# Compression-based methods for nonparametric prediction and estimation of some characteristics of time series

Boris Ryabko
Institute of Computational Technology 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 on-line prediction for time series. We show that any universal code (or a universal data compressor) can be used as a basis for constructing asymptotically optimal methods for this problem for a certain class of stationary and ergodic processes.**

***Index Terms*– density estimation, prediction of random processes, source coding, stationary ergodic source, universal coding.**

## I. Introduction

Since C.Shannon published his famous paper "A mathematical theory of communication" [40] the ideas and results of Information Theory and, particularly, the theory of source coding have begun to play an important role in mathematical statistics, see [16], [17], [20], [24], [30], [31].

Nowadays the most known practical application of source coding is archivers, which have shown their high efficiency as compressors of real data. It is perhaps less known that methods of data compression and, especially, universal codes play an important role in hypothesis testing [35], [36] and prediction of time series [37]. Moreover, their applications in [7], [8] created a new rapidly growing line of investigation in clustering and classification.

In this paper we show that any universal code can be applied to nonparametric prediction and some related problems for a certain class of stationary and ergodic time series. It is important to note that the problems of prediction and estimation of characteristics of time series has attracted attention of many researchers, see [1], [2], [4], [18], [20], [25], [26], [28], [30], [43].

We consider finite-alphabet and real-valued time series and the following problems: i) prediction, ii) estimation of the limiting probabilities for finite-alphabet time series and iii) estimation of the density for real-valued time series. We use a so-called on-line prediction model (or dynamic forecasting) which was suggested in [33] and is quite popular now [2], [18], [20], [25], [26], [28]. According to this model there is a stationary ergodic process $X_1, X_2, \dots$ with unknown limiting probabilities $P(X_1 \dots X_n)$, $n \geq 1$, for the case of finite-alphabet time series, or an unknown density function $p(x_1 \dots x_n)$, which is assumed to exist for all $n \geq 1$, for the case of real-valued time series.

The prediction problem is as follows: at the time $t$ we are given a realization of the process $X_1, X_2, \dots, X_t$ and have to estimate the conditional probabilities $P(x_{t+1}|x_1 \dots x_t)$ (or the density $p(x_{t+1}|x_1 \dots x_t)$). The point is that if one knows these probabilities (or the density), one has all the information about $X_{t+1}$, that is why the problem of estimation of these probabilities (or the density) is a fundamental problem of time series analysis. Clearly, the more precise is the estimation, the better is the prediction. At the next time instant $t+1$ we have to estimate the probabilities $P(x_{t+2}|x_1 \dots x_{t+1})$ (or the density $p(x_{t+2}|x_1 \dots x_{t+1})$), etc. It is shown in [33] (see also [18]) that for any predictor there exists a stationary and ergodic source such that the error $|P^*(x_{t+1}|x_1 \dots x_t) - P(x_{t+1}|x_1 \dots x_t)|$ (or $|p^*(x_{t+1}|x_1 \dots x_t) - p(x_{t+1}|x_1 \dots x_t)|$, correspondingly) does not go to 0, when the length of observed sequence $t$ goes to infinity. (Here $P^*(\,)$ and $p^*(\,)$ are the estimations.) More precisely, for any method of prediction there exists such a stationary ergodic process that with probability 1

$$\limsup_{t\to\infty} |P^*(x_{t+1}|x_1 \dots x_t) - P(x_{t+1}|x_1 \dots x_t)| > 0; \quad (1)$$

$$\limsup_{t\to\infty} |p^*(x_{t+1}|x_1 \dots x_t) - p(x_{t+1}|x_1 \dots x_t)| > 0;$$

see [18], [33]. In other words, there is no method whose error goes to 0 for every stationary ergodic time series (when $t$ goes to infinity). On the other hand, it will be proven that there exists a method of prediction for which the following Cesaro average

$$\frac{1}{t}\Big(\sum_{m=0}^{t-1} |p(x|x_1 \dots x_m) - p^*(x|x_1 \dots x_m)|\Big), \quad (2)$$

$t \to \infty$, with probability 1 goes to 0 for any stationary ergodic source (the similar equation is true for the conditional probabilities [33]). So, there are no consistent estimates if the consistency is considered in the sense (1), but there are consistent estimates in the Cesaro (average) sense of (2).

It is shown in [1] that no procedure can consistently estimate the one-dimensional marginal density of every stationary ergodic process for which such a density exists. In other words, it is impossible to construct estimates for the density functions $p(x_1 \dots x_n)$, $n \geq 1$. On the other hand, it will be shown later that there are the consistent estimates in the Cesaro (average)

sense. That is why we will consider such estimates based on estimations of the conditional probabilities and the densities in order to estimate the (unconditional) ones as follows:

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

$$p^*(x_1...x_t) = \prod_{i=1}^{t} p^*(x_i|x_1...x_{i-1}). \quad (3)$$

It will be shown that in a certain sense those equations define reasonable estimations of the unknown probabilities and densities (See Theorems 1 and 2 below).

Let us briefly describe universal codes. Informally, a universal code compresses a sequence generated by a stationary and ergodic source with a finite alphabet till the Shannon entropy (per letter), which, in turn, is a lower bound for a compression ratio. We will show that universal codes can be directly applied to the prediction and some related problems for a certain class of a real-valued time series. It is worth noting that everyday methods of data compression (or archivers) like *zip*, *arj, rar,* etc., can be used as a tool for the density estimation and prediction, because the modern archivers are based on different universal codes and results of coding theory (see, for ex., [14], [21], [23], [30], [39]). Some examples of applications of real data compressors to prediction of currency rates are given in [37].

## II. Definitions and Preliminaries

First we consider a finite alphabet sources. Let $P$ be a stationary and ergodic source generating letters from a finite alphabet $A$. The Shannon entropy of the source is defined as follows:

$$H(p) = \lim_{m\to\infty} -\frac{1}{m} \sum_{v\in A^m} p(v) \log p(v), \quad (4)$$

where $A^m$ is a set of all words of the length $m$, $\log \equiv \log_2$ .

A data compression method (or code) $\varphi$ is defined as a set of mappings $\varphi_n$ such that $\varphi_n : A^n \to \{0,1\}^*$, $n = 1, 2, \dots$ and for each pair of different words $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2)...\varphi_n(u_r), r \geq 1$, of encoded words from the set $A^n, n \geq 1$, could be uniquely decoded into $u_1u_2...u_r$. Such codes are called uniquely decodable. For example, let $A = \{a, b\}$, the code $\psi_1(a) = 0, \psi_1(b) = 00$, obviously, is not uniquely decodable. It is well known that if a code $\varphi$ is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft's inequality): $\sum_{u\in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, for ex., [15]). It will be convenient to reformulate this property as follows:

**Claim 1.** *Let $\varphi$ be a uniquely decodable code over an alphabet $A$. Then for any integer $n$ there exists a measure $\mu_\varphi$ on $A^n$ such that*

$$-\log \mu_\varphi(u) \ \leq \ |\varphi(u)| \quad (5)$$

*for any $u$ from $A^n$ .*

(Obviously, Claim 1 is true for the measure $\mu_\varphi(u) = 2^{-|\varphi(u)|}/\Sigma_{u\in A^n} 2^{-|\varphi(u)|}$). In what follows we call uniquely decodable codes just "codes".

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:

$$\lim_{t\to\infty} |U(x_1 \dots x_t)|/t = H(P) \quad (6)$$

with probability 1, and

$$\lim_{t\to\infty} E(|U(x_1 \dots x_t)|)/t = H(P), \quad (7)$$

where $H(P)$ is the Shannon entropy of $P$, $E(f)$ is a mean value of $f$. It is worth noting that there exist codes for which (6) and (7) are proven; see, for example, [33].

The well known Shannon-MacMillan-Breiman theorem claims that for any stationary and ergodic source $P$

$$\lim_{t\to\infty} -\log P(x_1 \dots x_t)/t = H(P) \quad (8)$$

with probability 1, see [5], [15]. This theorem plays a key role in our consideration, because we can see from (6) and (8) that for any universal code $U$

$$\lim_{t\to\infty} ( \ |U(x_1 \dots x_t)|/t - \log P(x_1 \dots x_t)/t \ ) = 0.$$

So, in fact the length of universal code is a reasonable estimation of a logarithm of (unknown) probability $P(\,)$.

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

$$KL_t(P, \gamma) = \sum_{a\in A^t} P(v) \log \frac{P(v)}{\gamma(v)}. \quad (9)$$

It is well-known that for any distributions $P$ and $\gamma$ the KL divergence is nonnegative and equals 0 if and only if $P(v) = \gamma(v)$ for all $v$, see, for ex., [15]. The following inequality (Pinsker's inequality)

$$\sum_{a\in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} ||P - Q||^2. \quad (10)$$

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 [9]. It will be convenient to combine all properties of the probability estimators, which are based on universal codes.

**Theorem 1.** *Let $U$ be a universal code and*

$$\mu_U(u) = 2^{-|U(u)|}/\Sigma_{v\in A^{|u|}} 2^{-|U(v)|}. \quad (11)$$

*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,*

$$ii) \quad \lim_{t\to\infty} \frac{1}{t} \sum_{u\in A^t} P(u) \log(P(u)/\mu_U(u)) \; = \; 0,$$

Now we briefly consider the problem of prediction for time series with a finite alphabet. Let an (unknown) source $P$ generate a message $x_1 \dots x_{t-1}x_t$, and the following letter $x_{t+1}$ needs to be predicted. As it was mentioned above, we consider the prediction as a set of estimations of unknown (conditional) probabilities. At first glance it seems natural to estimate the precision of some prediction method $\gamma$ by one of the two following values:

$$\log \frac{P(x_{t+1}|x_1 \cdots x_t)}{\gamma(x_{t+1}|x_1 \cdots x_t)}, \; \sum_{a\in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}, \tag{12}$$

where $\gamma(.|x_1 \cdots x_t)$ is an estimation (a probability distribution) and $x_1 \cdots x_t$ is a word generated by the unknown source. As we mentioned above, this measure of prediction error is not suitable for this problem. The point is that for any predictor $\gamma$ there exists a stationary and ergodic source such that both values in (12) do not go to 0, when $t \to \infty$ (with probability 1). (The proof is given in [33]; see also [18].) On the other hand, it is proven in [33] that there exists a predictor $R$ for which the following Cesaro averages go to 0 for any stationary and ergodic source:

$$t^{-1} \sum_{i=0}^{t-1} \log(P(x_{i+1}|x_1 \cdots x_i)/R(x_{i+1}|x_1 \cdots x_i)) \; ,$$

(with probability 1) and

$$t^{-1} \sum_{i=0}^{t-1} P(x_1 \cdots x_{i+1}) \log \frac{P(x_{i+1}|x_1 \cdots x_i)}{R(x_{i+1}|x_1 \cdots x_i)}.$$

Hence, for any predictor $\gamma$ it is natural to estimate its error by values

$$t^{-1} \sum_{i=0}^{t-1} \log(P(x_{i+1}|x_1 \cdots x_i)/\gamma(x_{i+1}|x_1 \cdots x_i)) \; ,$$

(with probability 1) and

$$t^{-1} \sum_{i=0}^{t-1} P(x_1 \cdots x_{i+1}) \log \frac{P(x_{i+1}|x_1 \cdots x_i)}{\gamma(x_{i+1}|x_1 \cdots x_i)},$$

which, in turn, are equal to the following expressions

$$t^{-1} \log \frac{P(x_1 \cdots x_t)}{\gamma(x_1 \cdots x_t)}, \quad t^{-1} P(x_1 \cdots x_t) \log \frac{P(x_1 \cdots x_t)}{\gamma(x_1 \cdots x_t)},$$

correspondingly. So, if we take a universal code $U$ and apply it for prediction, the Theorem 1 will be true for the corresponding measure $\mu_U$. In other words, from mathematical point of view the problems of probability estimation and prediction are completely the same and can be considered together.

## III. Time series with a density

Here the problems of the density estimation and prediction for a stationary ergodic time series with densities are considered.

We have seen that Shannon-MacMillan-Breiman theorem played a key role in the case of finite-alphabet processes. In this part we will use its generalization to the processes with densities. This result was proved by Barron [3] and was an extension of the $L^1$ convergence obtained in [27], [29], [19]. First we describe considered processes with some properties needed for the generalized Shannon-MacMillan-Breiman theorem to hold. In what follows, we restrict our attention to real valued processes, but the main results may be extended to processes taking values in a complete separable metric space.

Let $B$ denote the Borel subsets of $R$, and $B^k$ denote the Borel subsets of $R^k$, where $R$ is the set of real numbers. Let $R^\infty$ be the set of all infinite sequences $x = x_1, x_2 \dots$ with $x_i \in R$, and let $B^\infty$ denote the usual product sigma field on $R^\infty$, generated by the finite dimensional cylinder sets $\{A_1, \dots A_k, R, R, \dots\}$, where $A_i \in B, i = 1, \dots, k$. Each stochastic process $X_1, X_2, \dots, X_i \in R$, is defined by a probability distribution on $(R^\infty, B^\infty)$. Suppose that the joint distribution $P_n$ for $(X_1, X_2, \dots, X_n)$ has a probability density function $p(x_1 x_2 \dots x_n)$ with respect to the Lebesgue measure $\lambda_n$ on $R^n$. 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

$$h = \lim_{n\to\infty} -E(\log p(x_{n+1}|x_1 \dots x_n)), \tag{13}$$

where $E$ denotes expectation with respect to $P$; see [3]. The following generalization of the Shannon-MacMillan-Breiman theorem follows from [3]:

**Claim 2.** *If $\{X_n\}$ is a $P-$stationary ergodic process with density $p(x_1 \dots x_n) = dP_n/d\lambda_n$ and $h_n < \infty$ for some $n \geq m$, the sequence of relative entropy densities $-(1/n) \log p(x_1 \dots x_n)$ convergence almost surely to the relative entropy rate, i.e.,*

$$\lim_{n\to\infty} (-1/n) \log p(x_1 \dots x_n) = h \tag{14}$$

*with probability 1 (according to $P$).*

Now we return to the estimation problems. Let $\{\Pi_n\}, n \geq 1$, be an increasing sequence of finite partitions of $R$ that asymptotically generates the Borel sigma-field $B$ and let $x^{[k]}$ denote the element of $\Pi_k$ that contains the point $x$. (Informally, $x^{[k]}$ is obtained by quantizing $x$ to $k$ bits of precision.) For integers $s$ and $n$ we define the following approximation of the density

$$p^s(x_1 \dots x_n) = P(x_1^{[s]} \dots x_n^{[s]})/\lambda_n(x_1^{[s]} \dots x_n^{[s]}). \tag{15}$$

We also consider

$$h_s = \lim_{n\to\infty} -E(\log p^s(x_{n+1}|x_1 \dots x_n)). \tag{16}$$

Applying the claim 2 to the density $p^s(x_1 \dots x_t)$, we obtain that a.s.

$$\lim_{t\to\infty} -\frac{1}{t}\log p^s(x_1 \dots x_t) = h_s. \tag{17}$$

Let $U$ be a universal code, which is defined for any finite alphabet. (In fact, all known universal codes possess this property , see, for example, [21], [23], [33].) In order to describe our density estimate we first define a probability distribution $\{\omega = \omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_1 = 1 - 1/\log 3,\ \dots,\ \omega_i = 1/\log(i+1) - 1/\log(i+2),\ \dots . \tag{18}$$

(In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) Now we can define the density estimate $r_U$ as follows:

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i\, \mu_U(x_1^{[i]} \dots x_t^{[i]})/\lambda_t(x_1^{[i]} \dots x_t^{[i]}), \tag{19}$$

where the measure $\mu_U$ is defined by (11). (It is assumed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters. The only properties of universal codes which are used later are the limit equations (6) and (7). That is why, in principal, it is possible to use different universal codes for different $i$, say, one for even $i$ and another for odd $i$, etc. The only requirement is that the equations (6) and (7) should be valid for such a mixed code.)

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

**Theorem 2** . *Let $X_t$ be a stationary ergodic time series with densities $p(x_1 \dots x_t) = dP_t/d\lambda_t$, where $\lambda_t$ is the Lebesgue measure on $R^t$ and let*

$$\lim_{s\to\infty} h_s = h < \infty, \tag{20}$$

*where $h$ and $h_s$ are relative entropy rates, see (13), (16). Then*

$$\lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 ... x_t)}{r_U(x_1 ... x_t)} = 0 \tag{21}$$

*with probability 1 and*

$$\lim_{t\to\infty} \frac{1}{t}\, E\log\frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \;=\; 0\,. \tag{22}$$

*Proof:* First we prove that with probability 1 there exists the following limit $\lim_{t\to\infty} \frac{1}{t}\log(p(x_1 \dots x_t)/r_U(x_1 \dots x_t))$ and this limit is finite and nonnegative. For this purpose we will use some results of the martingale theory, see e.g. [13],[42, Chapter 7]. Let $A_n = \{x_1, \dots, x_n : p(x_1, \dots, x_n) \neq 0\}$. Define

$$z_n(x_1 \dots x_n) = r_U(x_1 \dots x_n)/p(x_1 \dots x_n) \tag{23}$$

for $(x_1, \dots, x_n) \in A_n$ and $z_n = 0$ elsewhere.

Similarly to [42, pp. 524-525] we obtain

$$E(z_n|x_1, \dots, x_{n-1}) = E\left(\frac{r_U(x_1 \dots x_n)}{p(x_1 \dots x_n)}\middle| x_1, \dots, x_{n-1}\right)$$

$$= \frac{r_U(x_1 \dots x_{n-1})}{p(x_1 \dots x_{n-1})} E\left(\frac{r_U(x_n|x_1 \dots x_{n-1})}{p(x_n|x_1 \dots x_{n-1})}\right)$$

$$= z_{n-1}\int_A \frac{r_U(x_n|x_1 \dots x_{n-1})dP(x_n|x_1 \dots x_{n-1})}{dP(x_n|x_1 \dots x_{n-1})/d\lambda_n(x_n|x_1 \dots x_{n-1})}$$

$$= z_{n-1}\int_A r_U(x_n|x_1 \dots x_{n-1})d\lambda_n(x_n|x_1 \dots x_{n-1}) \le z_{n-1}.$$

Thus, the stochastic sequence $(z_n, B^n)$ is, by definition, a nonnegative supermartingale with respect to $P$, with $E(z_n) \le 1$, see [42, Chapter 7]. Hence, Doob's submartingale convergence theorem implies that the limit $z_n$ exists and is finite with $P-$probability 1 (see [42, Theorem 7.4.1]). Since all terms are nonnegative so is the limit. Using the definition (23) with $P$-probability 1 we have

$$\lim_{n\to\infty} p(x_1 \dots x_n)/r_U(x_1 \dots x_n) > 0,$$

$$\lim_{n\to\infty} \log(p(x_1 \dots x_n)/r_U(x_1 \dots x_n)) > -\infty$$

and

$$\lim_{n\to\infty} n^{-1}\,\log(p(x_1 \dots x_n)/r_U(x_1 \dots x_n)) \ge 0. \tag{24}$$

Now we note that for any integer $s$ the following obvious equality is true: $r_U(x_1 \dots x_t) = \omega_s\mu_U(x_1^{[s]} \dots x_t^{[s]})/\lambda_t(x_1^{[s]} \dots x_t^{[s]})\ (1 + \delta(x_1 \dots x_t))$ for some $\delta(x_1 \dots x_t) > 0$. From this equality, (11) and (19) we immediately obtain that a.s.

$$\lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \;\le \lim_{t\to\infty} \frac{-\log\omega_s}{t}$$

$$+ \lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{\mu_U(x_1^{[s]} \dots x_t^{[s]})/\lambda_t(x_1^{[s]} \dots x_t^{[s]})}$$

$$\le \lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/\lambda_t(x_1^{[s]} \dots x_t^{[s]})}. \tag{25}$$

The right part can be presented as follows:

$$\lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/\lambda_t(x_1^{[s]} \dots x_t^{[s]})}$$

$$= \lim_{t\to\infty} \frac{1}{t}\log\frac{p^s(x_1 \dots x_t)\ \lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \tag{26}$$

$$+ \lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{p^s(x_1 \dots x_t)}.$$

Having taken into account that $U$ is a universal code, (15) and the theorem 1, we can see that the first term is equal to zero. From (14) and (17) we can see that a.s. the second term is equal to $h_s - h$. This equality is valid for any integer $s$ and, according to (20), the second term equals zero, too, and we obtain that

$$\lim_{t\to\infty} \frac{1}{t}\log\frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \;\le 0.$$

Having taken into account (24), we can see that the first statement is proven.

From (25) and (26) we can can see that

$$E\,\log\frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \le E\,\log\frac{p^s(x_1, \dots, x_t)\ \lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}}$$

$$+E\ \log\frac{p(x_1\dots x_t)}{p^s(x_1,\dots,x_t)}. \quad (27)$$

The first term is the average redundancy of the universal code for a finite- alphabet source, hence, according to the theorem 1, it tends to 0. The second term tends to $h_s - h$ for any $s$ and from (20) we can see that it is equals to zero. The second statement is proven. ■

We have seen that the requirement (20) plays an important role in the proof. The natural question is whether there exist processes for which (20) is valid. The answer is positive. For example, let a process possessed values in the interval $[-1,1]$, $\lambda_n$ be Lebesgue measure on $R^n$ and the considered process is Markovian with conditional density

$$p(x|y)=\begin{cases}1/2+\alpha\ \ sign(y)\ \sin(\pi x), & \text{if } x<0\,;\\ 1/2+\alpha\ \ sign(y)\ \sin(\pi x), & \text{if } x\geq 0\ ,\end{cases}$$

where $\alpha\in(0,1/2)$ is a parameter and

$$sign(y)=\begin{cases}-1, & \text{if } y<0,\\ 1, & \text{if } y\geq 0\ .\end{cases}$$

In words, the density depends on a sign of the previous value. It is easy to see that (20) is true for any $\alpha\in(0,1/2)$.

So we have seen that there exist time series for which the equality (20) is valid. The next question is whether there exist general conditions which guarantee that (20) is true. The equation in (20) in a certain sense looks like a definition of the Riemann integral. We will suggest one simple condition which guarantees validity of (20). This conditions directly follows from well-known properties of the Riemann integral whose definition can be found, for example, in [6].

**Claim 3.** *Let there be a time series generating elements from a finite interval $I\subset R$. If the following properties are valid: i) the memory of the time series is finite (i.e. there exists such $m$ that $p(x|x_1....x_t)=p(x|x_1....x_m)$ if $t\geq m$), ii) $-p(x|x_1....x_m)\log p(x|x_1....x_m)$ is upper bounded by a constant $C$ and is piecewise continuous function, then (20) is valid.*

*Comment.* The parameters $I$, $m$ and $C$ may be unknown.

*Proof:* From the definition (16) and the property i) we obtain that $h_s=-E(\log p^s(x|x_1\dots x_m))$. The value $-E(\log p^s(x|x_1\dots x_m))$ equals $-\int p^s(x|x_1\dots x_m)\log p^s(x|x_1\dots x_m)d\lambda_{m+1}$. From the definition (16) we can see that the function $\log p^s(x|x_1\dots x_m))$ has the same values for all $x\in\Delta$, where $\Delta$ is an element of the partition $\Pi_s^{m+1}$. Hence, taking into account the definition of $p^s()$ (15) we obtain the following inequalities:

$$\sum_{\Delta\in\Pi_s^{m+1}}\min_{(x,x_1,\dots,x_m)\in\Delta}p(x|x_1\dots x_m)\log p(x|x_1\dots x_m)\lambda(\Delta)$$

$$\leq\int p^s(x|x_1\dots x_m)\log p^s(x|x_1\dots x_m)d\lambda_{m+1}\leq$$

$$\sum_{\Delta\in\Pi_s^{m+1}}\max_{(x,x_1,\dots,x_m)\in\Delta}p(x|x_1\dots x_m)\log p(x|x_1\dots x_m)\lambda(\Delta),$$

where $\lambda(\Delta)$ is the volume of $\Delta$. Having taken into account the definition of Riemann integral [6, Addition, \$4], the properties ii) and the fact that the densities are defined on the finite interval $I$, we obtain that

$$\lim_{s\to\infty}\int p^s(x|x_1\dots x_m)\log p^s(x|x_1\dots x_m)d\lambda_{m+1}=$$

$$\int p(x|x_1\dots x_m)\log p(x|x_1\dots x_m)d\lambda_{m+1}.$$

■

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

**Theorem 3.** *Let $f$ be an integrable function, whose absolute value is bounded by a certain constant $b$ and all conditions of the theorem 2 are true. Then the following equality is valid:*

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

$$\int f(x)\,r_U(x|x_1...x_m)d\lambda_m)^2)=0, \quad (28)$$

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

$$\int f(x)\,r_U(x|x_1...x_m)\,d\lambda_m|)=0.$$

*Proof:* The last inequality of the following chain follows from the Pinsker's one, whereas all others are obvious.

$$(\int f(x)\,p(x|x_1...x_m)\,d\lambda_m-\int f(x)\,r_U(x|x_1...x_m)\,d\lambda_m)^2$$

$$=(\int f(x)\,(p(x|x_1...x_m)-r_U(x|x_1...x_m))\,d\lambda_m)^2$$

$$\leq b^2(\int(p(x|x_1...x_m)-r_U(x|x_1...x_m))\,d\lambda_m)^2$$

$$\leq b^2(\int|p(x|x_1...x_m)-r_U(x|x_1...x_m)|d\lambda_m)^2$$

$$\leq const\int p(x|x_1...x_m)\ \log\frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)}d\lambda_m.$$

From these inequalities we obtain:

$$E(\sum_{m=0}^{t-1}(\int f(x)\,p(x|x_1...x_m)\,d\lambda_m-$$

$$\int f(x)\,r_U(x|x_1...x_m)\,d\lambda_m)^2)\leq \quad (29)$$

$$\sum_{m=0}^{t-1}const\,E(\int p(x|x_1...x_m)\ \log\frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)}d\lambda_m).$$

The last term can be presented as follows:

$$\sum_{m=0}^{t-1} E(\int p(x|x_1...x_m)\ \log \frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)} d\lambda_m) =$$

$$\sum_{m=0}^{t-1} \int p(x_1...x_m)$$

$$\int p(x|x_1...x_m) \log \frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)} d\lambda_1 d\lambda_m$$

$$= \int p(x_1...x_t) \log(p(x_1...x_t)/r_U(x_1...x_t)) d\lambda_t.$$

From this equality, (29) and Corollary 1 we obtain (28). ii) can be derived from (29) and the Iensen inequality for the function $x^2$. ■

*Comment.* In fact, the statements i) and ii) are equivalent. Our proof is similar to the method from [38], see Lemma 2 there.

## ACKNOWLEDGMENT

I am grateful to Daniil Ryabko for finding a short proof of the Theorem 2 and help with applying of the martingale theory.

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

## REFERENCES

[1] T. M. Adams, A. B. Nobel, "On Density Estimation from Ergodic Processes," *Ann. Probab.*, v. 26, n.2, pp. 794-804, 1998.

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

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

[4] A.R.Barron, L.Györfi and E.C. van der Meulen, "Distribution Estimation Consistent in Total Variation and in Two Types of Information Divergence," *IEEE Transactions on Information Theory*, 1992 v.38, n.5, pp.1437-1454.

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

[6] L.Bers, *Calculus, Part II.* Holt, Rinehart and Winston Inc., 1969.

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

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

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

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

[11] G.A. Darbellay and I.Vajda, *Entropy expressions for multivariate continuous distributions.* Research Report no 1920, UTIA, Academy of Science, Prague, 1998. (library@utia.cas.cz).

[12] G.A.Darbellay and I.Vajda, "Estimatin of the mutual information with data-dependent partitions," *IEEE Trans. Inform. Theory*, v. 48, n. 5, pp. 1061-1081, 2002.

[13] J.L. Doob, *Stochastic Processes.* John Wiley & Sons, New York, 1990.

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

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

[16] P.D. Grünwald and J. Langford, "Suboptimal behavior of Bayes and MDL in classification under misspecification," *Machine Learning* , v.66, n.2-3, pp. 119-149, 2007.

[17] P.D. Grünwald, *The Minimum Description Length Principle.* MIT Press, 2007.

[18] L. Györfi , G. Morvai, S.J. Yakowitz, "Limits to Consistent On-Line Forecasting for Ergodic Time Series," *IEEE Transactions on Information Theory*, v.44, n.2, pp. 886 –892.

[19] J.Kieffer, "A simple proof of the Moy-Perez generalization of the Shannon-MacMillan theorem," *Pacific J. Math.*, v.51, pp. 203-206, 1974.

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

[21] J.C.Kieffer and En-Hui Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Transactions on Information Theory*, v.46, n.3, pp. 737 - 754, 2000.

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

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

[24] S. Kullback, *Information Theory and Statistics*. Wiley, New York, 1959.

[25] D.S. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Inform. Theory,* 44, n.1, 117-133, 1998.

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

[27] S.C.Moy, "Generalisations of Shannon-MacMillan theorem," *Pacific J. Math.*, v.11, pp. 705-714, 1961.

[28] A.B. Nobel, "On optimal sequential prediction," *IEEE Trans. Inform. Theory*, v. 49, n.1, pp. 83-98. 2003.

[29] A. Perez, "Extensions of Shannon-MacMillan's limit theorem to more general stohastic processes," *Trans. Third Prague Conference on Information Thepry, Statistical Decision Functions, and Random Processes*, 1964, pp. 545-574. Czehoslovak Academy of Sciences, Prague.

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

[31] J. Rissanen, *Information and complexity in statistical modeling.* Springer Verlag, 2007.

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

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

[34] B. Ya. Ryabko, "The complexity and effectiveness of prediction algorithms," *J. Complexity*, v. 10, no. 3, 281–295, 1994.

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

[36] B. Ya. Ryabko and V.A. Monarev, "Using information theory approach to randomness testing," *Journal of Statistical Planning and Inference*, v. 133, n.1, pp. 95-110, 2005.

[37] B.Ryabko and V.Monarev, "Experimental Investigation of Forecasting Methods Based on Data Compression Algorithms," *Problems of Information Transmission*, v.41, n.1, pp. 65-69, 2005.

[38] D.Ryabko and M.Hutter, "Sequence prediction for non-stationary processes," In proceedings: *Combinatorial and Algorithmic Foundations of Pattern and Association Discovery, Dagstuhl Seminar*, Germany, 2006,. http://www.dagstuhl.de/06201/ see also http://arxiv.org/pdf/cs.LG/0606077

[39] S. A.Savari, "A probabilistic approach to some asymptotics in noiseless communication," *IEEE Transactions on Information Theory*, v. 46, n.4, pp. 1246-1262, 2000.

[40] Shannon C. E., "A mathematical theory of communication", *Bell Sys. Tech. J.*, vol. 27, pp. 379-423 and pp. 623-656, 1948.

[41] P.C.Shields, "The interactions between ergodic theory and information theory," *IEEE Transactions on Information Theory*, v. 44, n. 6, pp. 2079 - 2093, 1998.

[42] A.N. Shiryaev, *Probability*, (second edition), Springer, 1995.

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

**Author's biography:** Boris Ryabko
received the M.S. degree from Novosibirsk state university in 1971, Ph.D. degree from Institute of Mathematics, Novosibirsk in 1981 and Dr.Sci. degree from Inst. of Problems of Information Transmission, Moscow, 1989. He has been Professor of applied mathematics and computer science since 1986. Now

he is a head of Department of Appl. Math and Cybernetics and Pro-rector of Science at the Siberian State University of Telecommunication and Informatics.

His research interests include Applied Mathematics, Information and Coding Theory, Cryptography and Mathematical Biology. He published more than 150 scientific articles and 4 books in these fields.