable on a Turing machine. Then, for $n \to \infty$*

$$L_J(x^{(n)}) \leq 1 - K(x^{(n)})/n + O(\log n/n), \tag{3}$$

where $K(x^{(n)})$ is the Kolmogorov complexity of the word $x^{(n)}$. Moreover, there is a sequence of prediction methods $J_s \in T$, $s = 1, 2, \ldots$, such that for every $x \in \{0,1\}^\infty$ and $n \to \infty$,

$$\lim_{s\to\infty} L_{J_s}(x^{(n)}) = 1 - K(x^{(n)})/n + O(\log n/n). \tag{4}$$

Proof. We base the proof on the concept of the universal semicomputable measure introduced in [22]. Let us give a definition. The measure μ defined on $\{0,1\}^\infty$ is called computable if there exist general recursive functions $F(y, n)$ and $G(y, n)$ defined on all y, $\{0,1\}^*$, and a positive integer n such that the number $F(y, n)/G(y, n)$ approximates $\mu(y, n)$ with accuracy 2^{-n}. (Recall, that a function $f(y, n)$ is called generally recursive if there exists an algorithm that is realizable, say, on a Turing machine such that the result of applying the algorithm to the pair (y, n) equals some nonnegative integer.) A measure v is called semicomputable if there is a generally recursive function $\Phi(y, n)$ and $\Gamma(y, n)$ such that the function

$$\beta_v(y, n) = \Phi(y, n)/\Gamma(y, n) \tag{5}$$

decreases monotonically over n, and

$$\lim_{n\to\infty} p_v(y, n) = v(y) \tag{6}$$

for all y $\{0,1\}^*$, where $\Phi(y, n)$ and $\Gamma(y, n)$ may be chosen in such a way that for all y and n the following equation holds:

$$\beta_v(y, n) \geq \beta_v(y0, n) + \beta_v(y1, n) \tag{7}$$

(see [22 Theorem 3.2]).

The semicomputable measure is defined on $\{0,1\}^\infty \cup \{0,1\}^*$. As is proved in [22], there is the universal semicomputable measure R such that for any semicomputable measure v a constant C_v exists such that

$$R(y) \geq v(y)/C_v \tag{8}$$

for all y, $\{0,1\}^*$. In [22] it is also shown tht for all y, $\{0,1\}^*$,

$$|K(y) - (-\log R(y))| = O(\log|y|). \tag{9}$$

Now, consider an arbitrary prediction method J, T. For every $x_1x_2 \cdots x_n$, $\{0,1\}^*$ we define the measure T_J by means of the equality

$$\pi_J(x_1x_2 \cdots x_n) = \prod_{i=1}^{n} P(x_i/x_1x_2 \cdots x_{i-1}). \tag{10}$$

Obviously, π_J is a computable measure. From (8) we have

$$R(x) \geq \pi_J(x)/C_J.$$

Using the logarithm and taking (2), (10) into account, we have

$$n^{-1} \log R(x) \geq L_J(x_1 x_2 \cdots x_n) - 1 - \log C_J/n.$$

From (9) it follows that

$$L_J(x_1 x_2 \cdots x_n) \leq 1 + K(x^{(n)})/n + O\left(\frac{\log n}{n}\right)$$

that gives us (3).

In order to prove (4) we take two general recursive functions $\Phi(y, n)$ and $\Gamma(y, n)$ defined for all $y \in \{0,1\}^*$ and integers n. These functions specify a semicomputable measure R. According to (5)–(7), $\beta_R(y, n) = \Phi(y, n)/\Gamma(y, n)$ and the following relationships hold:

$$\lim_{n-\infty} \beta_R(y, n) = R(y) \tag{11}$$

$$\beta_R(y, n) \geq \beta_R(y0, n) + \beta_R(y1, n). \tag{12}$$

For every integer n and all $y \in \{0,1\}^*$, $a \in \{0,1\}$ we define the prediction method $J_n \in T$ by

$$P_{J_n}(a/y) = \beta_R(ya, n)/(\beta_R(y0, n) + \beta_R(y1, n)).$$

From the equality above and definition (2) we have

$$L_{J_n}(y_1 y_2 \cdots y_\tau) = 1 + \frac{1}{\tau} \log \left(\prod_{i=1}^{\tau} \frac{\beta_R(y_1 y_2 \cdots y_i, n)}{\beta_R(y_{i-1}0, n) + \beta_R(y_{i-1}1, n)} \right).$$

From (12) it follows that

$$\begin{aligned} L_{J_n}(y_1 y_2 \cdots y_\tau) &= 1 + \frac{1}{\tau} \log \left(\prod_{i=1}^{\tau} \frac{\beta_R(y_1 y_2 \cdots y_i, n)}{\beta_R(y_1 y_2 \cdots y_{i-1}, n)} \right) \\ &= 1 + \frac{1}{\tau} \log \beta_R(y_1 y_2 \cdots y_\tau). \end{aligned}$$

For $n \to \infty$ it follows from (11) and (9) that

$$\begin{aligned} \lim_{n \to \infty} L_{J_n}(y_1 \cdots y_\tau) &= 1 + \frac{1}{\tau} \log R(y_1 \cdots y_\tau) \\ &= 1 - \frac{1}{\tau} K(y_1 \cdots y_\tau) + O\left(\frac{\log \tau}{\tau}\right). \end{aligned}$$

Taking into account that (3) is valid for any prediction $J \in T$ and, hence, for J_n, $n = 1, 2, \ldots$, from the latter relationship and (3) we have (4). The theorem is proven.

For every $\alpha \in [0,1]$ a set $\tilde{M}_\alpha$ of x words, $x \in \{0,1\}^\infty$ is defined as

$$\tilde{M}_\alpha = \left\{x: \sup_{M \in T} \overline{\lim_{n \to \infty}} L_M(x^{(n)}) \geq \alpha\right\}. \tag{13}$$

In other words, $\tilde{M}_\alpha$ consists of those x for which such a game strategy exists that the gambler's capital should asymptotically increase as $2^{\alpha n}$, where n is the number of games. We can readily show that for any $\alpha > 0$ the Lebesgue measure of the set $\tilde{M}_\alpha$ equals 0; therefore it is necessary to use other characteristics for comparison of $\tilde{M}_\alpha$ for different α. It turns out that we can evaluate the "bulk" of $\tilde{M}_\alpha$ by means of the Hausdorff dimension. In order to define the Hausdorff dimension of subsets from $\{0,1\}^\infty$, take the mapping $\sigma: \{0,1\}^\infty \to [0,1]$ that assigns to the word $x = x_1 x_2 \cdots \in \{0,1\}^\infty$ the number $\sigma(x) \in [0,1]$ whose binary expansion has the form $0.x_1 x_2 \ldots$. For every $X \subset \{0,1\}^\infty$ denote by $DH(X)$ the Hausdorff dimension of the set $\sigma(X) \subset [0,1]$. (For the definition of the Hausdorff dimension see, for example, [2].)

THEOREM 2. *For any $\alpha \in [0,1]$ the equality $DH(\tilde{M}_\alpha) = 1 - \alpha$ is valid.*

Proof. For every $\beta \in [0,1]$ the set $\tilde{N}_\beta$ is defined as

$$\tilde{N}_\beta = \{x: x \in \{0,1\}^\infty, \underline{\lim_{n \to \infty}} K(x^{(n)})/n \leq \beta\}. \tag{14}$$

In [16, 17] it is shown that

$$DH(\tilde{N}_\beta) = \beta. \tag{15}$$

It readily follows from the Theorem 1 that for all $x \in \{0,1\}^\infty$

$$\sup_{M \in T} \overline{\lim_{n \to \infty}} L_M(x^{(n)}) \leq 1 - \underline{\lim_{n \to \infty}} K(x^{(n)})/n.$$

From the latter expression and from (13), (14) it follows that for any $\alpha \in [0,1]$, $\tilde{M}_\alpha \subset \tilde{N}_{1-\alpha}$ is valid. Hence, $DH(\tilde{M}_\alpha) \leq DH(\tilde{N}_{1-\alpha})$. That yields, along with (15),

$$DH(\tilde{M}_\alpha) \leq 1 - \alpha. \tag{16}$$

In order to prove the inequality opposite to (16), we take an arbitrary $\alpha \in [0,1]$ and the number $\pi \in [0,1]$ that is the solution of the equation

$$-(\pi \log \pi + (1 - \pi) \log (1 - \pi)) = 1 - \alpha. \tag{17}$$

(It is easy to see that such a π always exists). We shall consider a set $R_\pi \subset [0,1]$ consisting of the words $x = x_1 x_2 \cdots$ such that the threshold rate of the 0 occurring in x does exist and equals π (the rate of 1 being equal to $(1 - \pi)$). (Such words are obtained with probability 1 during the tossing of an asymmetrical coin, where "0" falls out with probability π and "1" with $(1 - \pi)$). It is known [2] that

$$DH(R_\pi) = -(\pi \log \pi + (1 - \pi) \log(1 - \pi))$$

and it follows from (17) that

$$DH(R_\pi) = 1 - \alpha. \tag{18}$$

In [9] it is shown that there exists a strategy J such that for $x \in R_\pi$ the equality $\lim_{n\to\infty} L_J(x^{(n)}) = 1 + (\pi \log \pi + (1 - \pi) \log(1 - \pi))$ is valid. (This strategy is very simple; we should stake the share of capital π on "0" and the share of capital $(1 - \pi)$ on "1"). Hence, $R_\pi \subset \tilde{M}_\alpha$. Taking this into account and from (18), it follows that $DH(M_\alpha) \geq 1 - \alpha$. This fact and (16) give the proof of the theorem.

Now, we consider the problem of optimal prediction. A method J is asymptotically optimal for the class of methods A if for any $x \in \{0,1\}^n$ and any $J_\alpha \in A$ the inequality

$$L_{J_\alpha}(x) - L_J(x) = o(1),$$

holds (for $n - \infty$), where $o(1) \to 0$ for $n \to \infty$. In other words, the prediction method J is no worse asymptotically than any $J_\alpha \in A$. It turns out that the asymptotically optimal method for the class T of methods that are realizable on Turing machines does not exist. More precisely, it could be shown that there does not exist an algorithm giving an asymptotically optimal method for T. (Such a method could be based on the universal measure R from [22] or the Occam measure from [12]. These measures, however, are not realizable algorithmically.) Therefore an asymptotically optimal prediction can be realized only for the classes of methods that are more restricted than T. One of those classes consists of the prediction methods that are realizable by finite automata.

3. Finite-Automata Prediction

Now we give the definition of the prediction method that is realizable by a finite automation. Let α be a finite automation with a finite number of

states, where i_0 is the initial state. The automation α is assumed to have input and output. If the letter $a \in A$ is at the input, then after some time we should have $|A|$ nonnegative numbers at the output whose sum equals 1. (We may consider all of them rational.) In order to define $P_\alpha(a/x_1x_2 \cdots x_n)$ for the word $x_1x_2 \cdots x_n \in A^*$ and $a \in A$, we should, first, put the automation α into operation from the state i_0. After some time it should type numbers $\{P(a);\ a \in A\}$. Then, $x_1x_2 \cdots x_n$ have to be fed at the input in consecutive order. After x_1 the automation types $\{P(a/x_1);\ a \in A\}$; after x_2 it types $\{P(a/x_1x_2);\ a \in A\}$; etc. After x_n it types $\{P(a/x_1x_2 \cdots x_n);\ a \in A\}$.

Here we describe the prediction method ρ whose efficiency for any word $x \in A^\infty$ asymptotically is as high as with any method that is realizable by a finite automation. We give some definitions. Let $u = u_1u_2 \cdots u_n$, $v = v_1v_2 \cdots v_k$ by two words of A^* and $k \le n$, and let $s = \lfloor n/k \rfloor$. By $\tau_v(u)$ we denote the rate of the word v occurring in the sequence $u_1u_2 \cdots u_k$, $u_2u_3 \cdots u_{k+1}, \ldots, u_{n-k+1} \cdots u_n$ and by $\nu_v(u)$ we denote the rate of v occurring in the sequence $u_1u_2 \cdots u_k, u_{k+1} \cdots u_{2k}, \ldots, u_{(s-1)k} \cdots u_{sk}$. For example, $\tau_{00}(000100) = 3$, $\nu_{00}(000100) = 2$. For $k = 0, 1, 2, \ldots$ denote the mapping ρ_k that assigns to each word of A^* the values

$$\rho_k(u) = \begin{cases} |A|^{-|u|} & \text{for } |u| \le k \\ \left(\dfrac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}}\right)^{|A|^k} \cdot \dfrac{1}{|A|^k} = \displaystyle\prod_{\alpha \in A^k} \frac{\prod_{a \in A} \Gamma(\tau_{\alpha a}(u) + 1/2)}{\Gamma(\tilde{\tau}_\alpha(u) + |A|/2)} & \text{for } |u| > k, \end{cases}$$

where $\tilde{\tau}_\alpha(u) = \sum_{a \in A} \tau_{\alpha a}(u)$ and $\Gamma(\ )$ is a gamma function. Now, define the probability distribution λ on the set of nonnegative integers, $0, 1, 2, \ldots$, using the code from [14]. Let

$$\log^{(0)}(x) = x, \qquad \log^{(i)}(x) = \log_2(\log^{(i-1)}(x)) \qquad \text{for } i \ge 1$$

and

$$m(x) = i,$$

so that $0 \le \log^{(i)}(x) < 1$. For $n = 0, 1, 2, \ldots$ define

$$w(n) = \sum_{i=1}^{m(n)} \lfloor \log^{(i)}(n) \rfloor + m(n) + 1, \qquad \lambda(n) = 2^{-w(n)}.$$

It is known that

$$-\log \lambda(n) = \log n + O(\log \log n)$$

when $n \to \infty$ [14].

We define the function p: $A^* \to (0,1)$ by

$$p(u) = \sum_{k=0}^{\infty} \lambda(k)\, p_k(u).$$

The method of prediction is defined by

$$p(a/x_1x_2 \cdots x_t) = p(x_1x_2 \cdots x_t a)/p(x_1x_2 \cdots x_t),$$

where the right part is the value of a stake on letter "a" which has to occur at the moment $(t + 1)$, provided there are letters $x_1x_2 \cdots x_t$ at previous t moments.

The value $p_k(u)$ is known in universal coding theory. There, $[-\log p_k(u)]$ is the length of codewords of the optimal code on the set of kth-order Markovian sources [13].

As an example, consider the prediction computation on $x = x_1 \cdots x_4 = 0101$, $A = \{0,1\}$ by the method p. Taking into account that

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{1}{2} \cdot \frac{3}{2} \cdot \cdots \cdot \frac{(2n-1)}{2}\sqrt{\pi}; \qquad n \geq 1; \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\Gamma(n) = (n-1)!,$$

we have

$$p_0(0101) = \frac{\Gamma(1)}{\Gamma(1/2)^2} \frac{\Gamma(2 + 1/2) \cdot \Gamma(2 + 1/2)}{\Gamma(4+1)} = \frac{1/2 \cdot 3/2 \cdot 1/2 \cdot 3/2}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{3}{128}$$

$$p_1(0101) = \left(\frac{\Gamma(1)}{\Gamma(1/2)^2}\right)^2 \cdot \frac{1}{2}\left(\frac{\Gamma(1/2)\Gamma(2 + 1/2)}{\Gamma(3)}\right)\left(\frac{\Gamma(1/2)\Gamma(1 + 1/2)}{\Gamma(2)}\right) = \frac{3}{32}$$

$$p_2(0101) = \left(\frac{\Gamma(1)}{\Gamma(1/2)^2}\right)^4 \cdot \frac{1}{2^2}\left(\frac{\Gamma(1/2)\,\Gamma(1/2)}{\Gamma(1)}\right)^2\left(\frac{\Gamma(1 + 1/2)\Gamma(1/2)}{\Gamma(2)}\right)^2 = \frac{1}{16}$$

$$p_3(0101) = p_4(0101) = \cdots = \tfrac{1}{16}; \qquad \lambda(0) = \tfrac{1}{2}; \qquad \lambda(1) = \tfrac{1}{4};$$

$$\lambda(2) = \tfrac{1}{16}; \qquad p(0101) = \sum_{k=0}^{\infty} \lambda(k) p_k(0101) = \frac{1}{2} \cdot \frac{3}{128} + \frac{1}{4} \cdot \frac{3}{32}$$

$$+ \left(\sum_{k=2}^{\infty} \lambda(k)\right) \frac{1}{16} = \frac{9}{128} + (1 - (\lambda(0) + \lambda(1))) \cdot \frac{1}{16} = 13 \times 2^{-8}.$$

Similarly, we have

$$p(01010) = 33 \times 2^{-10}; \qquad p(01011) = 19 \times 2^{-10}.$$

Hence, the prediction follows:

$$P_{p^*}(0/0101) = 33/52 \approx 2/3; \qquad P_{p^*}(1/0101) = 19/52 \approx 1/3.$$

Thus, the stake on "0" after "0101" is almost twice as high as the stake on "1." For $k \geq 1$ we define

$$\hat{H}_k(u) = \frac{1}{k} \sum_{v \in A^k} \left(\frac{\nu_v(u)}{|u|/k}\right) \log \left(\frac{\nu_v(u)}{|u|/k}\right).$$

From [13] it is known that for all $k = 0, 1. \ldots, |u| \to \infty$, the inequality

$$-\frac{1}{|u|} \log p_k(u) - \hat{H}_{k+1}(u) \leq \frac{C(k) \log (u)}{|u|} \tag{19}$$

holds, where $C(k)$ does not depend on u.

THEOREM 3. *Let A be a finite alphabet, let x be the word in A^∞, and let α be a method of prediction that is realizable by a finite automation and applicable to all $x^{(n)}$, $n \geq 1$. Then there exists a constant $C(\alpha)$ such that*

$$L_\alpha(x^{(n)}) - L_{p^*}(x^{(n)}) \leq \frac{C(\alpha) \log n}{n}.$$

Otherwise, the precision of a prediction by the method p^* is asymptotically not worse than for any finite-automata prediction.

Proof. First, we shall give an auxiliary assertion. Let $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ be two vectors of dimension n, $n \geq 1$ such that $p_i \geq 0$, $q_i \geq 0$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n q_i \leq 1$. Then

$$\sum_{i=1}^n p_i \log \frac{p_i}{q_i} \geq 0. \tag{20}$$

(Here, as usual, $0 \log 0 = 0$). This inequality is widely known in information theory (see, for example, [8]).

Here we consider $x \in A^\infty$ and the prediction method α that is realized by a finite automation with a set of states S, $|S| < \infty$. We suppose that there is an initial state i_0 among the set of states S of the automation α. The automation α is in that state at the initial moment. The prediction of letter $x_t \in A$ was denoted as $P_\alpha(x_t/x_1 \cdots x_{t-1}; i)$, provided that the automation was in the state $i \in S$ at the initial moment and, then, that the word $x_1x_2 \cdots x_{t-1}$ is at the input. Note, that $P_\alpha(x_t/x_1x_2 \cdots x_{t-1}; i_0) = P_\alpha(x_t/x_1x_2 \cdots$

x_{t-1}). For $i \in S$ and $x_1x_2 \cdots x_t \in A^t$ we define the value of

$$P_\alpha(x_1x_2 \cdots x_t; i) = P_\alpha(x_1; i) \cdot P_\alpha(x_2/x_1; i) \cdots P_\alpha(x_t/x_1 \cdots x_{t-1}; i) \quad (21)$$

and we let

$$\pi_\alpha(x_1x_2 \cdots x_t) = \max\{P_\alpha(x_1x_2 \cdots x_t; i)\ i \in S\}. \quad (22)$$

Note at once that

$$\sum_{x \in A^t} \pi_\alpha(x) \sum_{x \in A^t} \left(\sum_{i \in S} P_\alpha(x; i) \right) = \sum_{i \in S} \sum_{x \in A^t} P_\alpha(x; i) \leq |S|.$$

Thus,

$$\sum_{x \in A^t} \pi_\alpha(x)/|S| \leq 1. \quad (23)$$

Now we take integers n, k, N such that $n = k \cdot N$ and evaluate $L_\alpha(x^{(n)}) - L_{\mathrm{p}}(x^{(n)})$. The following train of expressions is valid:

$$\begin{aligned}
&L_\alpha(x^{(n)}) - L_{\mathrm{p}}(x^{(n)}) \\
&\quad = \left(\log |A| + n^{-1} \sum_{i=1}^{n} \log P_\alpha(x_i/x_1 \cdots x_{i-1}) \right) \\
&\qquad - \left(\log |A| + n^{-1} \sum_{i=1}^{n} P_{\mathrm{p}}(x_i/x_1 \cdots x_{i-1}) \right) \\
&\quad = (k \cdot N)^{-1} \sum_{i=0}^{k-1} \log P_\alpha(x_{iN+1} \cdots x_{(i+1)N}) - (k \cdot N)^{-1} \log P_{\mathrm{p}}(x^{(n)}) \\
&\quad \leq N^{-1} \sum_{y \in A^N} \left(\frac{\nu_y(x)^{(n)}}{k} \right) \log \pi_\alpha(y) - \frac{1}{k \cdot N} \left[\log(\lambda(N) \mathrm{p}_N(x^{(n)})) \right. \\
&\qquad \left. + \log \left(1 + \sum_{i=0; i \neq N}^{\infty} (\lambda(i) \mathrm{p}_i(x^{(n)}) / (\lambda(N) \mathrm{p}_N(x^{(n)})) \right) \right] \\
&\quad \leq \frac{1}{N} \sum_{y \in A^N} \left(\frac{\nu_y(x^{(n)})}{k} \right) \log \pi_\alpha(y) - \frac{1}{kN} [\log \mathrm{p}_N(x^{(n)}) + \log \lambda(N)] \\
&\quad \leq \frac{1}{N} \sum_{y \in A^N} \left(\frac{\nu_y(x^{(n)})}{k} \right) \log \pi_\alpha(y) \\
&\qquad + \frac{1}{k} \left(k \hat{H}_{N+1}(x^{(n)}) + \log \lambda(N) + \frac{C(N) \log n}{N} \right)
\end{aligned}$$

$$= \frac{1}{N} \sum_{y \in A^N} \frac{\nu_y(x^{(n)})}{k} \log \left(\frac{\pi_\alpha(y)}{|S|} \right) + \hat{H}_{N+1}(x^{(n)}) + \frac{\log |S|}{N} + \frac{C(N) \log n}{k \cdot N}$$

$$+ \frac{\log \lambda(N)}{k \cdot N} = \frac{1}{N} \sum_{y \in A^N} \left(\frac{\nu_y(x^{(n)})}{k} \right) \times \log \left(\frac{\pi_\alpha(y)/|S|}{\nu_y(x^{(n)})/k} \right) + \frac{\log |S|}{N}$$

$$+ \frac{\log \lambda(N)}{k \cdot N} + \frac{C(N) \log(k \cdot N)}{k \cdot N} \leq N^{-1} \log |S| + (k \cdot N)^{-1} \log(\lambda(N)$$

$$+ (k \cdot N)^{-1}(C(N) \log(k \cdot N)).$$

Here, first two equalities follow from the definitions (2) and (21). The first inequality follows from (23), the second from the monotonicity of $\log x$, the third from (19). Equalities are evident. The last inequality follows from (20) and (23). By increasing N and k we may make the last sum infinitesimal. The theorem is proved.

4. The Efficiency of Algorithmic Prediction for the Classes of Random Processes

In this section we investigate the efficiency of method p that is optimal in the class of finite automata, using some sets of random processes. We establish a connection between the precision of prediction and the Shannon entropy of a random process. The results of this section are obtained in [18] and are presented without proof. Let us begin with definitions. We define stationary and ergodic processes on the set A^∞ in the usual way (see, for example, [2, 8]). Denote the set of all stationary processes on A^∞ and the set of ergodic processes by $E(A)$ and $M(A) \subset E(A)$, respectively. Denote by $M_0(A) \subset M(A)$ the set of Bernoullian processes and by $M_k(A) \subset M(A)$, $k \geq 1$, the set of Markovian processes of connectivity (memory) k. In other words, for $\xi \in M_k(A)$, $k \geq 0$, the equality

$$P\{\xi(t) = a_t / \xi(t-1) = a_{t-1}, \ldots, \xi(1) = a_1\}$$
$$= P\{\xi(t) = a_t / \xi(t-1) = a_{t-1}, \ldots, \xi(t-k) = a_{t-k}\}$$

holds for all integers $t \geq k$ and $a_1, a_2, \ldots, a_t \in A$. (Here, $\xi(t)$ is a random value equal to the value of the process at the moment t.) By $h(\xi)$ we denote the Shannon entropy of the process ξ for $\xi \in E(A)$ (see, for example, [2, 8] for the definition).

Theorem 4. *For any $k \geq 0$ and a random process $\xi \in M(A)$ with probability equal to* 1,

$$\lim_{t \to \infty} L_p(x_1 x_2 \cdots x_t) = \log |A| - h(\xi).$$

(The limit does exist with probability equal to 1.*) On the other hand, any prediction method* α *realizable on the Turing machine is asymptotically no better than* p*, so the following inequality is valid with probability equal to* 1*:*

$$\lim_{t\to\infty} L_\alpha(x_1x_2\cdots x_t) \le \log|A| - h(\xi).$$

The proof readily follows from the results obtained in [18].

The theorem presented shows that the prediction method "collects statistics" on the sequence $x_1x_2\cdots x_t$ efficiently enough and uses that information to increase the precision of prediction, attaining (in the limit) the maximal precision (and the maximal wealth in the game).

For the classes of Bernoulli and Markov processes, method p is also optimal in a somewhat stronger sense. We define the average precision for $\xi \in M(A)$ and the prediction method α by the equality

$$\tilde{L}_\alpha(\xi, t) = E_\xi(L_\alpha(x_1x_2\cdots x_t)),$$

where $t \ge 0$ is an integer and $E_\xi(\cdot)$ is an expectation over ξ. The following theorem is valid.

THEOREM 5. *The average precision of prediction according to the method* p *is maximal in the classes of Bernoullian and Markovian processes. More precisely, for* $k \ge 0$ *and* $\xi \in M_k(A)$,

$$\tilde{L}_p(\xi, t) \ge \log|A| - h(\xi)) - \frac{(|A|-1)|A|^k}{2t}\log t + O(1/t).$$

On the other hand, for the prediction according to any method α *that is realizable on a Turing machine, the following inequality is valid:*

$$\sup_{\xi\in M_k(A)} [\log A - h(\xi) - L_\alpha(\xi, t)] \ge \frac{(|A|-1)|A|^k}{2t}\log t + O(1/t).$$

The proof is also given in [18].

REFERENCES

1. ALGOET, P. H. (1986), Universal algorithms for gambling, data compression, and portfolio selection, *in* "IEEE International Symposium of Information Theory, Ann Arbor, Michigan, 1986," p. 76.
2. BILLINGSLEY, P. (1965), "Ergodic Theory and Information," Wiley, New York.
3. BARRON, A. R., AND COVER, T. M. (1991), A bound on the financial value of information, *IEEE Trans. Inform. Theory* **37**, No. 4, 1067–1071.

4. CLARKE, B. S., AND BARRON, A. R. (1990), Information-theoretic asymptotics of Bayes methods, *IEEE Trans. Inform. Theory* **36,** No. 3, 453–471.
5. COVER, T., AND THOMAS, J. (1991), "Elements of Information Theory," Wiley, New York.
6. FEDER, M. (1991), Gambling using a finite-state machine, *IEEE Trans. Inform. Theory* **37,** No. 5, 1459–1465.
7. FEDER, M., MERHAV, N., AND GUTMAN, M. (1992), Universal prediction of individual sequences, *IEEE Trans. Inform. Theory* **38,** No. 4, 1258–1270.
8. GALLAGER, R. G. (1968), "Information Theory and Reliable Communication," Wiley, New York, 1968.
9. KELLY, J. L. (1956), A new interpretation of information rate, *Bell System Tech. J.* **35,** 917–926.
10. KOLMOGOROV, A. N. (1965), Three approaches to the concept of the amount of information, *Probl. Inform. Trans.* **1,** No. 1, 3–7. [Russian]
11. KOLMOGOROV, A. N. (1983), On logical foundation of probability theory, *in* "Proceedings, 4th USSR–Japan Symposium on Probability Theory and Statistics," Lect. Notes in Math., Vol. 1021, Springer-Verlag, New York/Berlin.
12. KRAMOSIL, I. (1985), Non-deterministic prediction based on algorithm complexity, *Probl. Control Inform. Theory* **14,** 461–476.
13. KRICHEVSKY, R. E., AND TROFIMOV, V. K. (1980), Optimal sample-based encoding sources, *in* "Third Czechoslovak–Soviet–Hungarian Seminar on Information Theory, Lublice, June, 23–27," pp. 131–138.
14. LEVENSTEIN, V. I. (1968), The redundancy and deceleration of a separative encoding of natural numbers, *in* "Problems of Cybernetics, No. 20, Moscow," pp. 173–179. [Russian]
15. RISSANEN, J. (1984), Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory* **30,** 629–636.
16. RYABKO, B. YA. (1984), Coding of combinatorial sources and Hausdorff dimention, *Soviet Math. Dokl.* **30,** No. 1, 219–222.
17. RYABKO, B. YA. (1986), Coding of combinatorial sources, Hausdorff dimention and Kolmogorov complexity, *Probl. Inform. Trans.* **22,** No. 3, 16–26. [Russian]
18. RYABKO, B. YA. (1988). The prediction of the random sequences and universal coding, *Probl. Inform. Trans.* **24,** No. 2, 18–26. [Russian]
19. SOLOMONOFF, R. I. (1964), A formal theory of inductive inference, *Inform. and Control* **7,** 1–22.
20. SCHNORR, C. P. (1971), A unified approach to the definition of random sequences, *Math. Systems Theory* **5,** No. 3, 246–258.
21. ZIV, J., AND LEMPEL, A. (1978), Compression of individual sequences via variable rate coding, *IEEE Trans. Inform. Theory* **24,** No. 5, 530–536.
22. ZVONKIN, A. K., AND LEVIN, L. A. (1970), The complexity of finite objects and the algorithmic concepts of information and randomness, *Russian Math. Surveys* **25,** No. 6, 83–124.