Fundamenta Informaticae 83 (2008) 1–20
IOS Press

Applications of Kolmogorov Complexity and Universal Codes to Nonparametric Estimation of Characteristics of Time Series

Boris Ryabko[*†]
Siberian State University of Telecommunications and Informatics
Institute of Computational Technologies of Siberian Branch of Russian Academy of Science
Kirov Street,86, 630102, Novosibirsk, Russia
boris@ryabko.net

Abstract. We consider finite-alphabet and real-valued time series and the following four problems: i) estimation of the (limiting) probability $P(x_0 \dots x_s)$ for every s and each sequence $x_0 \cdots x_s$ of letters from the process alphabet (or estimation of the density $p(x_0, \dots, x_s)$ for real-valued time series), ii) the 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, where $x_1 x_2 \cdots x_t$ are given, iii) regression and iv) classification (or so-called problems with side information).

We show that Kolmogorov complexity (KC) and universal codes (or universal data compressors), whose codeword length can be considered as an estimation of KC, 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 to the Shannon entropy of the source.)

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

1. Introduction

Kolmogorov complexity was suggested in [21] and nowadays plays an important role in the theory of algorithms and information theory. It is closely connected with such deep theoretical issues as definition

*Research was supported by Russian Foundation for Basic Research (grant no. 06-07-89025).
†Address for correspondence: Siberian State University of Telecommunications and Informatics, Institute of Computational Technologies of Siberian Branch of Russian Academy of Science, Kirov Street,86, 630102, Novosibirsk, Russia

of randomness and logical foundation of probability theory (see [5, 6, 16, 25, 40, 42, 43, 44]). In this paper we show that Kolmogorov complexity and universal codes (or data compression methods) can be applied to some problems in the framework of mathematical statistics. It is important to note that nowadays there are many different classes of universal codes and such of them as CTW code [45, 46], Grammar-based codes [19], LZ-codes [9] and some others have shown there great efficiency as data compressors and predictors [3, 18].

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. It is supposed that the probability distribution (or distribution of limiting probabilities) $P(x_1 = a_{i_1}, x_2 = a_{i_2}, \ldots, x_t = a_{i_t})$ (or the density $p(x_1, x_2, \ldots, x_t)$) is unknown, but we are given either one sample $x_1 \ldots x_t$ or several (r) independent samples $x^1 = x^1_1 \ldots x^1_{t_1}, \ldots, x^r = x^r_1 \ldots x^r_{t_r}$ generated by the source. (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. That is why precise estimations of the probability distribution and the density can be used for prediction, regression, estimation, etc. In this paper we follow this 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, 11, 18, 19, 26, 30, 41]. We show that Kolmogorov complexity and 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, Kolmogorov complexity K and any universal code U the following equalities are valid with probability 1:

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

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

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 $|U(x_1 \cdots x_t)|$ and Kolmogorov complexity $K(x_1 \cdots x_t)$ can be used as estimates of the logarithm of the unknown probability and, obviously, $2^{-|U(x_1 \cdots x_t)|}$ and $2^{-K(x_1 \cdots x_t)}$ can be considered as the estimations of $P(x_1 \cdots x_t)$. They can be viewed as non-parametric estimations 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) and universal codes were applied for solving prediction problems [32].

We would like to emphasize that, on the one hand, all results will be 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., [13, 19, 28, 29, 30, 45, 46]) and have 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 estimation, 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 [7, 8] 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 statistical hypotheses concerned with stationary and ergodic time series [35, 36, 37].

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

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 $x_1 x_2 \cdots$ of letters from some set (or alphabet) A. It will be convenient at first to describe briefly the prediction problem. Let the source generate a message $x_1 \ldots x_{t-1} x_t,\ x_i \in A$ for all i, and the following letter x_{t+1} needs to be predicted. This problem can be traced back to Laplace who considered the problem how to estimate the probability that the sun will rise tomorrow, given that it has risen every day since Creation (see [14]). 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:

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

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 [32] and now is often called on-line prediction or universal prediction [1, 18, 26, 27]. 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 [17, 34]) and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor γ defines a measure by the following equation

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

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 (or estimation of the measure) 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 next natural question is how to estimate the precision of a prediction and of 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 as follows:

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

where γ is the estimation of an unknown condition probability. It is well-known that for any distributions P and γ the KL divergence is nonnegative and equals 0 if and only if $P(a) = \gamma(a)$ for all a, see, for ex., [15]. The following inequality (Pinsker's inequality)

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

connects the KL divergence with the so-called variation distance

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

where P and Q are distributions over A, see [10]. 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

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

It is shown in [33] that the error of Laplace predictor L_0 goes to 0 for any i.i.d. source P. More precisely, it is proven that

$$\rho^t(P\|L_0) < (|A| - 1) \log e/(t + 1) \tag{6}$$

for any source P, [33], see also [38]. 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 [32] (see also [1]) that such predictors do not exist for the class of all stationary and ergodic sources (generating 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 may not go to 0, if the predictor is applied to an arbitrary stationary and ergodic source, that is why it is difficult to use (3) and (5) to compare different predictors.

On the other hand, it is shown in [32], 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

$$\bar{\rho}_{\gamma,P}(x_1...x_t) = t^{-1} \left(\log(P(x_1...x_t)/\gamma(x_1...x_t)\right) \tag{7}$$

and

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

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:

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

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).

A natural problem is to find a predictor whose error is minimal (for i.i.d. sources). This problem was considered and solved by Krichevsky in [22], see also [23]. He suggested the following predictor:

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

where, as before, $\nu_{x_1 \cdots x_t}(a)$ is the count of letter a occurring in the word $x_1 \dots x_t$. 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:

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

It is known that

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

where $\Gamma(\,)$ is the gamma function (see for definition, for ex., [20]). So, (11) can be presented as follows:

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

For this predictor

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

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 [22, 23].

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 ($A^* = \bigcup_{i=1}^{\infty} A^i$).

The time shift T on A^∞ is defined as $T(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots)$. A process P is called stationary if it is T-invariant: $P(T^{-1}B) = P(B)$ for every Borel set $B \subset A^\infty$. A stationary process is called ergodic if every T-invariant set has probability 0 or 1: $P(B) = 0$ or 1 whenever $T^{-1}B = B$. For more details on stationarity and ergodicity see [4, 15].

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-memory 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 [38]. 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 (13), we can present the Krichevsky predictors for $M_m(A)$ as follows:

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

where $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va),\ x = x_1...x_t$. It is worth noting that the representation (12) can be more convenient for carrying out calculations. Let us consider an example. For the word $OOIOIIOOIIIOIO$ 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 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 $\{\omega = \omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_1 = 1 - 1/\log 3,\ \ldots,\ \omega_i\ = 1/\log(i+1) - 1/\log(i+2),\ \ldots\ . \tag{16}$$

(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:

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

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

$$H(P) = \lim_{m\to\infty} -\frac{1}{m} \sum_{v \in A^m} P(v) \log P(v). \tag{18}$$

Theorem 2.1. ([32])
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,

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

2.3. Kolmogorov complexity 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) φ is defined as a set of mappings φ_n such that $\varphi_n : A^n \to \{0,1\}^*$, $n = 1, 2, \ldots$ 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 $u_1 u_2 ... u_r$. 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 φ 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 2.1. Let φ be a uniquely decodable code over an alphabet A. Then for any integer n there exists a measure μ_φ on A^n such that

$$-\log \mu_\varphi(u) \;\leq\; |\varphi(u)| \tag{19}$$

for any u from A^n .

Clearly, the statement holds for the measure $\mu_\varphi(u) = 2^{-|\varphi(u)|} / \Sigma_{u \in A^n} \, 2^{-|\varphi(u)|}$.

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

In this paper we will use the so-called prefix Kolmogorov complexity, whose precise definition can be found in [16, 25]. Its main properties can be described as follows. There exists a uniquely decodable code κ such that i) there is an algorithm of decoding (i.e. there is a Turing machine, which maps $\kappa(u)$ to u for any $u \in A^*$) and ii) for any uniquely decodable code ψ, whose decoding is algorithmically realizable, there exists a constant C_ψ such that

$$|\kappa(u)| - |\psi(u)| < C_\psi \tag{20}$$

for any $u \in A^*$. The prefix Kolmogorov complexity $K(u)$ is defined as the length of $\kappa(u)$: $K(u) = |\kappa(u)|$. The code κ is not unique, but the second property means that codelengths of two codes κ_1 and κ_2, for which i) and ii) are true, are equal up to a constant: $|\,|\kappa_1(u)| - |\kappa_2(u)|\,| < C_{1,2}$ for any word u (and the constant $C_{1,2}$ does not depend on u, see (20).) So, $K(u)$ is defined up to a constant.

In what follows we call this value "Kolmogorov complexity" and uniquely decodable codes just "codes".

We can see from ii) that the code κ is asymptotically (up to a constant) the best method of data compression, but it turns out that there is no algorithm that can calculate the codeword $\kappa(u)$ (and even $K(u)$). That is why the code κ (and Kolmogorov complexity) cannot be used for practical data compression directly. On the other hand, so-called universal codes can be realized and, in a certain sense, can be used instead of the optimal code κ, if they are applied for compression of sequences generated by a stationary and ergodic source. For their description we recall that (as it is known in Information Theory) sequences $x_1...x_t$, generated by a source P, can be "compressed" to $-\log P(x_1...x_t)$ bits and, on the other hand, there is no code ψ for which the average codeword length ($\Sigma_{x_1...x_t \in A^t} P(x_1...x_t)|\psi(x_1...x_t)|$) is less than $-\Sigma_{x_1...x_t \in A^t} P(x_1...x_t) \log P(x_1...x_t)$. The universal codes can reach the lower bound $-\log P(x_1...x_t)$ asymptotically for any stationary and ergodic source P with probability 1. The formal definition is as follows: a code U is universal if for any stationary and ergodic source P the following equalities are valid:

$$\lim_{t \to \infty} |U(x_1 \dots x_t)|/t = H(P) \tag{21}$$

with probability 1, and

$$\lim_{t \to \infty} E(|U(x_1 \dots x_t)|)/t = H(P), \tag{22}$$

where $H(P)$ is the Shannon entropy of P, $E(f)$ is the expected value of f. So, informally speaking, universal codes estimate the probability characteristics of the source P and use them for efficient "compression". Both equalities are true if we replace the length of the code $(|U(x_1 \dots x_t)|)$ by $K(x_1 \dots x_t)$, because, in a certain sense, Kolmogorov complexity is the length of codeword of the best universal code. That is why in what follows we will speak about universal codes taking into account that all statements are true for Kolmogorov complexity.

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 3.1. Let U be a universal code and

$$\mu_U(u) = 2^{-|U(u)|} / \Sigma_{v \in A^{|u|}} \, 2^{-|U(v)|}. \tag{23}$$

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,

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

$$iii) \ \lim_{t\to\infty} \frac{1}{t} \sum_{u \in A^t} P(u) \, |P(u) - \mu_U(u)| \ = \ 0.$$

Proof:
The proof is based on Shannon-MacMillan-Breiman Theorem which states that for any stationary and ergodic source P

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

with probability 1, see [4, 15]. From this equality and (21) we obtain the statement i). The second statement follows from the definition of Shannon entropy (18) and (22), whereas iii) follows from ii) and the Pinsker's inequality (4). □

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), which can be applied to estimation. For example, the above described measure R is based on the code from [31, 32] and can be applied for probability estimation. More precisely, Theorem 3.1 (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 describes these properties (whereas all other theorems are valid for all universal codes and corresponding measures, including the measure R).

Theorem 3.2. 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:
$$\frac{1}{t} \sum_{u \in A^t} P(u) \log(P(u)/R(u)) \ \le \ \frac{(|A|-1)|A|^k \log t}{2\,t} + O(\frac{1}{t}),$$

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

iii) Let Θ be a set of stationary and ergodic processes such that there exists a measure μ_Θ for which the estimation error of the probability goes to 0 uniformly:

$$\lim_{t\to\infty} \sup_{P\in\Theta} \Big(\frac{1}{t} \sum_{u\in A^t} P(u) \log(P(u)/\mu_\Theta(u)) \Big) = 0.$$

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

$$\lim_{t\to\infty} \sup_{P\in\Theta} \Big(\frac{1}{t} \sum_{u\in A^t} P(u) \log(P(u)/R(u)) \Big) = 0.$$

The proof can be found in [31, 32].

3.2. Prediction

As we mentioned above, any universal code U can be applied for prediction. Namely, the measure μ_U (23) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1...x_t) = \mu_U(x_1...x_t x_{t+1})/\mu_U(x_1...x_t). \tag{24}$$

Theorem 3.3. Let U be a universal code and P be any stationary and ergodic process. Then

$$i)\ \lim_{t\to\infty} \frac{1}{t} \,\{E(\log \frac{P(x_1)}{\mu_U(x_1)}) + E(\log \frac{P(x_2|x_1)}{\mu_U(x_2|x_1)}) + \ldots + E(\log \frac{P(x_t|x_1...x_{t-1})}{\mu_U(x_t|x_1...x_{t-1})})\} \;=\; 0,$$

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

and

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

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

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

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} \sum_{x_1\ldots x_i\in A^i} P(x_1 \ldots x_i) (\sum_{a\in A} |P(a|x_1 \ldots x_i) - \mu_U(a|x_1 \ldots x_i)|)^2 \le$$

$$\lim_{t\to\infty} \frac{const}{t} \sum_{i=0}^{t-1} \sum_{x_1\ldots x_i\in A^i} P(x_1 \ldots x_i) \sum_{a\in A} P(a|x_1 \ldots x_i) \log \frac{P(a|x_1 \ldots x_i)}{\mu_U(a|x_1 \ldots x_i)} =$$

$$\lim_{t\to\infty} (\frac{const}{t} \sum_{x_1\ldots x_t\in A^t} P(x_1 \ldots x_t) \log(P(x_1 \ldots x_t)/\mu(x_1 \ldots x_t))).$$

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. $\square$

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:

$$\lim_{t\to\infty} \log \frac{P(x_{t+1}|x_1...x_t)}{R(x_{t+1}|x_1...x_t)} = 0$$

with probability 1, where $R(x_{t+1}|x_1...x_t) = R(x_1...x_t x_{t+1})/R(x_1...x_t)$; see [33].

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 [39].

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. 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 the parts 3.1 and 3.2. More precisely, for any universal code U and the corresponding measure μ_U (23) 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 3.4. Let U be a universal code and P be any stationary and ergodic process. Then

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

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

$$ii)\ \lim_{t\to\infty} E(\frac{1}{t}\ \sum_{i=0}^{t-1}(P(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1}))-$$

$$\mu_U(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1}))^2) \quad = \ 0\,,$$

and

$$iii)\ \lim_{t\to\infty} E(\frac{1}{t}\ \sum_{i=0}^{t-1}|P(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1}))-$$

$$\mu_U(x_{i+1}|(x_1, y_1), ..., (x_i, y_i), y_{i+1})|) \quad = \ 0\,.$$

Proof:
The following inequality follows from the nonnegativity of the KL divergency (see (4)), whereas equality is obvious.

$$E(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}) + E(\log \frac{P(x_2|(x_1,y_1),y_2)}{\mu_U(x_2|(x_1,y_1),y_2)}) + \ldots \le E(\log \frac{P(y_1)}{\mu_U(y_1)})$$

$$+E(\log \frac{P(x_1|y_1)}{\mu_U(x_1|y_1)}) + E(\log \frac{P(y_2|(x_1,y_1)}{\mu_U(y_2|(x_1,y_1)}) + E(\log \frac{P(x_2|(x_1,y_1),y_2)}{\mu_U(x_2|(x_1,y_1),y_2)}) + \ldots$$

$$= E(\log \frac{P(x_1,y_1)}{\mu_U(x_1,y_1)}) + E(\log \frac{P((x_2,y_2)|(x_1,y_1))}{\mu_U((x_2,y_2)|(x_1,y_1))}) + \ldots.$$

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

$$\lim_{t\to\infty} \frac{1}{t} E(\log \frac{P(x_1,y_1)}{\mu_U(x_1,y_1)}) + E(\log \frac{P((x_2,y_2)|(x_1,y_1))}{\mu_U((x_2,y_2)|(x_1,y_1))}) + \ldots$$

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

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. □

3.4. The case of several independent samples

Now we extend our consideration to the case where the sample is presented as several independent samples $x^1 = x^1_1 \ldots x^1_{t_1}$, $x^2 = x^2_1 \ldots x^2_{t_2}, ..., x^r = x^r_1 \ldots x^r_{t_r}$ generated by a source. More precisely, we will suppose that all sequences were independently 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 $x^1 \diamond x^2 \diamond \ldots \diamond x^r$ and define $\nu_{x^1 \diamond x^2 \diamond \ldots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. 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:

$$K_m(x^1 \diamond x^2 \diamond ... \diamond x^r) = \tag{25}$$

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

whereas the definition of R is the same (see (17)). (Here, as before, $\bar\nu_{x^1 \diamond x^2 \diamond \ldots \diamond x^r}(v) = \sum_{a\in A} \nu_{x^1 \diamond x^2 \diamond \ldots \diamond x^r}(va)$. Note, that $\bar\nu_{x^1 \diamond x^2 \diamond \ldots \diamond x^r}(\,) = \sum_{i=1}^r t_i$ 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 $y = y_1 \ldots y_4 = 0101$ and $x = x_1 \ldots x_3 = 101$, 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 $z_1 z_2 \ldots$ can be considered as an independent sequence, generated by the source) and predict $x_4 x_5$ (i.e. estimate conditional probability $P(x_4 x_5 | x_1 \ldots x_3 = 101, y_1 \ldots y_4 = 0101)$. For solving both problems we will use the measure R (see (17)). First we consider the case where $P(z_1z_2)$ is to be estimated without knowledge of sequences x and y. From (11) and (15) we obtain:

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

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

Having taken into account the definitions of ω_i (16) and the measure R (17), we can calculate $R(z_1z_2)$ as follows:

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

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

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

Let us now estimate the probability $P(z_1z_2)$ taking into account that there are two independent samples $y = y_1 \ldots y_4 = 0101$ and $x = x_1 \ldots x_3 = 101$. 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 three-letters and four-letters words are 0. Then we calculate :

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

$$\approx 0.0293, \quad K_2(\,0101 \diamond 101) \approx 0.01172, \quad K_i(\,0101 \diamond 101) = 2^{-7},\ i \geq 3,$$

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

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

In order to avoid repetitions, we estimate only one probability $P(z_1z_2 = 01)$. Carrying out similar calculations, we obtain

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

$$R(z_1z_2 = 01|y_1 \ldots y_4 = 0101, x_1 \ldots x_3 = 101) =$$

$$R(0101 \diamond 101 \diamond 01)/R(\,0101 \diamond 101) \approx 0.32812.$$

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 $y = y_1 \ldots y_4 = 0101$ and $x = x_1 \ldots x_3 = 101$).

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 and, hence, the measure $\mu_U(x_1 \diamond x_2)$ is not defined. That is why we will not describe properties of R and not of universal codes in general. For the measure R all asymptotic properties are the same for the cases of one sample and several samples. More precisely, the following statement is true:

Claim 3.1. Let $x^1, x^2, ..., x^r$ be independent sequences generated by a stationary and ergodic source and t be a total length of those sequences ($t = \sum_{i=1}^{r} |x^i|$). Then, if $t \to \infty$, (and r is fixed) the statements of the Theorems 1-5 are valid, when applied to $x^1 \diamond x^2 \diamond ... \diamond x^r$ instead of $x_1 \dots x_t$. (In theorems 2, 4, 5 μ_U should be changed to 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 Λ. The probability distribution of X_t is unknown but it is known that the time series is stationary and ergodic. Let $\{\Pi_n\}, n \geq 1$, 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; see [12]). Suppose that the joint distribution P for $(X_1, \dots, X_n)$ has a probability density function $p(x_1, \dots, x_n)$ 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

$$p^s(x_1, \dots, x_n) = P(x_1^{[s]}, \dots, x_n^{[s]}) / \lambda_n(x_1^{[s]} \dots x_n^{[s]}). \tag{26}$$

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

$$h = \lim_{n\to\infty} E(\log p(x_{n+1}|x_1, \dots, x_n)), \tag{27}$$

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

$$h_s = \lim_{n\to\infty} E(\log p^s(x_{n+1}|x_1, \dots, x_n)). \tag{28}$$

It is shown in [2] that almost surely

$$\lim_{t\to\infty} \frac{1}{t} \log p(x_1 \dots x_t) = h. \tag{29}$$

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

$$\lim_{t\to\infty} \frac{1}{t} \log p^s(x_1, \dots, x_t) = h_s. \tag{30}$$

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

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i 2^{-|U(x_1^{[i]} \dots x_t^{[i]})|} / \lambda_t(x_1^{[i]} \dots x_t^{[i]}) \,. \tag{31}$$

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

$$\lim_{s\to\infty} h_s = h < \infty, \tag{32}$$

where h and h_s are relative entropy rates, see (27), (28). Then

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 ... x_t)}{r_U(x_1 ... x_t)} = 0 \tag{33}$$

with probability 1 and

$$\lim_{t\to\infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) \; = \; 0\,. \tag{34}$$

Proof:
First we note that for any integer s the following obvious equality is true: $r_U(x_1 \dots x_t) = \omega_s \, \mu_U(x_1^{[s]} \dots x_t^{[s]})/\lambda_t(x_1^{[s]} \dots x_t^{[s]})\,(1+\delta)$ for some $\delta > 0$. From this equality, (23) and (33) we immediately obtain that a.s.

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \;\le \lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/\lambda_t(x_1^{[s]} \dots x_t^{[s]})}\,. \tag{35}$$

The right part can be presented as follows:

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

$$= \lim_{t\to\infty} \frac{1}{t} \log \frac{p^s(x_1 \dots x_t)\;\lambda_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} + \lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{p^s(x_1 \dots x_t)}. \tag{36}$$

Having taken into account that U is the universal code, (26) and Theorem 1, we can see that the first term in the limit $t \to \infty$ equals to zero. From (29) and (30) 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 (32), the second term equals to zero too, and we obtain (34). The first statement is proven.

From (35) and (36) we can can see that

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

$$+E\,\log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}. \tag{37}$$

The first term is the average redundancy of the universal code for a finite- alphabet source, hence, according to Theorem 1, it tends to 0. The second term tends to $h_s - h$ for any s and from (32) we can see that it is equal to zero. The second statement is proven. □

We have seen that the requirement (32) plays an important role in the proof. A natural question is whether there exist processes for which (32) is valid. The answer is positive. For example, let Λ be an interval $[-1,1]$, λ_n be Lebesgue measure and a considered process is Markovian with conditional density

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

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

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

It is easy to see that (32) is true for any $\alpha \in (0,1)$.

The following theorem describes properties of conditional probabilities $r_U(x|x_1...x_m) = r_U(x_1...x_m x) \;/ r_U(x_1...x_m)$ 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 4.2. Let $B_1, B_2, ...$ be a sequence of measurable sets. Then the following equalities are true:

$$i) \lim_{t\to\infty} E(\frac{1}{t}\sum_{m=0}^{t-1}(P(x_{m+1}\in B_{m+1}|x_1...x_m) - R_U(x_{m+1}\in B_{m+1}|x_1...x_m))^2) = 0\,, \tag{38}$$

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

Proof:
Obviously,

$$E(\frac{1}{t}\sum_{m=0}^{t-1}(P(x_{m+1}\in B_{m+1}|x_1...x_m) - R_U(x_{m+1}\in B_{m+1}|x_1...x_m))^2) \leq \tag{39}$$
$$\frac{1}{t}\sum_{m=0}^{t-1}E(|P(x_{m+1}\in B_{m+1}|x_1...x_m) - R_U(x_{m+1}\in B_{m+1}|x_1...x_m)|+$$
$$|P(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m) - R_U(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)|)^2.$$

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

$$\frac{1}{t}\sum_{m=0}^{t-1}E(|P(x_{m+1}\in B_{m+1}|x_1...x_m) - R_U(x_{m+1}\in B_{m+1}|x_1...x_m)|+ \tag{40}$$
$$|P(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m) - R_U(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)|)^2 \leq$$
$$\frac{const}{t}\sum_{m=0}^{t-1}E((\log\frac{P(x_{m+1}\in B_{m+1}|x_1...x_m)}{R_U(x_{m+1}\in B_{m+1}|x_1...x_m)} + \log\frac{P(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)}{R_U(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)}) \leq$$
$$\frac{const}{t}\sum_{m=0}^{t-1}(\int p(x_1...x_m)(\int p(x_{m+1}|x_1...x_m))\log\frac{p(x_{m+1}|x_1...x_m)}{r_U(x_{m+1}|x_1...x_m)}d\lambda)d\lambda_m).$$

Having taken into account that the last term is equal to $\frac{const}{t}E(\log \frac{p(x_1...x_t)}{r_U(x_1...x_t)})$, from (39), (40) and (34) we obtain (38). ii) can be derived from i) and the Jensen inequality for the function x^2. □

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

$$i) \lim_{t\to\infty} \frac{1}{t} E(\sum_{m=0}^{t-1} (\int f(x) p(x|x_1...x_m) d\lambda_m - \int f(x) r_U(x|x_1...x_m) d\lambda_m)^2) = 0, \quad (41)$$

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

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

$$(\int f(x)p(x|x_1...x_m)d\lambda_m - \int f(x)r_U(x|x_1...x_m)d\lambda_m)^2 =$$
$$(\int f(x)(p(x|x_1...x_m) - r_U(x|x_1...x_m))d\lambda_m)^2$$
$$\le M^2(\int (p(x|x_1...x_m) - r_U(x|x_1...x_m))d\lambda_m)^2$$
$$\le M^2(\int |p(x|x_1...x_m) - r_U(x|x_1...x_m)|d\lambda_m)^2 \le$$
$$const \int p(x|x_1...x_m) \log(p(x|x_1...x_m)/r_U(x|x_1...x_m)d\lambda_m.$$

From these inequalities we obtain:

$$\sum_{m=0}^{t-1} E(\int f(x)p(x|x_1...x_m)d\lambda_m - \int f(x)r_U(x|x_1...x_m)d\lambda_m)^2) \le \quad (42)$$
$$\sum_{m=0}^{t-1} const\, E(\int p(x|x_1...x_m) \log(p(x|x_1...x_m)/r_U(x|x_1...x_m))d\lambda_m.$$

The last term can be presented as follows:

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

From this equality, (42) and (34) we obtain (41). ii) can be derived from (42) and the Jensen inequality for x^2. □

References

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

[2] Barron, A.R.: The strong ergodic theorem for dencities: generalized Shannon-McMillan-Breiman theorem. *The annals of Probability*, 13(4), 1985, 1292-1303.

[3] Begleiter, R., El-Yaniv, R.: Superior Guarantees for Sequential Prediction and Lossless Compression via Alphabet Decomposition. *Journal of Machine Learning Research,* 7, 2006, 379-411.

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

[5] Calude, C. S., Staiger, L., Svozil, K.: Randomness Relative to Cantor Expansions. *CDMTCS Research Report Series*, 2003, http://www.cs.auckland.ac.nz/CDMTCS//researchreports/213cris.pdf

[6] Calude, C.S., Staiger, L., Terwijn, S.A.: On Partial Randomness. *CDMTCS Research Report Series,* 2004, http://www.cs.auckland.ac.nz/CDMTCS//researchreports/239cris.pdf

[7] Cilibrasi, R., Vitanyi, P.M.B.: Clustering by Compression. *IEEE Transactions on Information Theory,* 51(4), 2005.

[8] Cilibrasi R., de Wolf R., Vitanyi P.M.B.: Algorithmic Clustering of Music. *Computer Music Journal,* 28(4), 2004, 49-67.

[9] Cover, T.M., Thomas J.A.: *Elements of Information Theory.* JohnWiley and sons, 1991.

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

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

[12] Darbellay, G.A., Vajda, I.: Estimatin of the mutual information with data-dependent partitions. *IEEE Trans. Inform. Theory.* 48(5), 1999, 1061-1081.

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

[14] Feller, W.: *An Introduction to Probabability Theory and Its Applications*, vol.1. John Wiley & Sons, New York, 1970.

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

[16] Hutter, M.: *Universal Artificial Intelligence. Sequential Decisions based on algorithmic probability,* Springer-Verlag, 2005.

[17] Kelly, J.L.: A new interpretation of information rate, *Bell System Tech. J.*, 35, 1956, 917-926.

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

[19] Kieffer, J.C., En-Hui Yang.: Grammar-based codes: a new class of universal lossless source codes. *IEEE Transactions on Information Theory,* 46 (3),2000, 737 - 754.

[20] Knuth, D.E.: *The art of computer programming.* Vol.2. Addison Wesley, 1981.

[21] Kolmogorov, A.N.: Three approaches to the quantitative definition of information, *Problems of Inform. Transmission,* 1, 1965, 3-11.

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

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

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

[25] Li, M., Vitanyi, P.: *An Introduction to Kolmogorov Complexity and Its Applications,* Springer-Verlag, New York, 2nd Edition, 1997.

[26] Modha, D.S., Masry, E.: Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory,* 44(1), 1998, 117-133.

[27] Nobel, A.B.: On optimal sequential prediction. *IEEE Trans. Inform. Theory,* 49(1), 2003, 83-98.

[28] Rissanen, J.: Generalized Kraft inequality and arithmetic coding, *IBM J. Res. Dev.*, 20(5), 1976, 198-203.

[29] Rissanen, J.: Modeling by shortest data description. *Automatica,* v.14, 1978, 465-471.

[30] Rissanen, J.: Universal coding, information, prediction, and estimation., *IEEE Trans. Inform. Theory,* 30(4), 1984, 629-636.

[31] Ryabko, B.Ya.: Twice-universal coding. *Problems of Information Transmission,* 20(3), 1984, 173-177.

[32] Ryabko, B.Ya.: Prediction of random sequences and universal coding. *Problems of Inform. Transmission,* 24(2), 1988, 87-96.

[33] Ryabko, B.Ya.: A fast adaptive coding algorithm. *Problems of Inform. Transmission,* 26(4), 1990, 305-317.

[34] Ryabko, B. Ya.: The complexity and effectiveness of prediction algorithms. *J. Complexity,* 10 (3), 1994, 281-295.

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

[36] Ryabko, B., Astola, J., Gammerman, A.: Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science,* 359, 2006, 440-448.

[37] Ryabko, B. Ya., Monarev, V.A.: Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference,* 2005, 133(1), 95-110.

[38] Ryabko, B., Topsoe, F.: On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources. *Journal of Complexity,* 18(1), 2002, 224-241.

[39] Ryabko, D., Hutter, M.: Sequence prediction for non-stationary processes. In proceedings: IEEE International Symposium on Information Theory (ISIT 2007), 2007, 2346-2350.

see also http://arxiv.org/pdf/cs.LG/0606077

[40] Staiger, L.: The Kolmogorov complexity of infinite words. *CDMTCS Research Report Series,* 2006, http://www.cs.auckland.ac.nz/CDMTCS//researchreports/279ludwig.pdf

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

[42] Uspenskii, V.A., Semenov, A.L., Shen, A.K.: Can an individual sequence of zeros and ones be random?, *Russian Mathematical Surveys,* 45, 1990.

[43] Vereshchagin, N., Vitanyi, P.M.B.: Kolmogorov's structure functions with application to the foundations of model selections. In: Proc. *43th Symposium on Foundations of Computer Science,* 2002, 751- 760.

[44] Vitanyi, P.M.B., Li, M.: Minimum description length induction, Bayesianism, and Kolmogorov complexity, *IEEE Trans. Inform. Theory*, 46(2), 2000, 446-464.

[45] Willems, F.M.J., Shtarkov, Y.M., Tjalkens, Tj.J.: The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Inform. Theory*, 41, 1995, 653-664.

[46] Willems, F.M.J., Shtarkov, Y.M., Tjalkens, Tj.J.: Reflections on 'The Context-Tree Weighting Method: Basic Properties', *Newsletter of the IEEE Information Theory Society*, 47(1), 1997.