Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series

Boris Ryabko
Institute of Computational Technologies of Siberian Branch of Russian Academy of Science
Siberian State University of Telecommunications and Informatics, Novosibirsk, Russia
boris@ryabko.net

***Abstract*— We address the problem of nonparametric estimation of characteristics for stationary and ergodic time series. We consider finite-alphabet time series and the real-valued ones and the following problems: estimation of the (limiting) probability $P(u_0 \dots u_s)$ for every s and each sequence $u_0 \cdots u_s$ of letters from the process alphabet (or estimation of the density $p(x_0, \dots, x_s)$ for real-valued time series), so-called on-line prediction, where the conditional probability $P(x_{t+1}/x_1 x_2 \dots x_t)$ (or the conditional density $p(x_{t+1}/x_1 x_2 \dots x_t)$) should be estimated (in the case where $x_1 x_2 \cdots x_t$ is known), regression and classification (or so-called problems with side information). We show that any universal code (or a universal data compressor) can be used as a basis for constructing asymptotically optimal methods for the above problems.**

I. Introduction

J. Rissanen [23], [24], [25] has discovered some deep connections between universal coding (or universal data compression) and mathematical statistics. In this paper we apply this approach to some statistical problems concerned with time series. We address the problem of nonparametric estimation of characteristics of stationary and ergodic time series.

We consider a stationary and ergodic source, which generates sequences $x_1 x_2 \cdots$ of elements (letters) from some set (alphabet) A, which is either finite or real-valued. Of course, if someone knows the probability distribution (or the density) he has all information about the source and can solve all problems in the best way. Hence, generally speaking, precise estimations of the probability distribution and the density can be used for prediction, regression estimation, etc. In this paper we follow the scheme: we consider the problems of estimation of the probability distribution or the density estimation. Then we show how the solution can be applied to other problems, paying the main attention to the problem of prediction, because of its practical applications and importance for probability theory, information theory, statistics and other theoretical sciences, see [1], [12], [13], [15], [19], [20], [25], [36]. We show that universal codes (or data compressors) can be applied directly to the problems of estimation, prediction, regression and classification. It is not surprising, because for any stationary and ergodic source p generating letters from a finite alphabet and any universal code U the following equality is valid with probability 1:

$$\lim_{t\to\infty} \frac{1}{t}(-\log p(x_1 \cdots x_t) - |U(x_1 \cdots x_t)|) \;=\; 0,$$

where $x_1 \cdots x_t$ is generated by p. (Here and below $\log = \log_2$, $|v|$ is the length of v, if v is a word and the number of elements of v if v is a set.) So, in fact, the length of the universal code ($|U(x_1 \cdots x_t)|$) can be used as an estimate of the logarithm of the unknown probability and, obviously, $2^{-|U(x_1 \cdots x_t)|}$ can be considered as the estimation of $p(x_1 \cdots x_t)$. In fact, a universal code can be viewed as a non-parametrical estimation of (limiting) probabilities for stationary and ergodic sources. This was recognized shortly after the discovery of universal codes (for the set of stationary and ergodic processes with finite alphabets [26]) and universal codes were applied for solving prediction problem [27].

We would like to emphasize that, on the one hand, all results are obtained in the framework of classical probability theory and mathematical statistics and, on the other hand, everyday methods of data compression (or archivers) can be used as a tool for density estimation, prediction and other problems, because they are practical realizations of universal codes. It is worth noting that the modern data compressors (like *zip*, *arj, rar,* etc.) are based on deep theoretical results of the theory of source coding (see, for ex., [10], [16], [18], [25], [34]) and have been demonstrated high efficiency in practice as compressors of texts, DNA sequences and many other types of real data. In fact, archivers can find many kinds of latent regularities, that is why they look like a promising tool for prediction and other problems. Moreover, recently universal codes and archivers were efficiently applied to some problems which are very far from data compression: first, their applications in [5], [6] created a new and rapidly growing line of investigation in clustering and classification and, second, universal codes were used as a basis for non-parametric tests for the main statistical hypotheses concerned with stationary and ergodic time series [30], [31].

Proofs of the theorems can be found in the extended version of the paper in $arxiv.org$, $cs.IT/0701036$.

II. Predictors and universal data compressors

We consider a source with unknown statistics which generates sequences $x_1 x_2 \cdots$ of letters from some set (or alphabet)

978-1-4244-2271-5/08/$25.00 ©2008 IEEE

A. It will be convenient at first to describe briefly the prediction problem. This problem can be traced back to Laplace. He suggested the following predictor:

$$L_0(a|x_1 \cdots x_t) = (\nu_{x_1 \cdots x_t}(a) + 1)/(t + |A|), \qquad (1)$$

where $\nu_{x_1 \cdots x_t}(a)$ denote the count of letter a occurring in the word $x_1 \ldots x_{t-1} x_t$. For example, if $A = \{0, 1\}$, $x_1 ... x_5 = 01010$, then the Laplace prediction is as follows: $L_0(x_6 = 0|01010) = (3+1)/(5+2) = 4/7, L_0(x_6 = 1|01010) = (2+1)/(5+2) = 3/7$. In other words, $3/7$ and $4/7$ are estimations of the unknown probabilities $P(x_{t+1} = 0|x_1 \ldots x_t = 01010)$ and $P(x_{t+1} = 1|x_1 \ldots x_t = 01010)$.

We can see that Laplace considered prediction as a set of estimations of unknown (conditional) probabilities. This approach to the problem of prediction was developed in [27] and now is often called on-line prediction or universal prediction [1], [12], [20]. As we mentioned above, it seems natural to consider conditional probabilities to be the best prediction, because they contain all information about the future behavior of the stochastic process. Moreover, this approach is deeply connected with game-theoretical interpretation of prediction (see [14], [29]) and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor γ defines a measure by following equation

$$\gamma(x_1 ... x_t) = \prod_{i=1}^{t} \gamma(x_i|x_1 ... x_{i-1}). \qquad (2)$$

For example, $L_0(0101) = \frac{1}{2}\frac{1}{3}\frac{1}{2}\frac{2}{5} = \frac{1}{30}$. And, vice versa, any measure γ defines a predictor: $\gamma(x_i|x_1 ... x_{i-1}) = \gamma(x_1 ... x_{i-1} x_i)/\gamma(x_1 ... x_{i-1})$. The same is true for a density: a predictor is defined by conditional density and, vice versa, the density is equal to the product of conditional densities:

$$p(x_i|x_1 ... x_{i-1}) = p(x_1 ... x_{i-1} x_i)/p(x_1 ... x_{i-1}),$$

$$p(x_1 ... x_t) = \prod_{i=1}^{t} p(x_i|x_1 ... x_{i-1}).$$

The next natural question is how to estimate the precision or of the prediction and an estimation of probability. Mainly we will estimate the error of prediction by the Kullback-Leibler (KL) divergence between a distribution p and its estimation. Consider an (unknown) source p and some predictor γ. The *error* is characterized by the KL divergence

$$\rho_{\gamma,p}(x_1 \cdots x_t) = \sum_{a \in A} p(a|x_1 \cdots x_t) \log \frac{p(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}. \qquad (3)$$

It is well-known that for any distributions p and γ the K-L divergence is nonnegative and equals 0 if and only if $p(a) = \gamma(a)$ for all a, see, for ex., [11]. The following inequality (Pinsker's inequality)

$$\sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} ||P - Q||^2. \qquad (4)$$

connects the KL divergence with a so-called variation distance

$$||P - Q|| = \sum_{a \in A} |P(a) - Q(a)|,$$

where P and Q are distributions over A, see [7]. For fixed t, $\rho_{\gamma,p}(\,)$ is a random variable, because $x_1, x_2, \cdots, x_t$ are random variables. We define the average error at time t by

$$\rho^t(p\|\gamma) = E\ (\rho_{\gamma,p}(\cdot)) = \sum_{x_1 \cdots x_t \in A^t} p(x_1 \cdots x_t)\ \rho_{\gamma,p}(x_1 \cdots x_t). \qquad (5)$$

The following inequality shows that the error of Laplace predictor L_0 goes to 0 for any i.i.d. source p :

$$\rho^t(p\|L_0) < (|A| - 1)/(t + 1) \qquad (6)$$

([28]; see also [32]). So, we can see from this inequality that the average error of the Laplace predictor L_0 (estimated either by the KL divergence or the variation distance) goes to zero for any unknown i.i.d. source, when the sample size t grows. Moreover, it can be easily shown that the error (3) (and the corresponding variation distance) goes to zero with probability 1, when t goes to infinity. Obviously, such a property is very desirable for any predictor and for larger classes of sources, like Markov, stationary and ergodic, etc. However, it is proven in [27] (see also [1], [12], [20]) that such predictors do not exist for the class of all stationary and ergodic sources (generated letters from a given finite alphabet). More precisely, for any predictor γ there exists a source p and $\delta > 0$ such that with probability 1 $\rho_{\gamma,p}(x_1 \cdots x_t) \geq \delta$ infinitely often when $t \to \infty$. So, the error of any predictor does not go to 0, if the predictor is applied to all stationary and ergodic sources, that is why it is difficult to use (3) and (5) for comparison of different predictors.

On the other hand, it is shown in [27] that there exists a predictor R, such that the following Cesaro average $t^{-1} \sum_{i=1}^{t} \rho_{R,p}(x_1 \cdots x_t)$ goes to 0 (with probability 1) for any stationary and ergodic source p, where t goes to infinity. That is why we will focus our attention on such averages and by analogy with (5) we define

$$\bar{\rho}_{\gamma,p}(x_1 ... x_t) = t^{-1} (\log(p(x_1 ... x_t)/\gamma(x_1 ... x_t)), \qquad (7)$$

$$\bar{\rho}_t(\gamma, p) = t^{-1} \sum_{x_1 ... x_t \in A^t} p(x_1 ... x_t) \log \frac{p(x_1 ... x_t)}{\gamma(x_1 ... x_t)}, \qquad (8)$$

where, as before, $\gamma(x_1 ... x_t) = \prod_{i=1}^{t} \gamma(x_i|x_1 ... x_{i-1})$.

From these definitions and (6) we obtain the following estimation of the error of the Laplace predictor L_0 for any i.i.d. source:

$$\bar{\rho}_t(L_0, p) < ((|A| - 1)\ \log t + c)/t, \qquad (9)$$

where c is a certain constant. So, we can see that the average error of the Laplace predictor goes to zero for any i.i.d. source (which generates letters from a known finite alphabet). As a matter of fact, the Laplace probability $L_0(x_1 ... x_t)$ is a consistent estimate of the unknown probability $p(x_1 ... x_t)$. The natural problem is to find a predictor whose error is minimal

(for i.i.d. sources). This problem was considered and solved by Krichevsky [17], [18], see also [37], [38]. He suggested the following predictor:

$$K_0(a|x_1\cdots x_t) = (\nu_{x_1\cdots x_t}(a) + 1/2)/(t + |A|/2), \quad (10)$$

where, as before, $\nu_{x_1\cdots x_t}(a)$ denote the count of letter a occurring in the word $x_1 \ldots x_t$. We can see that the Krychevsky predictor is quite close to the Laplace's one (1). For example, if $A = \{0, 1\}$, $x_1...x_5 = 01010$, then $K_0(x_6 = 0|01010) = (3+1/2)/(5+1) = 7/12, K_0(x_6 = 1|01010) = (2+1/2)/(5+1) = 5/12$ and $K_0(01010) = \frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8}\frac{1}{2} = \frac{3}{256}$.

The Krichevsky measure K_0 can be presented as follows:

$$K_0(x_1...x_t) = \frac{\prod_{a\in A}(\prod_{j=1}^{\nu_{x_1\cdots x_t}(a)}(j-1/2))}{\prod_{i=0}^{t-1}(i+|A|/2)}. \quad (11)$$

It is known that $(r+1/2)((r+1)+1/2)...(s-1/2) = \frac{\Gamma(s+1/2)}{\Gamma(r+1/2)}$, where $\Gamma(\,)$ is the gamma function. So, (11) can be presented as follows:

$$K_0(x_1...x_t) = \frac{\prod_{a\in A}(\Gamma(\nu_{x_1\ldots x_t}(a)+1/2)\,/\Gamma(1/2)\,)}{\Gamma(t+|A|/2)\,/\Gamma(|A|/2)}. \quad (12)$$

For this predictor $\bar{\rho}_t(K_0, p) < ((|A|-1)\,\log t + c)/(2t)$, where c is a constant, and, moreover, in a certain sense this average error is minimal: for any predictor γ there exists such a source p^* that $\bar{\rho}_t(\gamma, p^*) \geq ((|A|-1)\,\log t + c)/(2t)$, see [17], [18].

Now we briefly describe consistent estimations of unknown probabilities and efficient on-line predictors for general stochastic processes (or sources of information). Denote by A^t and A^* the set of all words of length t over A and the set of all finite words over A correspondingly ($A^* = \bigcup_{i=1}^{\infty} A^i$). By $M_\infty(A)$ we denote the set of all stationary and ergodic sources, which generate letters from A and let $M_0(A) \subset M_\infty(A)$ be the set of all i.i.d. processes. Let $M_m(A) \subset M_\infty(A)$ be the set of Markov sources of order (or with memory, or connectivity) not larger than m, $m \geq 0$. Let $M^*(A) = \bigcup_{i=0}^{\infty} M_i(A)$ be the set of all finite-order sources.

The Laplace and Krichevsky predictors can be extended to general Markov processes. The trick is to view a Markov source $p \in M_m(A)$ as resulting from $|A|^m$ i.i.d. sources. We illustrate this idea by an example from [32]. So assume that $A = \{O, I\}$, $m = 2$ and assume that the source $p \in M_2(A)$ has generated the sequence

$$OOIOIIOOIIIOIO.$$

We represent this sequence by the following four subsequences:

$$**I*****I*****,$$

$$***O*I***I***O,$$

$$****I**O****I*,$$

$$******O***IO**.$$

These four subsequences contain letters which follow OO, OI, IO and II, respectively. By definition, $p \in M_m(A)$ if $p(a|x_1\cdots x_t) = p(a|x_{t-m+1}\cdots x_t)$, for all $0 < m \leq t$, all $a \in A$ and all $x_1\cdots x_t \in A^t$. Therefore, each of the four generated subsequences may be considered to be generated by a Bernoulli source. Further, it is possible to reconstruct the original sequence if we know the four ($= |A|^m$) subsequences and the two ($= m$) first letters of the original sequence.

Any predictor γ for i.i.d. sources can be applied for Markov sources. Indeed, in order to predict, it is enough to store in the memory $|A|^m$ sequences, one corresponding to each word in A^m. Thus, in the example, the letter x_3 which follows OO is predicted based on the Bernoulli method γ corresponding to the x_1x_2- subsequence ($= OO$), then x_4 is predicted based on the Bernoulli method corresponding to x_2x_3, i.e. to the OI- subsequence, and so forth. When this scheme is applied along with either L_0 or K_0 we denote the obtained predictors as L_m and K_m, correspondingly and define the probabilities for the first m letters as follows: $L_m(x_1) = L_m(x_2) = \ldots = L_m(x_m) = 1/|A|$, $K_m(x_1) = K_m(x_2) = \ldots = K_m(x_m) = 1/|A|$. For example, having taken into account (12), we can present the Krichevsky predictors for $M_m(A)$ as follows:

$$K_m(x_1...x_t) =$$

$$\begin{cases} \frac{1}{|A|^t}, & \text{if } t \leq m, \\ \frac{1}{|A|^m}\prod_{v\in A^m}\frac{\prod_{a\in A}((\Gamma(\nu_x(va)+1/2)\,/\,\Gamma(1/2))}{(\Gamma(\bar{\nu}_x(v)+|A|/2)\,/\,\Gamma(|A|/2))}, & \text{if } t > m \end{cases}, \quad (13)$$

where $\bar{\nu}_x(v) = \sum_{a\in A}\nu_x(va)$, $x = x_1...x_t$. It is worth noting that the representation (11) can be more convenient for carrying out calculations. Let us consider an example. For the word $OOIOIIOOIIIOIO$ considered in the previous example, we obtain $K_2(OOIOIIOOIIIOIO) = 2^{-2}\ \frac{1}{2}\frac{3}{4}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}$.

Let us define the measure R, which, in fact, is a consistent estimator of probabilities for the class of all stationary and ergodic processes with a finite alphabet. First we define a probability distribution $\{\omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_i = 1/\log(i+1) - 1/\log(i+2),\ i = 1, 2, \ldots. \quad (14)$$

(In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) The measure R is defined as follows:

$$R(x_1...x_t) = \sum_{i=0}^{\infty}\omega_{i+1}\,K_i(x_1...x_t). \quad (15)$$

It is worth noting that this construction can be applied to the Laplace measure (if we use L_i instead of K_i) and any other family of measures.

The main properties of the measure R are connected with the Shannon entropy, which is defined as follows

$$H(p) = \lim_{m\to\infty} -\frac{1}{m}\sum_{v\in A^m} p(v)\log p(v). \quad (16)$$

Theorem 1. [27]. *For any stationary and ergodic source p the following equalities are valid: i) $\lim_{t\to\infty}\frac{1}{t}\log(1/R(x_1\cdots x_t)) = H(p)$ with probability 1, ii) $\lim_{t\to\infty}\frac{1}{t}\sum_{u\in A^t} p(u)\log(1/R(u)) = H(p)$.*

Now we consider universal codes. By definition, a code U is universal if for any stationary and ergodic source p the following equalities are valid:

$$\lim_{t\to\infty} |U(x_1\dots x_t)|/t = H(p) \tag{17}$$

with probability 1, and

$$\lim_{t\to\infty} E(|U(x_1\dots x_t)|)/t = H(p), \tag{18}$$

where $H(p)$ is the Shannon entropy of p, $E(f)$ is a mean value of f.

III. Finite-alphabet processes

A. The estimation of (limiting) probabilities

The following theorem shows how universal codes can be applied for probability estimations.

Theorem 2. *Let U be a universal code and*

$$\mu_U(u) = 2^{-|U(u)|}/\Sigma_{v\in A^{|u|}}\, 2^{-|U(v)|}. \tag{19}$$

Then, for any stationary and ergodic source p the following equalities are valid: i) $\lim_{t\to\infty} \frac{1}{t}(-\log p(x_1\cdots x_t) - (-\log \mu_U(x_1\cdots x_t))) = 0$ with probability 1, ii) $\lim_{t\to\infty} \frac{1}{t}\sum_{u\in A^t} p(u)\log(p(u)/\mu_U(u)) = 0$, iii) $\lim_{t\to\infty} \frac{1}{t}\sum_{u\in A^t} p(u)\,|p(u)-\mu_U(u)| = 0$.

So, we can see that, in a certain sense, the measure μ_U is a consistent (nonparametric) estimation of the (unknown) measure p.

Nowadays there are many efficient universal codes (and universal predictors connected with them), see [13], [15], [21], [25], [27], [34], which can be applied to estimation. For example, the above described measure R is based on the code from [26], [27] and can be applied for probability estimation. More precisely, Theorem 2 (and the following theorems) are true for R, if we replace μ_U by R.

It is important to note that the measure R has some additional properties, which can be useful for applications. The following theorem will be devoted to description of these properties (whereas all other theorems are valid for all universal codes and corresponding them measures, including the measure R).

Theorem 3. [27], [28]. *For any Markov process p with memory k*

i) the error of the probability estimator, which is based on the measure R, is upper-bounded as follows: $\frac{1}{t}\sum_{u\in A^t} p(u)\log(p(u)/R(u)) \le \frac{(|A|-1)|A|^k\log t}{2t} + O(\frac{1}{t})$,

ii) in a certain sense the error of R is asymptotically minimal: for any measure μ there exists a $k-$memory Markov process p_μ such that $\frac{1}{t}\sum_{u\in A^t} p_\mu(u)\log(p_\mu(u)/\mu(u)) \ge \frac{(|A|-1)|A|^k\log t}{2t} + O(\frac{1}{t})$,

iii) *Let Θ be such a set of stationary and ergodic processes that there exists a measure μ_Θ for which the estimation error of the probability goes to 0 uniformly: $\lim_{t\to\infty} \sup_{p\in\Theta} (\frac{1}{t}\sum_{u\in A^t} p(u)\log(p(u)/\mu_\Theta(u))) = 0$. Then the error of estimator, which is based on the measure R, goes to 0 uniformly, too: $\lim_{t\to\infty} \sup_{p\in\Theta} (\frac{1}{t}\sum_{u\in A^t} p(u)\log(p(u)/R(u))) = 0$.*

B. Prediction

As we mentioned above, any universal code U can be applied for prediction. Namely, the measure μ_U (19) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1...x_t) = \mu_U(x_1...x_tx_{t+1})/\mu_U(x_1...x_t). \tag{20}$$

Theorem 4. *Let U be a universal code and p be any stationary and ergodic process. Then $i)$ $\lim_{t\to\infty}\frac{1}{t}\{E(\log\frac{p(x_1)}{\mu_U(x_1)}) + E(\log\frac{p(x_2|x_1)}{\mu_U(x_2|x_1)}) + \dots + E(\log\frac{p(x_t|x_1\dots x_{t-1})}{\mu_U(x_t|x_1\dots x_{t-1})})\} = 0$,*

$$ii)\ \lim_{t\to\infty} E(\frac{1}{t}\sum_{i=0}^{t-1}(p(x_{i+1}|x_1...x_i)-\mu_U(x_{i+1}|x_1...x_i))^2) = 0\,,$$

$$iii)\ \lim_{t\to\infty} E(\frac{1}{t}\sum_{i=0}^{t-1}|p(x_{i+1}|x_1...x_i)-\mu_U(x_{i+1}|x_1...x_i)|) = 0\,.$$

Comment. In fact, the statements ii) and iii) are equivalent, because one of them follows from the other. For details see Lemma 2 in [33].

The above-described measure R has one additional property, if it is used for prediction.

Theorem 5 . ([28]) *for any Markov process p ($p\in M^*(A)$) the following is true: $\lim_{t\to\infty}\log\frac{p(x_{t+1}|x_1\dots x_t)}{R(x_{t+1}|x_1\dots x_t)} = 0$ with probability 1, where $R(x_{t+1}|x_1...x_t) = R(x_1...x_tx_{t+1})/R(x_1...x_t)$.*

IV. Real-valued time series

Let X_t be a time series with each X_t taking values in some interval Λ. The probability distribution of X_t is unknown but it is known that the time series is stationary and ergodic. Let $\{\Pi_n\}, n\ge 1$, be an increasing sequence of finite partitions that asymptotically generates the Borel sigma-field on Λ, and let $x^{[k]}$ denote the element of Π_k that contains the point x. (Informally, $x^{[k]}$ is obtained by quantizing x to k bits of precision.) Suppose that the joint distribution P_n for $(X_1,\dots,X_n)$ has a probability density function $p_n(x_1,\dots,x_n)$ with respect to a sigma-finite measure λ_n. (For example, λ_n can be Lebesgue measure, counting measure, etc.) For integers s and n we define the following approximation of the density $p^s(x_1,\dots,x_n) = P(x_1^{[s]},\dots,x_n^{[s]})/\lambda_n(x_1^{[s]}\dots x_n^{[s]})$. Let $p(x_{n+1}|x_1,\dots,x_n)$ denote the conditional density given by the ratio $p(x_1,\dots,x_{n+1})/p(x_1,\dots,x_n)$ for $n>1$. It is known that for stationary and ergodic processes there exists a so- called relative entropy rate h defined by $h=\lim_{n\to\infty} E(\log p(x_{n+1}|x_1,\dots,x_n))$, where E denotes expectation with respect to P; see [3]. We also consider $h_s = \lim_{n\to\infty} E(\log p^s(x_{n+1}|x_1,\dots,x_n))$.

It is shown by Barron [3] that almost surely $\lim_{t\to\infty}\frac{1}{t}\log p(x_1\dots x_t) = h$. Applying the same theorem to the density $p^s(x_1,\dots,x_t)$, we obtain that a.s. $\lim_{t\to\infty}\frac{1}{t}\log p^s(x_1,\dots,x_t) = h_s$.

Let U be a universal code, which is defined for any finite alphabet. We define the corresponding density r_U as follows:

$$r_U(x_1\dots x_t) = \sum_{i=0}^{\infty}\omega_i 2^{-|U(x_1^{[i]}\dots x_t^{[i]})|}/\lambda_t(x_1^{[i]}\dots x_t^{[i]})\,. \tag{21}$$

(It is supposed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters.)

It turns out that, in a certain sense, the density $r_U(x_1 \dots x_t)$ estimates a unknown density $p(x_1, \dots, x_t)$.

Theorem 6. *Let X_t be a stationary ergodic process with densities $p(x_1 \dots x_t) = dP_t/d\lambda_t$ such that*

$$\lim_{s\to\infty} h_s = h < \infty, \tag{22}$$

where h and h_s are relative entropy rates. Then $\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} = 0$ with probability 1 and $\lim_{t\to\infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) = 0$.

The following theorem describes properties of conditional probabilities $r_U(x|x_1 \dots x_m) = r_U(x_1 \dots x_m x)/r_U(x_1 \dots x_m)$ which, in turn, is connected with the prediction problem. We will see that the conditional density $r_U(x|x_1 \dots x_m)$ is a reasonable estimation of $p(x|x_1 \dots x_m)$.

Theorem 7. *Let $B_1, B_2, \dots$ be a sequence of measurable sets. Then the following equalities are true:* i) $\lim_{t\to\infty} E(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m))^2) = 0$, *ii)* $E(\frac{1}{t} \sum_{m=0}^{t-1} |P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m))| = 0$.

We have seen that in a certain sense the estimation r_U approximates the density p. The following theorem shows that r_U can be used instead of p for estimation of average values of certain functions.

Theorem 8. *Let f be an integrable function, whose absolute value is bounded by a certain constant M. Then the following equalities are valid:*

i) $\lim_{t\to\infty} \frac{1}{t} E(\sum_{m=0}^{t-1} (\int f(x) p(x|x_1 \dots x_m) d\lambda_m - \int f(x)\, r_U(x|x_1 \dots x_m)\, d\lambda_m)^2) = 0$,

ii) $\lim_{t\to\infty} \frac{1}{t} E(\sum_{m=0}^{t-1} |\int f(x) p(x|x_1 \dots x_m) d\lambda_m - \int f(x) r_U(x|x_1 \dots x_m) d\lambda_m|) = 0$.

ACKNOWLEDGMENT

Research was supported by Russian Foundation for Basic Research (grant no. 06-07-89025).

REFERENCES

[1] P. Algoet. "Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information," *IEEE Trans. Inform. Theory,* v. 45, 1999, pp. 1165-1185.

[2] A. R. Barron *Monotonic central limit theorem for densities,* Department of Statistics Technical Report n. 50, Stanford University, Stanford, California, 1984.

[3] A.R. Barron , "The strong ergodic theorem for dencities: generalized Shannon-McMillan-Breiman theorem," *The Annals of Probability,* v.13, n.4, 1985, pp. 1292-1303.

[4] P. Billingsley, *Ergodic theory and information,* John Wiley & Sons, 1965.

[5] R. Cilibrasi, P.M.B. Vitanyi, "Clustering by Compression," *IEEE Transactions on Information Theory,* v. 51, n.4. 2005.

[6] R. Cilibrasi, R. de Wolf , P.M.B. Vitanyi , "Algorithmic Clustering of Music," *Computer Music Journal,* v. 28, n. 4, 2004, pp. 49-67.

[7] I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Budapesht, Akadémiai Kiadó, 1981.

[8] I. Csiszár, P. Shields, "The consistency of the BIC Markov order estimation," *Annals of Statistics,* v. 6, 2000, pp. 1601-1619.

[9] G.A. Darbellay, I. Vajda, "Estimatin of the mutual information with data-dependent partitions," *IEEE Trans. Inform. Theory,* v. 48, 1999, pp. 1061-1081.

[10] M. Effros, K. Visweswariah, S.R. Kulkarni, S. Verdu, "Universal lossless source coding with the Burrows Wheeler transform," *IEEE Trans. Inform. Theory,* v. 45, 1999, pp. 1315-1321.

[11] R.G. Gallager , *Information Theory and Reliable Communication,* John Wiley & Sons, New York, 1968.

[12] L.Gyorfi, G. Morvai, S.J. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on nformation Theory,* v. 44, n. 2, 1998, pp. 886 - 892.

[13] P. Jacquet, W. Szpankowski, L. Apostol "Universal predictor based on pattern matching," *IEEE Trans. Inform. Theory,* v.48, 2002, pp. 1462-1472.

[14] J.L. Kelly, "A new interpretation of information rate," *Bell System Tech. J.,* v. 35, 1956, pp. 917-926.

[15] J. Kieffer, *Prediction and Information Theory,* Preprint,1998. (available at ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/)

[16] J.C. Kieffer, En-Hui Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Transactions on Information Theory,* v. 46, 2000, pp. 737 - 754.

[17] R. Krichevsky "A relation between the plausibility of information about a source and encoding redundancy," *Problems Inform. Transmission,* v.4, n.3, 1968, pp. 48-57.

[18] R. Krichevsky *Universal Compression and Retrival*. Kluver Academic Publishers, 1993.

[19] D.S. Modha, E. Masry "Memory-universal prediction of stationary random processes," *IEEE Trans. Inform. Theory,* v. 44, 1998, pp.117-133.

[20] G. Morvai, S.J. Yakowitz, P.H. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Inform. Theory,* v. 43, 1997, pp. 483 - 498.

[21] A.B. Nobel, "On optimal sequential prediction", *IEEE Trans. Inform. Theory,* v. 49, 2003, pp. 83-98.

[22] D.S. Ornstein, B.Weiss, "How sampling reveals a process," The Annals of Probability, v.18, n.3, 1990, pp.905-930.

[23] J. Rissanen, *Stochastic Complexity in Statistical Inquiry Theory.* World Scientific Publishing Co., Singapore, 1989.

[24] J. Rissanen , "Modeling by shortest data description", *Automatica,* v.14, 1978, pp. 465-471.

[25] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory,* 30, n.4, 1984, pp. 629-636.

[26] B.Ya. Ryabko, "Twice-universal coding," *Problems of Information Transmission,* v. 20, n.3, 1984, pp. 173-177.

[27] B.Ya. Ryabko, "Prediction of random sequences and universal coding," *Problems of Inform. Transmission,* v. 24, n.2, 1988, pp. 87-96.

[28] B.Ya. Ryabko, "A fast adaptive coding algorithm," *Problems of Inform. Transmission,* v. 26, n.4, 1990, pp. 305–317.

[29] B. Ya. Ryabko, "The complexity and effectiveness of prediction algorithms," *J. Complexity,* v.10, 1994, pp. 281–295.

[30] B. Ryabko, J. Astola. "Universal Codes as a Basis for Time Series Testing," *Statistical Methodology,* v.3, 2006, pp.375-397.

[31] B. Ya. Ryabko, V.A. Monarev. "Using information theory approach to randomness testing," *Journal of Statistical Planning and Inference,* , v. 133, 2005, pp. 95-110.

[32] B. Ryabko, F. Topsoe, "On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources," *Journal of Complexity,* v. 18, 2002, pp. 224-241.

[33] D. Ryabko, M. Hutter, "Sequence prediction for non-stationary processes." In proceedings: IEEE International Symposium on Information Theory (ISIT 2007), 2007, pp. 2346-2350.
see also http://arxiv.org/pdf/cs.LG/0606077

[34] S. A. Savari, "A probabilistic approach to some asymptotics in noiseless communication," *IEEE Transactions on Information Theory,* v. 46, 2000, pp. 1246-1262.

[35] P.C.Shields, "The interactions between ergodic theory and information theory," *IEEE Transactions on Information Theory,* v. 44, 1998, pp. 2079 - 2093.

[36] W. Szpankowsky. *Average case analysis of algorithms on sequences.* John Wiley and Sons, New York, 2001.

[37] Q. Xie, A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory,* v.43, 1997, pp.646-657.

[38] Q. Xie, A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transaction on Information Theory,* v.46, 2000, pp.431-445.