

# Compression-based methods for nonparametric density estimation, prediction, regression and classification for time series.

Boris Ryabko \*

## Abstract

We address the problem of nonparametric estimation of characteristics for stationary and ergodic time series. We consider finite-alphabet time series and real-valued ones and the following four problems: i) estimation of the (limiting) probability  $P(u_0 \dots u_s)$  for every  $s$  and each sequence  $u_0 \dots u_s$  of letters from the process alphabet (or estimation of the density  $p(x_0, \dots, x_s)$  for real-valued time series), ii) 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  $x_1 x_2 \dots x_t$  is known), iii) regression and iv) classification (or so-called problems with side information).

We show that so-called archivers (or data compressors) can be used as a tool for solving these problems. In particular, firstly, it is proven that any so-called universal code (or universal data compressor) can be used as a basis for constructing asymptotically optimal methods for the above problems. (By definition, a universal code can "compress" any sequence generated by a stationary and ergodic source asymptotically till the Shannon entropy of the source.) And, secondly, we show experimentally that estimates, which are based on practically used methods of data compression, have a reasonable precision.

**AMS subject classification:** 60G10, 60J10, 62G07, 62G08, 62M20, 94A29.

**keywords:** time series, nonparametric estimation, prediction, universal coding, data compression, on-line prediction, Shannon entropy, stationary and ergodic process, regression.

## 1 Introduction

We consider a stationary and ergodic source, which generates sequences  $x_1 x_2 \dots$  of elements (letters) from some set (alphabet)  $A$ , which is either finite or real-valued. It is supposed that the probability distribution (or distribution of limiting probabilities)  $P(x_1 = a_{i_1}, x_2 = a_{i_2}, \dots, x_t = a_{i_t})$  (or the density  $p(x_1, x_2, \dots, x_t)$ ) is unknown,

---

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

but we are given either one sample  $x_1 \dots x_t$  or several ( $r$ ) non-overlapping samples  $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^r = x_1^r \dots x_{t_r}^r$  generated by the source. (Here non-overlapping means that the sequences either are parts of different realizations or belongs to non-overlapping parts of one realization, say, a realization with gaps. Generally speaking, they cannot be combined into one sample for a stationary and ergodic source, as it can be done for an i.i.d. one.)

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, 14, 15, 17, 24, 25, 27, 41].

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:

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

where  $x_1 \dots x_t$  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 ( $|U(x_1 \dots x_t)|$ ) can be used as an estimate of the logarithm of the unknown probability and, obviously,  $2^{-|U(x_1 \dots x_t)|}$  can be considered as the estimation of  $p(x_1 \dots x_t)$ . 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 [29]) and universal codes were applied for solving prediction problem [30].

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, 18, 22, 27, 37]) 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 [4, 5] 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 [33, 34].

The outline of the paper is as follows. The section 2 contains description of the Laplace predictor and its generalizations, a review of known results and description of one universal code. The sections 3 and 4 are devoted to processes with finite and real-valued alphabets, correspondingly. The last part contains some examples and simulations.

## 2 Predictors and universal data compressors

### 2.1 The Laplace measure and on-line prediction for i.i.d. processes

We consider a source with unknown statistics which generates sequences  $x_1x_2\cdots$  of letters from some set (or alphabet)  $A$ . Let the source generate a message  $x_1\ldots x_{t-1}x_t\ldots$ ,  $x_i \in A$  for all  $i$ , and the following letter  $x_{t+1}$  needs to be predicted.

It will be convenient at first to describe briefly the prediction problem. This problem can be traced back to Laplace [11]. He considered the problem how to estimate the probability that the sun will rise tomorrow, given that it has risen every day since Creation. In our notation the alphabet  $A$  contains two letters 0 ("*the sun rises*") and 1 ("*the sun does not rise*"),  $t$  is the number of days since Creation,  $x_1\ldots x_{t-1}x_t = 00\ldots 0$ .

Laplace suggested the following predictor:

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

see [11], 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  $A = \{0, 1\}$ ,  $x_1\ldots x_5 = 01010$ , 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 [30] and now is often called on-line prediction or universal prediction [1, 14, 25]. 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 [16, 32]) and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor  $\gamma$  defines a measure by following equation

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

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  $\gamma$  (or estimation of the measure) defines a predictor:  $\gamma(x_i|x_1\ldots x_{i-1}) = \gamma(x_1\ldots x_{i-1}x_i)/\gamma(x_1\ldots x_{i-1})$ . The

same is true for a density (and its estimation): a predictor is defined by conditional density and, vice versa, the density is equal to the product of conditional densities:

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

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  $\gamma$ . The *error* is characterized by the KL divergence

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

It is well-known that for any distributions  $p$  and  $\gamma$  the K-L divergence is nonnegative and equals 0 if and only if  $p(a) = \gamma(a)$  for all  $a$ , see, for ex., [13]. The following inequality (Pinsker's inequality)

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

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 [6]. For fixed  $t$ ,  $\rho_{\gamma,p}(\cdot)$  is a random variable, because  $x_1, x_2, \dots, x_t$  are random variables. We define the average error at time  $t$  by

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

It is shown in [31] that the error of Laplace predictor  $L_0$  goes to 0 for any i.i.d. source  $p$ . More precisely, it is proven that

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

for any source  $p$ , ([31]; see also [35]). 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 [30] (see also [1, 14, 25]) 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  $\gamma$  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  $t \rightarrow \infty$ . 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 [30] 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

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

and

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

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

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:

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

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  $L_0(x_1 \dots x_t)$  is a consistent estimate of the unknown probability  $p(x_1 \dots x_t)$ .

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 [21], see also [22]. He suggested the following predictor:

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

where, as before,  $\nu_{x_1 \dots x_t}(a)$  denote the count of letter  $a$  occurring in the word  $x_1 \dots x_t$ . We can see that the Krichevsky predictor is quite close to the Laplace's one (1). For example, if  $A = \{0, 1\}$ ,  $x_1 \dots x_5 = 01010$ , 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:

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

It is known that

$$(r + 1/2)((r + 1) + 1/2) \dots (s - 1/2) = \frac{\Gamma(s + 1/2)}{\Gamma(r + 1/2)}, \quad (12)$$

where  $\Gamma(\cdot)$  is the gamma function (see for definition, for ex., [19]). So, (11) can be presented as follows:

$$K_0(x_1 \dots x_t) = \frac{\prod_{a \in A} (\Gamma(\nu_{x_1 \dots x_t}(a) + 1/2) / \Gamma(1/2))}{\Gamma(t + |A|/2) / \Gamma(|A|/2)}. \quad (13)$$

For this predictor

$$\bar{\rho}_t(K_0, p) < ((|A| - 1) \log t + c)/(2t), \quad (14)$$

where  $c$  is a constant, and, moreover, in a certain sense this average error is minimal: for any predictor  $\gamma$  there exists such a source  $p^*$  that

$$\bar{\rho}_t(\gamma, p^*) \geq ((|A| - 1) \log t + c)/(2t),$$

see [21], [22].

## 2.2 Consistent estimations and on-line predictors for Markov and ergodic processes

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 [35]. So assume that  $A = \{O, I\}$ ,  $m = 2$  and assume that the source  $p \in M_2(A)$  has generated the sequence

*OOIOIIIOOIIIOIO.*

We represent this sequence by the following four subsequences:

\* \* I \* \* \* \* I \* \* \* \* ,

\* \* \* O \* I \* \* \* I \* \* \* O ,

\* \* \* \* I \* \* O \* \* \* \* I \* ,

\* \* \* \* \* O \* \* \* I O \* \* .

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  $\gamma$  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  $\gamma$  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:  $L_m(x_1) = L_m(x_2) = \dots = L_m(x_m) = 1/|A|$ ,  $K_m(x_1) = K_m(x_2) = \dots = K_m(x_m) = 1/|A|$ . For example, having taken into account (13), we can present the Krichevsky predictors for  $M_m(A)$  as follows:

$$K_m(x_1 \dots x_t) = \begin{cases} \frac{1}{|A|^t}, & \text{if } t \leq m, \\ \frac{1}{|A|^m} \prod_{v \in A^m} \frac{\prod_{a \in A} ((\Gamma(\nu_x(va) + 1/2) / \Gamma(1/2))}{(\Gamma(\bar{\nu}_x(v) + |A|/2) / \Gamma(|A|/2))}, & \text{if } t > m, \end{cases} \quad (15)$$

where  $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va)$ ,  $x = x_1 \dots x_t$ . It is worth noting that the representation (12) can be more convenient for carrying out calculations. Let us consider an example. For the word  $OOIOIIIOOIIIOIO$  considered in the previous example, we obtain  $K_2(OOIOIIIOOIIIOIO) = 2^{-2} \frac{1}{2} \frac{3}{4} \frac{1}{2} \frac{1}{4} \frac{1}{2} \frac{3}{8} \frac{1}{2} \frac{1}{4} \frac{1}{2} \frac{1}{2} \frac{1}{2}$ .

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  $\{\omega = \omega_1, \omega_2, \dots\}$  on integers  $\{1, 2, \dots\}$  by

$$\omega_1 = 1 - 1/\log 3, \dots, \omega_i = 1/\log(i+1) - 1/\log(i+2), \dots \quad (16)$$

(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:

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t). \quad (17)$$

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

$$H(p) = \lim_{m \rightarrow \infty} -\frac{1}{m} \sum_{v \in A^m} p(v) \log p(v). \quad (18)$$

**Theorem 1.** [30]. *For any stationary and ergodic source  $p$  the following equalities are valid:*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} \log(1/R(x_1 \dots x_t)) = H(p)$$

with probability 1,

$$ii) \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} p(u) \log(1/R(u)) = H(p).$$

## 2.3 Nonparametric estimations and data compression

One of the goals of the paper is to show how practically used data compressors can be used as a tool for nonparametric estimation, prediction and other problems. That

is why a short description of universal data compressors (or universal codes) will be given here.

A data compression method (or code)  $\varphi$  is defined as a set of mappings  $\varphi_n$  such that  $\varphi_n : A^n \rightarrow \{0, 1\}^*$ ,  $n = 1, 2, \dots$  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  $\varphi_n(u_1)\varphi_n(u_2)\dots\varphi_n(u_r)$ ,  $r \geq 1$ , of encoded words from the set  $A^n$ ,  $n \geq 1$ , could be uniquely decoded into  $u_1u_2\dots u_r$ . 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., [13]. 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*

$$-\log \mu_\varphi(u) \leq |\varphi(u)| \quad (19)$$

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

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

It is worth noting that, in fact, any measure  $\mu$  defines a code for which the length of the codeword associated with a word  $u$  is (close to)  $-\log \mu(u)$ .

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:

$$\lim_{t \rightarrow \infty} |U(x_1 \dots x_t)|/t = H(p) \quad (20)$$

with probability 1, and

$$\lim_{t \rightarrow \infty} E(|U(x_1 \dots x_t)|)/t = H(p), \quad (21)$$

where  $H(p)$  is the Shannon entropy of  $p$ ,  $E(f)$  is a mean value of  $f$ . In fact, (21) and (20) are valid for known universal codes, but there exist codes for which only one equality is valid.

## 3 Finite-alphabet processes

### 3.1 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*

$$\mu_U(u) = 2^{-|U(u)|} / \sum_{v \in A^{|u|}} 2^{-|U(v)|}. \quad (22)$$

*Then, for any stationary and ergodic source  $p$  the following equalities are valid:*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} (-\log p(x_1 \dots x_t) - (-\log \mu_U(x_1 \dots x_t))) = 0$$



with probability 1,

$$\begin{aligned} \text{ii)} \quad & \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} p(u) \log(p(u)/\mu_U(u)) = 0, \\ \text{iii)} \quad & \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{u \in A^t} p(u) |p(u) - \mu_U(u)| = 0. \end{aligned}$$

*Proof* is based on Shannon-MacMillan-Breiman Theorem which states that for any stationary and ergodic source  $p$

$$\lim_{t \rightarrow \infty} -\log p(x_1 \dots x_t)/t = H(p)$$

with probability 1, see [3, 13]. From this equality and (20) we obtain the statement i). The second statement follows from the definition of Shannon entropy (18) and (21), whereas iii) follows from ii) and the Pinsker's inequality (4).

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

Nowadays there are many efficient universal codes (and universal predictors connected with them), see [15, 17, 26, 27, 30, 37], which can be applied to estimation. For example, the above described measure  $R$  is based on the code from [29, 30] and can be applied for probability estimation. More precisely, Theorem 2 (and the following theorems) are true for  $R$ , if we replace  $\mu_U$  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.** *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:*

$$\frac{1}{t} \sum_{u \in A^t} p(u) \log(p(u)/R(u)) \leq \frac{(|A| - 1)|A|^{k-1} \log t}{2t} + O\left(\frac{1}{t}\right),$$

*ii) in a certain sense the error of  $R$  is asymptotically minimal: for any measure  $\mu$  there exists a  $k$ -memory Markov process  $p_\mu$  such that*

$$\frac{1}{t} \sum_{u \in A^t} p_\mu(u) \log(p_\mu(u)/\mu(u)) \geq \frac{(|A| - 1)|A|^{k-1} \log t}{2t} + O\left(\frac{1}{t}\right),$$

*iii) Let  $\Theta$  be such a set of stationary and ergodic processes that there exists a measure  $\mu_\Theta$  for which the estimation error of the probability goes to 0 uniformly:*

$$\lim_{t \rightarrow \infty} \sup_{p \in \Theta} \left( \frac{1}{t} \sum_{u \in A^t} p(u) \log(p(u)/\mu_\Theta(u)) \right) = 0.$$

*Then the error of estimator, which is based on the measure  $R$ , goes to 0 uniformly, too:*

$$\lim_{t \rightarrow \infty} \sup_{p \in \Theta} \left( \frac{1}{t} \sum_{u \in A^t} p(u) \log(p(u)/R(u)) \right) = 0.$$

*Proof* can be found in [30, 31].

### 3.2 Prediction

As we mentioned above, any universal code  $U$  can be applied for prediction. Namely, the measure  $\mu_U$  (22) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1 \dots x_t) = \mu_U(x_1 \dots x_t x_{t+1}) / \mu_U(x_1 \dots x_t). \quad (23)$$

**Theorem 4.** *Let  $U$  be a universal code and  $p$  be any stationary and ergodic process. Then*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} \{E(\log \frac{p(x_1)}{\mu_U(x_1)}) + E(\log \frac{p(x_2|x_1)}{\mu_U(x_2|x_1)}) + \dots + E(\log \frac{p(x_t|x_1 \dots x_{t-1})}{\mu_U(x_t|x_1 \dots x_{t-1})})\} = 0,$$

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

and

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

*Proof* i) immediately follows from the second statement of the previous theorem and properties of log. The statement ii) can be proved as follows:

$$\begin{aligned} & \lim_{t \rightarrow \infty} E(\frac{1}{t} \sum_{i=0}^{t-1} (p(x_{i+1}|x_1 \dots x_i) - \mu_U(x_{i+1}|x_1 \dots x_i))^2) \leq \\ & \lim_{t \rightarrow \infty} E(\frac{1}{t} \sum_{i=0}^{t-1} (\sum_{a \in A} |p(a|x_1 \dots x_i) - \mu_U(a|x_1 \dots x_i)|)^2) \leq \\ & \lim_{t \rightarrow \infty} E(\frac{const}{t} \sum_{i=0}^{t-1} \sum_{a \in A} p(a|x_1 \dots x_i) \log(p(a|x_1 \dots x_i)/\mu_U(a|x_1 \dots x_i))) = \\ & \lim_{t \rightarrow \infty} (\frac{const}{t} \sum_{i=0}^{t-1} p(x_1 \dots x_i) \sum_{a \in A} p(a|x_1 \dots x_i) \log(p(a|x_1 \dots x_i)/\mu_U(a|x_1 \dots x_i))) = \\ & \lim_{t \rightarrow \infty} (\frac{const}{t} \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log(p(x_1 \dots x_t)/\mu(x_1 \dots x_t))). \end{aligned}$$

Here the first inequality is obvious, the second follows from the Pinsker's inequality (4), the others from properties of expectation and log. iii) can be derived from ii) and the Jensen inequality for the function  $x^2$ . Theorem is proven.

*Comment 1.* The measure  $R$  described above has one additional property, if it is used for prediction. Namely, for any Markov process  $p$  ( $p \in M^*(A)$ ) the following is true:

$$\lim_{t \rightarrow \infty} \log \frac{p(x_{t+1}|x_1 \dots x_t)}{R(x_{t+1}|x_1 \dots x_t)} = 0$$

with probability 1, where  $R(x_{t+1}|x_1 \dots x_t) = R(x_1 \dots x_t x_{t+1})/R(x_1 \dots x_t)$ ; see [31].

*Comment 2.* In fact, the statements ii) and iii) are equivalent, because one of them follows from the other. For details see Lemma 2 in [36].

### 3.3 Problems with side information

Now we consider so-called problems with side information, which are described as follows: there is a stationary and ergodic source, whose alphabet  $A$  is presented as a product  $A = X \times Y$ . We are given a sequence  $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$  and so-called side information  $y_t$ . The goal is to predict, or estimate,  $x_t$ . This problem arises in statistical decision theory, pattern recognition, and machine learning, see [25]. Obviously, if someone knows the conditional probabilities  $p(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)$  for all  $x_t \in X$ , he has all information about  $x_t$ , available before  $x_t$  is known. That is why we will look for the best (or, at least, good) estimations for this conditional probabilities. Our solution will be based on results obtained in two previous subparagraphs. More precisely, for any universal code  $U$  and the corresponding measure  $\mu_U$  (22) we define the following estimate for the problem with side information:

$$\mu_U(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t) = \frac{\mu_U((x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (x_t, y_t))}{\sum_{x_t \in X} \mu_U((x_1, y_1), \dots, (x_{t-1}, y_{t-1}), (x_t, y_t))}.$$

**Theorem 5.** *Let  $U$  be a universal code and  $p$  be any stationary and ergodic process. Then*

$$\begin{aligned} i) \quad & \lim_{t \rightarrow \infty} \frac{1}{t} \left\{ E\left(\log \frac{p(x_1 | y_1)}{\mu_U(x_1 | y_1)}\right) + E\left(\log \frac{p(x_2 | (x_1, y_1), y_2)}{\mu_U(x_2 | (x_1, y_1), y_2)}\right) + \dots \right. \\ & \left. + E\left(\log \frac{p(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)}{\mu_U(x_t | (x_1, y_1), \dots, (x_{t-1}, y_{t-1}), y_t)}\right) \right\} = 0, \\ ii) \quad & \lim_{t \rightarrow \infty} E\left(\frac{1}{t} \sum_{i=0}^{t-1} (p(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1})) - \right. \\ & \left. \mu_U(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1}))^2 \right) = 0, \end{aligned}$$

and

$$\begin{aligned} iii) \quad & \lim_{t \rightarrow \infty} E\left(\frac{1}{t} \sum_{i=0}^{t-1} |p(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1})) - \right. \\ & \left. \mu_U(x_{i+1} | (x_1, y_1), \dots, (x_i, y_i), y_{i+1}))| \right) = 0. \end{aligned}$$

*Proof.* The following inequality follows from the nonnegativity of the K-L divergence (see (4)), whereas equality is obvious.

$$\begin{aligned} & E\left(\log \frac{p(x_1 | y_1)}{\mu_U(x_1 | y_1)}\right) + E\left(\log \frac{p(x_2 | (x_1, y_1), y_2)}{\mu_U(x_2 | (x_1, y_1), y_2)}\right) + \dots \leq \\ & E\left(\log \frac{p(y_1)}{\mu_U(y_1)}\right) + E\left(\log \frac{p(x_1 | y_1)}{\mu_U(x_1 | y_1)}\right) + E\left(\log \frac{p(y_2 | (x_1, y_1))}{\mu_U(y_2 | (x_1, y_1))}\right) + E\left(\log \frac{p(x_2 | (x_1, y_1), y_2)}{\mu_U(x_2 | (x_1, y_1), y_2)}\right) + \dots \\ & = E\left(\log \frac{p(x_1, y_1)}{\mu_U(x_1, y_1)}\right) + E\left(\log \frac{p((x_2, y_2) | (x_1, y_1))}{\mu_U((x_2, y_2) | (x_1, y_1))}\right) + \dots \end{aligned}$$

Now we can apply the first statement of the theorem 4 to the last sum as follows:

$$\lim_{t \rightarrow \infty} \frac{1}{t} E(\log \frac{p(x_1, y_1)}{\mu_U(x_1, y_1)}) + E(\log \frac{p((x_2, y_2)|(x_1, y_1))}{\mu_U((x_2, y_2)|(x_1, y_1))}) + \dots$$

$$E(\log \frac{p((x_t, y_t)|(x_1, y_1) \dots (x_{t-1}, y_{t-1}))}{\mu_U((x_t, y_t)|(x_1, y_1) \dots (x_{t-1}, y_{t-1}))}) = 0.$$

From this equality and last inequality we obtain the proof of i). The proof of the second statement can be obtained from the similar representation for ii) and the second statement of the theorem 4. iii) can be derived from ii) and the Jensen inequality for the function  $x^2$ . Theorem is proven.

### 3.4 The case of several independent samples

Now we extend our consideration to the case where the sample is presented as several non-overlapping sequences  $x^1 = x_1^1 \dots x_{t_1}^1$ ,  $x^2 = x_1^2 \dots x_{t_2}^2, \dots, x^r = x_1^r \dots x_{t_r}^r$  generated by a source. More precisely, we will suppose that all sequences were created by one stationary and ergodic source. (As it was mentioned above, it is impossible just to combine all samples into one, if the source is not i.i.d.) We denote this sample by  $x^1 \diamond x^2 \dots \diamond x^r$  and define  $\nu_{x^1 \diamond x^2 \dots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$ . For example, if  $x^1 = 0010, x^2 = 011$ , then  $\nu_{x^1 \diamond x^2}(00) = 1$ . The definition of  $K_m$  and  $R$  can be extended to this case:

$$K_m(x^1 \diamond x^2 \dots \diamond x^r) = \tag{24}$$

$$\left( \prod_{i=1}^r |A|^{-\min\{m, t_i\}} \right) \prod_{v \in A^m} \frac{\prod_{a \in A} ((\Gamma(\nu_{x^1 \diamond x^2 \dots \diamond x^r}(va) + 1/2) / \Gamma(1/2))}{(\Gamma(\bar{\nu}_{x^1 \diamond x^2 \dots \diamond x^r}(v) + |A|/2) / \Gamma(|A|/2))},$$

whereas the definition of  $R$  is the same (see (17)). (Here, as before,  $\bar{\nu}_{x^1 \diamond x^2 \dots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond x^2 \dots \diamond x^r}(va)$ . Note, that  $\bar{\nu}_{x^1 \diamond x^2 \dots \diamond x^r}() = \sum_{i=1}^r t_i$  if  $m = 0$ .)

The following example is intended to show the difference between the case of many samples and one. Let there be two independent samples  $y = y_1 \dots y_4 = 0101$  and  $x = x_1 \dots x_3 = 101$ , generated by a stationary and ergodic source with the alphabet  $\{0, 1\}$ . One wants to estimate the (limiting) probabilities  $P(z_1 z_2)$ ,  $z_1, z_2 \in \{0, 1\}$  (here  $z_1 z_2 \dots$  can be considered as an independent sequence, generated by the source) and predict  $x_4 x_5$  (i.e. estimate conditional probability  $P(x_4 x_5 | x_1 \dots x_3 = 101, y_1 \dots y_4 = 0101)$ ). For solving both problems we will use the measure  $R$  (see (17)). First we consider the case where  $P(z_1 z_2)$  is to be estimated without knowledge of sequences  $x$  and  $y$ . From (11) and (15) we obtain:

$$K_0(00) = K_0(11) = \frac{1/2}{1} \frac{3/2}{1+1} = 3/8, \quad K_0(01) = K_0(10) = \frac{1/2}{1+0} \frac{1/2}{1+1} = 1/8,$$

$$K_i(00) = K_i(01) = K_i(10) = K_i(11) = 1/4; \quad i \geq 1.$$

Having taken into account the definitions of  $\omega_i$  (16) and the measure  $R$  (17), we can calculate  $R(z_1 z_2)$  as follows:

$$R(00) = \omega_1 K_0(00) + \omega_2 K_1(00) + \dots = (1 - 1/\log 3) 3/8 + (1/\log 3 - 1/\log 4) 1/4 +$$

$$(1/\log 4 - 1/\log 5) 1/4 + \dots = (1 - 1/\log 3) 3/8 + (1/\log 3) 1/4 \approx 0.296.$$

Analogously,  $R(01) = R(10) \approx 0.204$ ,  $R(11) \approx 0.296$ .

Let us now estimate the probability  $P(z_1 z_2)$  taking into account that there are two independent samples  $y = y_1 \dots y_4 = 0101$  and  $x = x_1 \dots x_3 = 101$ . First of all we note that such estimates are based on the formula for conditional probabilities:

$$R(z|x \diamond y) = R(x \diamond y \diamond z)/R(x \diamond y).$$

First we estimate the frequencies :

$$\nu_{0101 \diamond 101}(0) = 3, \nu_{0101 \diamond 101}(1) = 4, \nu_{0101 \diamond 101}(00) = \nu_{0101 \diamond 101}(11) = 0, \nu_{0101 \diamond 101}(01) = 3,$$

$$\nu_{0101 \diamond 101}(10) = 2, \nu_{0101 \diamond 101}(010) = 1, \nu_{0101 \diamond 101}(101) = 2, \nu_{0101 \diamond 101}(0101) = 1,$$

whereas frequencies of all other tree-letters and four-letters words are 0. Then we calculate :

$$K_0(0101 \diamond 101) = \frac{1}{2} \frac{3}{4} \frac{5}{6} \frac{7}{8} \frac{1}{10} \frac{3}{12} \frac{5}{14} \approx 0.00244, K_1(0101 \diamond 101) = (2^{-1})^2 \frac{1}{2} \frac{3}{4} \frac{5}{6} \frac{1}{2} \frac{3}{4} \frac{1}{1} \approx 0.0293,$$

$$K_2(0101 \diamond 101) \approx 0.01172, \quad K_i(0101 \diamond 101) = 2^{-7}, \quad i \geq 3,$$

$$R(0101 \diamond 101) = \omega_1 K_0(0101 \diamond 101) + \omega_2 K_1(0101 \diamond 101) + \dots \approx$$

$$0.369 \cdot 0.00244 + 0.131 \cdot 0.0293 + 0.06932 \cdot 0.01172 + 2^{-7} / \log 5 \approx 0.0089.$$

In order to avoid repetitions, we estimate only one probability  $P(z_1 z_2 = 01)$ . Carrying out similar calculations, we obtain

$$R(0101 \diamond 101 \diamond 01) \approx 0.00292,$$

$$R(z_1 z_2 = 01 | y_1 \dots y_4 = 0101, x_1 \dots x_3 = 101) = R(0101 \diamond 101 \diamond 01) / R(0101 \diamond 101) \approx 0.32812.$$

If we compare this value and the estimation  $R(01) \approx 0.204$ , which is not based on the knowledge of samples  $x$  and  $y$ , we can see that the measure  $R$  uses additional information quite naturally (indeed, 01 is quite frequent in  $y = y_1 \dots y_4 = 0101$  and  $x = x_1 \dots x_3 = 101$ ).

Such generalization can be applied for many universal codes, but, generally speaking, there exist codes  $U$  for which  $U(x_1 \diamond x_2)$  is not defined for independent samples  $x_1$  and  $x_2$  and, hence, the measure  $\mu_U(x_1 \diamond x_2)$  is not defined. That is why we will not describe properties of any universal code, but for  $R$  only. For the measure  $R$  all asymptotic properties are the same for a case of one sample and several ones. More precisely, the following statement is true:

**Claim 2.** *Let  $x^1 \diamond x^2 \diamond \dots \diamond x^r$  be non-overlapping samples generated by a stationary and ergodic source and  $t$  be a total length of those samples ( $t = \sum_{i=1}^r |x^i|$ ). Then, if  $t \rightarrow \infty$ , (and  $r$  is fixed) the statements of the Theorems 1-5 are valid, when applied to  $x^1 \diamond x^2 \diamond \dots \diamond x^r$  instead of the one sample  $x_1 \dots x_t$ . (In theorems 2, 4, 5  $\mu_U$  should be changed in  $R$ .)*

The proofs are analogous to the proofs of the Theorems 1-5.

## 4 Real-valued time series

Let  $X_t$  be a time series with each  $X_t$  taking values in some interval  $\Lambda$ . The probability distribution of  $X_t$  is unknown but it is known that the time series is stationary and ergodic. Let  $\{\Pi_n\}, n \geq 1$ , be an increasing sequence of finite partitions that asymptotically generates the Borel sigma-field on  $\Lambda$ , 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.) 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  $\lambda_n$ . (For example,  $\lambda_n$  can be Lebesgue measure, counting measure, etc.) For integers  $s$  and  $n$  we define the following approximation of the density

$$p^s(x_1, \dots, x_n) = P(x_1^{[s]}, \dots, x_n^{[s]}) / \lambda_n(x_1^{[s]} \dots x_n^{[s]}). \quad (25)$$

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

$$h = \lim_{n \rightarrow \infty} E(\log p(x_{n+1}|x_1, \dots, x_n)), \quad (26)$$

where  $E$  denotes expectation with respect to  $P$ ; see [2]. We also consider

$$h_s = \lim_{n \rightarrow \infty} E(\log p^s(x_{n+1}|x_1, \dots, x_n)). \quad (27)$$

It is shown by Barron [2] that almost surely

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log p(x_1 \dots x_t) = h. \quad (28)$$

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

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log p^s(x_1, \dots, x_t) = h_s. \quad (29)$$

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

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i 2^{-|U(x_1^{[i]} \dots x_t^{[i]})|} / \lambda_t(x_1^{[i]} \dots x_t^{[i]}). \quad (30)$$

(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 the 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  $\lim_{s \rightarrow \infty} h_s = h < \infty$ , where  $h$  and  $h_s$  are relative entropy rates, see (26), (27). Then the following equality is true with probability 1:*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \left\{ \log \frac{p(x_1)}{r_U(x_1)} + \dots + \log \frac{p(x_{n+1}|x_1 \dots x_n)}{r_U(x_{n+1}|x_1 \dots x_n)} + \dots + \log \frac{p(x_t|x_1 \dots x_{t-1})}{r_U(x_t|x_1 \dots x_{t-1})} \right\} = 0. \quad (31)$$

**Proof.**

It can be seen that (31) is equivalent to the following equality.

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} = 0. \quad (32)$$

First we note that for any integer  $s$  the following obvious equality is true:  $r_U(x_1 \dots x_t) = \omega_s 2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / \lambda_t(x_1^{[s]} \dots x_t^{[s]}) (1 + \delta)$  for some  $\delta > 0$ . From this equality and (32) we immediately obtain that a.s.

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \leq \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / \lambda_t(x_1^{[s]} \dots x_t^{[s]})}. \quad (33)$$

The right part can be presented as follows:

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / \lambda_t(x_1^{[s]} \dots x_t^{[s]})} &= \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p^s(x_1, \dots, x_t) \lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \\ &+ \lim_{t \rightarrow \infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}. \end{aligned} \quad (34)$$

Having taken into account that  $U$  is the universal code and (25), we can see that the first term equals to zero. From (28) and (29) we can see that a.s. the second term is equal to  $h - h_s$ . This equality is valid for any integer  $s$  and, according to the statement of the theorem,  $\lim_{s \rightarrow \infty} h_s = h$ . Hence, the second term equals to zero, too, and we obtain the proof of (32). The theorem is proven.

**Corollary 1.**

$$\lim_{t \rightarrow \infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) = 0.$$

**Proof.** Analogously to (33) and (34) we can obtain the following enequality

$$\begin{aligned} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} &\leq \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|} / \lambda_t(x_1^{[s]} \dots x_t^{[s]})} = \log \frac{p_t^s(x_1, \dots, x_t) \lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \\ &+ \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}. \end{aligned} \quad (35)$$

for any integer  $s$ . Hence,

$$\frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) \leq E(\frac{1}{t} \log \frac{p_t^s(x_1, \dots, x_t) \lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}}) + E(\frac{1}{t} \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}). \quad (36)$$

The first term is the average redundancy of a universal code for a finite-alphabet source, hence, it tends to 0 according to the definition of the universal code. The second term tends to  $h - h_s$  for any  $s$ , hence, it is equals to zero. Corollary 1 is proven.

**Corollary 2.**

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} \int (p(x_1 \dots x_t) - r_U(x_1 \dots x_t))^2 d\lambda_t = 0,$$

$$ii) \lim_{t \rightarrow \infty} \frac{1}{t} \int |p(x_1 \dots x_t) - r_U(x_1 \dots x_t)| d\lambda_t = 0.$$

**Proof** i) immediately follows from the corollary 1 and the Pinsker's inequality (4). ii) can be derived from i) and the Jensen inequality for the function  $x^2$ .

**Theorem 7 .** *Let  $B_1, B_2, \dots$  be a sequence of measurable sets. Then the following equalities are true:*

$$i) \lim_{t \rightarrow \infty} E\left(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m))^2\right) = 0, \quad (37)$$

$$ii) E\left(\frac{1}{t} \sum_{m=0}^{t-1} |P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)|\right) = 0.$$

**Proof.** Obviously,

$$E\left(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m))^2\right) \leq \quad (38)$$

$$\begin{aligned} & \frac{1}{t} \sum_{m=0}^{t-1} E(|P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)| + \\ & |P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)|)^2. \end{aligned}$$

From the Pinsker inequality (4) and convexity of the KL divergence (3) we obtain the following inequalities

$$\begin{aligned} & \frac{1}{t} \sum_{m=0}^{t-1} E(|P(x_{m+1} \in B_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)| + \quad (39) \\ & |P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m) - R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)|)^2 \leq \\ & \frac{const}{t} \sum_{m=0}^{t-1} E\left(\log \frac{P(x_{m+1} \in B_{m+1}|x_1 \dots x_m)}{R_U(x_{m+1} \in B_{m+1}|x_1 \dots x_m)} + \log \frac{P(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)}{R_U(x_{m+1} \in \bar{B}_{m+1}|x_1 \dots x_m)}\right) \leq \\ & \frac{const}{t} \sum_{m=0}^{t-1} \left(\int p(x_1 \dots x_m) \left(\int p(x_{m+1}|x_1 \dots x_m) \log \frac{p(x_{m+1}|x_1 \dots x_m)}{r_U(x_{m+1}|x_1 \dots x_m)} d\lambda\right) d\lambda_m\right). \end{aligned}$$

Having taken into account that the last term is equal to  $\frac{const}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)})$ , from (38) and (39) and Corollary 1 we obtain (37). ii) can be derived from i) and the Jensen inequality for the function  $x^2$ . The theorem is proven.

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:*

$$i) \lim_{t \rightarrow \infty} \frac{1}{t} E \left( \sum_{m=0}^{t-1} \left( \int f(x) p(x|x_1 \dots x_m) d\lambda_m - \int f(x) r_U(x|x_1 \dots x_m) d\lambda_m \right)^2 \right) = 0, \quad (40)$$

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

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

$$\begin{aligned} & \left( \int f(x) p(x|x_1 \dots x_m) d\lambda_m - \int f(x) r_U(x|x_1 \dots x_m) d\lambda_m \right)^2 = \\ & \left( \int f(x) (p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)) d\lambda_m \right)^2 \leq M^2 \left( \int (p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)) d\lambda_m \right)^2 \\ & \leq M^2 \left( \int |p(x|x_1 \dots x_m) - r_U(x|x_1 \dots x_m)| d\lambda_m \right)^2 \leq \\ & \text{const} \int p(x|x_1 \dots x_m) \log(p(x|x_1 \dots x_m)/r_U(x|x_1 \dots x_m)) d\lambda_m. \end{aligned}$$

From these inequalities we obtain:

$$\begin{aligned} & \sum_{m=0}^{t-1} E \left( \int f(x) p(x|x_1 \dots x_m) d\lambda_m - \int f(x) r_U(x|x_1 \dots x_m) d\lambda_m \right)^2 \leq \quad (41) \\ & \sum_{m=0}^{t-1} \text{const} E \left( \int p(x|x_1 \dots x_m) \log(p(x|x_1 \dots x_m)/r_U(x|x_1 \dots x_m)) d\lambda_m \right). \end{aligned}$$

The last term can be presented as follows:

$$\begin{aligned} & \sum_{m=0}^{t-1} E \left( \int p(x|x_1 \dots x_m) \log(p(x|x_1 \dots x_m)/r_U(x|x_1 \dots x_m)) d\lambda_m \right) = \\ & \sum_{m=0}^{t-1} \int p(x_1 \dots x_m) \int p(x|x_1 \dots x_m) \log(p(x|x_1 \dots x_m)/r_U(x|x_1 \dots x_m)) d\lambda d\lambda_m = \\ & \int p(x_1 \dots x_t) \log(p(x_1 \dots x_t)/r_U(x_1 \dots x_t)) d\lambda_t. \end{aligned}$$

From this equality, (41) and Corollary 1 we obtain (40). ii) can be derived from (41) and the Jensen inequality for  $x^2$ . Theorem is proven.

## 5 The Experiments

In this part we describe the results of some experiments and a simulation study carried out in order to evaluate the efficiency of the suggested algorithms, paying the main attention to the prediction problem. The obtained results show that, in general, the described approach can be used in applications.

## 5.1 Simulations

We constructed several artificial samples created by processes with known structure and tried to predict the next value ( $x_{n+1}$ ) of the process based on  $x_1, \dots, x_n$ . WinRAR archiver (<http://www.rarlab.com>) was chosen as a code for constructing predictors. The scheme of experiments is as follows. Let  $x_1 \dots x_n$  be the generated sequence. Denote by  $x^*$  the estimation of  $x_{n+1}$ . For each  $n$  we calculate the density  $r_U(x_1 \dots x_n)$  and the average value (according to this density), which is output as the predicted value  $x^*$ .

The first process was created according to the following formula:  $x_i = \sin(\pi * i/23)$ . In this experiment we used WinRAR with the medium quality of compression. After every experiment the error of the prediction  $r_i = |x^* - x_{n+1}|$  was evaluated. We compared these values with errors of the so-called inertial predictor, where the estimation of (unknown)  $x_{n+1}$  is defined as  $x_n$  (i.e.  $x^* = x_n$ ). The obtained results are given in the following table.

Number of experiments	Length of a sample sequence (n)	Suggested	Inertial
100	1000	0.37	0.41
100	2000	0.37	0.46
100	3000	0.34	0.45

The numbers given in the first line of the table mean that 100 experiments were carried out, the length of the observed data is equal to 1000 ( $n = 1000$ ), the mean value of the error ( $\sum_{i=1}^{100} r_i/100$ ) of prediction using the suggested method is 0.37, whereas the mean value of inertial prediction is 0.41.

The second was a "random mixture" of the four following functions:  $f_1(i) = [5 * \sin(\pi * i/16)]$ ,  $f_2(i) = [7 * \sin(\pi * i/ + \pi/5)]$ ,  $f_3(i) = [8 * \sin(\pi * i/3)]$ ,  $f_4(i) = [8 * \sin(\pi * i/23)]$ . More precisely, first the length of a segment was randomly chosen according to the Poisson distribution (with a parameter  $\lambda = 0.1$ ), then the function on each segment was chosen randomly (with the probability 1/4) and values of the segment were generated according to the chosen formula. The results of this experiment are given in the table below.

Number of experiments	Length of a sample sequence (n)	Suggested	Inertial
100	2000	1.43	2.2
100	5000	2.97	4.27
100	10000	3.07	3.4

## 5.2 Prediction of currency rate

To carry out this experiments we took values EURO/USD from Forex stock (<http://www.forex.com>). The scheme of the experiments is mainly the same as in the previous section. In these experiments we used WinRAR data compression method and predictor  $R$ . First of all we carried out few experiments to find best parameters

for prediction.  $R$  showed better results than WinRAR archiver. We took independent samples and carried out experiments as it was described above. The results are given in the table below.

Number of experiments	Length of a sample sequence (n)	Suggested	Inertial
100	600	0.0150	0.0175
100	600	0.0143	0.0165
100	600	0.0131	0.0162
100	600	0.0164	0.0175

So, we can see that predictors which are based on data compression methods have reasonable performance in practice.

## 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, pp. 1165-1185, 1999.
- [2] Barron A.R. The strong ergodic theorem for dencities: generalized Shannon-McMillan-Breiman theorem. *The annals of Probability*, v.13, n.4, pp. 1292-1303, 1985.
- [3] Billingsley P., *Ergodic theory and information*, 1965. John Wiley & Sons.
- [4] Cilibrasi R., Vitanyi P.M.B. Clustering by Compression. *IEEE Transactions on Information Theory*, v. 51, n.4. 2005.
- [5] Cilibrasi R., de Wolf R., and Vitanyi P.M.B. Algorithmic Clustering of Music. *Computer Music Journal*, v. 28, n. 4, pp. 49-67, 2004.
- [6] Csiszár I., Körner J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapesht, Akadémiai Kiadó, 1981.
- [7] Csiszár I., Shields P., 2000, The consistency of the BIC Markov order estimation. *Annals of Statistics*, v. 6, pp. 1601-1619.
- [8] Darbellay G.A., Vajda I., 1998. Entropy expressions for multivariate continuous distributions. Research Report no 1920, UTIA, Academy of Science, Prague (library@utia.cas.cz).
- [9] Darbellay G.A., Vajda I., 1999. Estimatin of the mutual information with data-dependent partitions. *IEEE Trans. Inform. Theory*. **48**(5), 1061-1081.
- [10] Effros, M., Visweswariah, K., Kulkarni, S.R., Verdu, S., Universal lossless source coding with the Burrows Wheeler transform. *IEEE Trans. Inform. Theory*. 45, 1315-1321.

- [11] Feller W., 1970. An Introduction to Probability Theory and Its Applications, vol.1. John Wiley & Sons, New York.
- [12] Fitingof B.M. Optimal encoding for unknown and changing statistical messages. Problems of Information Transmission, v.2, n. 2, pp. 3-11, 1966.
- [13] Gallager R.G., 1968. Information Theory and Reliable Communication. John Wiley & Sons, New York, 1968.
- [14] Györfi, L.; Morvai, G.; Yakowitz, S.J.; Limits to consistent on-line forecasting for ergodic time series. *IEEE Transactions on Information Theory*, v. 44, n. 2, pp. 886 - 892, 1998.
- [15] Jacquet P., Szpankowski W., Apostol L. Universal predictor based on pattern matching. *IEEE Trans. Inform. Theory*, v.48, pp. 1462-1472., 2002.
- [16] Kelly J.L. *A new interpretation of information rate*, Bell System Tech. J., V. 35, pp. 917-926, 1956.
- [17] Kieffer J., 1998. Prediction and Information Theory, Preprint, (available at <ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/> )
- [18] Kieffer, J.C., En-Hui Yang, 2000. Grammar-based codes: a new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46 (3), 737 - 754.
- [19] Knuth D.E. *The art of computer programming*. Vol.2. Addison Wesley, 1981.
- [20] Kolmogorov A.N. *Three approaches to the quantitative definition of information*. Problems Inform. Transmission, v.1, 1965, pp.3-11.
- [21] Krichevsky R. *A relation between the plausibility of information about a source and encoding redundancy* Problems Inform. Transmission, v.4, n.3, 1968, pp. 48-57.
- [22] Krichevsky R. *Universal Compression and Retrieval*. Kluwer Academic Publishers, 1993.
- [23] Kullback S. Information Theory and Statistics. Wiley, New York, 1959.
- [24] Modha D.S., Masry E. Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory*, 44, n.1, 117-133.
- [25] Morvai G. , Yakowitz S.J., Algoet P.H. , 1997. Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inform. Theory*, 43, 483 - 498.
- [26] Nobel A.B., 2003. On optimal sequential prediction. *IEEE Trans. Inform. Theory*, 49(1), 83-98.
- [27] Rissanen J., 1984. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, 30(4) 629-636.

- [28] Rukhin A. and others. *A statistical test suite for random and pseudorandom number generators for cryptographic applications*. NIST Special Publication 800-22 (with revision dated May,15,2001). <http://csrc.nist.gov/rng/SP800-22b.pdf>
- [29] Ryabko B.Ya., 1984. Twice-universal coding. Problems of Information Transmission, 20(3) 173-177.
- [30] Ryabko B.Ya., 1988. Prediction of random sequences and universal coding. Problems of Inform. Transmission, 24(2) 87-96.
- [31] Ryabko B.Ya., 1990. A fast adaptive coding algorithm. Problems of Inform. Transmission, 26(4) 305-317.
- [32] Ryabko, B. Ya. The complexity and effectiveness of prediction algorithms. J. Complexity 10 (1994), no. 3, 281-295.
- [33] B. Ryabko and J. Astola. "Universal Codes as a Basis for Time Series Testing " "Statistical Methodology" v.3, pp.375-397 ,2006,
- [34] B. Ya. Ryabko, V.A. Monarev. Using information theory approach to randomness testing. Journal of Statistical Planning and Inference,2005, v. 133, n.1, pp. 95-110.
- [35] Ryabko B., Topsoe F., 2002. On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources. Journal of Complexity, 18(1) 224-241.
- [36] Ryabko D., Hutter M., Sequence prediction for non-stationary processes. In proceedings: Combinatorial and Algorithmic Foundations of Pattern and Association Discovery. Dagstuhl Seminar 2006, Germany <http://www.dagstuhl.de/06201/> see also <http://arxiv.org/pdf/cs.LG/0606077>
- [37] Savari S. A., 2000. A probabilistic approach to some asymptotics in noiseless communication. IEEE Transactions on Information Theory 46(4): 1246-1262.
- [38] Shannon C. E., A mathematical theory of communication, Bell Sys. Tech. J., vol. 27, pp. 379-423 and pp. 623-656, 1948.
- [39] Shannon C.E. Communication theory of secrecy systems, Bell Sys. Tech. J., vol. 28, pp. 656-715, 1948.
- [40] Shields, P.C., The interactions between ergodic theory and information theory., IEEE Transactions on Information Theory, v. 44, n. 6, 1998, pp. 2079 - 2093
- [41] W. Szpankowski. *Average case analysis of algorithms on sequences*. John Wiley and Sons, New York, 2001.