Chapter 1

Applications of Universal Source Coding to Statistical Analysis of Time Series

Boris Ryabko*

Siberian State University of Telecommunications and Informatics and Institute of Computational Technology of Siberian Branch of Russian Academy of Science

boris@ryabko.net †

We show how universal codes can be used for solving some of the most important statistical problems for time series. By definition, a universal code (or a universal lossless data compressor) can compress any sequence generated by a stationary and ergodic source asymptotically to the Shannon entropy, which, in turn, is the best achievable ratio for lossless data compressors.

We consider finite-alphabet and real-valued time series and the following problems: estimation of the limiting probabilities for finite-alphabet time series and estimation of the density for real-valued time series, the on-line prediction, regression, classification (or problems with side information) for both types of the time series and the following problems of hypothesis testing: goodness-of-fit testing, or identity testing, and testing of serial independence. It is important to note that all problems are considered in the framework of classical mathematical statistics and, on the other hand, everyday methods of data compression (or archivers) can be used as a tool for the estimation and testing.

It turns out, that quite often the suggested methods and tests are more powerful than known ones when they are applied in practice.

Contents

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

†Kirov Street 86, Novosibirsk, 630102, Russia. URL: http://boris.ryabko.net/.

1.1. Introduction

Since C. Shannon published the paper "A mathematical theory of communication" ,[47] the ideas and results of Information Theory have played an important role in cryptography,[26,48] mathematical statistics,[3,8,25] and many other fields,[6,7] which are far from telecommunications. Universal coding, which is a part of Information Theory, also has been efficiently applied in many fields since its discovery.[13,21] Thus, application of results of universal coding, initiated in 1988 ,[35] created a new approach to prediction.[1,19,27,28] Maybe the most unexpected application of data compression ideas arises in experiments that show that some ant species are capable of compressing messages and are capable of adding and subtracting small numbers.[30,43]

In this chapter we describe a new approach to estimation, prediction and hypothesis testing for time series, which was suggested recently.[35,38,42] This approach is based on ideas of universal coding (or universal data compression). We would like to emphasize that everyday methods of data compression (or archivers) can be directly used as a tool for estimation and hypothesis testing. It is important to note that the modern archivers (like *zip*, *arj, rar,* etc.) are based on deep theoretical results of the source coding theory[10,20,24,32,46] and have shown their high efficiency in practice because archivers can find many kinds of latent regularities and use them for compression.

It is worth noting that this approach was applied to the problem of randomness testing.[42] This problem is quite important for practice; in particular, the National Institute of Standards and Technology of USA (NIST)has suggested "A statistical test suite for random and pseudorandom number generators for cryptographic applications",[33] which consists of 16 tests. It has turned out that tests which are based on universal codes are more powerful than the tests suggested by NIST .[42]

The outline of this chapter is as follows. The next section contains some necessary definitions and facts about predictors, codes, hypothesis testing and description of one universal code. The section 1.3 and 1.4 are devoted to problems of estimation and hypothesis testing, correspondingly, for the case of finite-alphabet time series. The case of infinite alphabets is considered in the 1.5 section. All proofs are given in Appendix, but some intuitive indication are given in the body of the chapter.

1.2. Definitions and Statements of the Problems

1.2.1. *Estimation and Prediction for I.I.D. Sources*

First we consider a source with unknown statistics which generates sequences $x_1x_2\cdots$ of letters from some set (or alphabet) A. It will be convenient now to describe briefly the prediction problem. Let the source generate a message $x_1 \dots x_{t-1}x_t,\ x_i \in A$ for all i, and the following letter x_{t+1} needs to be predicted. This problem can be traced back to Laplace[11,29] who considered the problem of estimation of 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, $x_1 \dots x_{t-1}x_t = 00\dots0$.

Laplace suggested the following predictor:

$$L_0(a|x_1\cdots x_t) = (\nu_{x_1\cdots x_t}(a)+1)/(t+|A|), \tag{1.1}$$

where $\nu_{x_1\cdots x_t}(a)$ denotes the count of letter a occurring in the word $x_1 \dots x_{t-1}x_t$. It is important to note that the predicted probabilities cannot be equal to zero even through a certain letter did not occur in the word $x_1 \dots x_{t-1}x_t$.

Example 1.1. Let $A = \{0,1\}$, $x_1...x_5 = 01010$, then the Laplace prediction is as follows: $L_0(x_6 = 0|x_1...x_5 = 01010) = (3+1)/(5+2) = 4/7, L_0(x_6 = 1|x_1...x_5 = 01010) = (2+1)/(5+2) = 3/7$. In other words, 3/7 and 4/7 are estimations of the unknown probabilities $P(x_{t+1} = 0|x_1 \dots x_t = 01010)$ and $P(x_{t+1} = 1|x_1 \dots x_t = 01010)$. (In what follows we will use the shorter notation: $P(0|01010)$ and $P(1|01010)$).

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 1988[35] and now is often called on-line prediction or universal prediction.[1,19,27,28] 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[17,37] and, in fact, all obtained results can be easily transferred from one model to the other.

Any predictor γ defines a measure (or an estimation of probability) by the following equation

$$\gamma(x_1...x_t) = \prod_{i=1}^{t} \gamma(x_i|x_1...x_{i-1}). \tag{1.2}$$

And, vice versa, any measure γ (or estimation of probability) defines a predictor:

$$\gamma(x_i|x_1 ... x_{i-1}) = \gamma(x_1 ... x_{i-1}x_i)/\gamma(x_1 ... x_{i-1}). \tag{1.3}$$

Example 1.2. Let us apply the Laplace predictor for estimation of probabilities of the sequences 01010 and 010101. From (1.2) we obtain $L_0(01010) = \frac{1}{2}\frac{1}{3}\frac{2}{4}\frac{2}{5}\frac{3}{6} = \frac{1}{60}$, $L_0(010101) = \frac{1}{60}\frac{3}{7} = \frac{1}{140}$. Vice versa, if for some measure (or a probability estimation) χ we have $\chi(01010) = \frac{1}{60}$ and $\chi(010101) = \frac{1}{140}$, then we obtain from (1.3) the following prediction, or the estimation of the conditional probability, $\chi(1|01010) = \frac{1/140}{1/60} = \frac{3}{7}$.

Now we concretize the class of stochastic processes which will be considered. Generally speaking, we will deal with so-called stationary and ergodic time series (or sources), whose definition will be given later, but now we consider may be the simplest class of such processes, which are called i.i.d. sources. By definition, they generate independent and identically distributed random variables from some set A. In our case A will be either some alphabet or a real-valued interval.

The next natural question is how to measure the errors of prediction and estimation of probability. Mainly we will measure these errors by the Kullback-Leibler (KL) divergence which is defined by

$$D(P, Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)}\ , \tag{1.4}$$

where $P(a)$ and $Q(a)$ are probability distributions over an alphabet A (here and below $\log \equiv \log_2$ and $0 \log 0 = 0$). The probability distribution $P(a)$ can be considered as unknown whereas $Q(a)$ is its estimation. It is well-known that for any distributions P and Q the KL divergence is nonnegative and equals 0 if and only if $P(a) = Q(a)$ for all a .[14] So, if the estimation Q is equal to P, the error is 0, otherwise the error is a positive number.

The KL divergence is connected with the so-called variation distance

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

via the the following inequality (Pinsker's inequality)

$$\sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} ||P - Q||^2. \tag{1.5}$$

Let γ be a predictor, i.e. an estimation of an unknown conditional probability and $x_1 \cdots x_t$ be a sequence of letters created by an unknown source P. The KL divergence between P and the predictor γ is equal to

$$\rho_{\gamma,P}(x_1 \cdots x_t) = \sum_{a \in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}, \tag{1.6}$$

For fixed t it is a random variable, because $x_1, x_2, \cdots, x_t$ are random variables. We define the average error at time t by

$$\rho^t(P\|\gamma) = E\ (\rho_{\gamma,P}(\cdot)) = \sum_{x_1\cdots x_t \in A^t} P(x_1 \cdots x_t)\ \rho_{\gamma,P}(x_1 \cdots x_t) \tag{1.7}$$

$$= \sum_{x_1\cdots x_t \in A^t} P(x_1 \cdots x_t)\ \sum_{a\in A} P(a|x_1 \cdots x_t) \log \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}.$$

Analogously, if $\gamma(\,)$ is an estimation of a probability distribution we define the errors *per letter* as follows:

$$\bar{\rho}_{\gamma,P}(x_1...x_t) = t^{-1}\ (\log(P(x_1...x_t)/\gamma(x_1...x_t)) \tag{1.8}$$

and

$$\bar{\rho}^t(P\|\gamma) = t^{-1} \sum_{x_1...x_t \in A^t} P(x_1...x_t) \log(P(x_1...x_t)/\gamma(x_1...x_t)), \tag{1.9}$$

where, as before, $\gamma(x_1...x_t) = \prod_{i=1}^t \gamma(x_i|x_1...x_{i-1})$. (Here and below we 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: $A^* = \bigcup_{i=1}^\infty A^i$.)

Claim 1.1 ([35]). *For any i.i.d. source P generating letters from an alphabet A and an integer t the average error (1.7) of the Laplace predictor and the average error of the Laplace estimator are upper bounded as follows:*

$$\rho^t(P\|L_0) \le ((|A|-1)\log e)/(t+1), \tag{1.10}$$

$$\bar{\rho}^t(P\|L_0) \le (|A|-1)\log t/t + O(1/t), \tag{1.11}$$

where $e \simeq 2.718$ is the Euler number.

So, we can see that the average error of the Laplace predictor goes to zero for any i.i.d. source P when the length t of the sample $x_1 \cdots x_t$ tends to infinity. Such methods are called universal, because the error goes to zero for any source, or process. In this case they are universal for the set of all i.i.d. sources generating letters from the finite alphabet A, but later we consider universal estimators for the set of stationary and ergodic sources. It is worth noting that the first universal code for which the estimation (1.11) is valid, was suggested independently by Fitingof[13] and Kolmogorov[21] in 1966.

The value

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

has one more interpretation connected with data compression. Now we consider the main idea whereas the more formal definitions will be given later. First we recall the definition of the Shannon entropy $h_0(P)$ for an i.i.d. source P

$$h_0(P) = -\sum_{a \in A} P(a) \log P(a). \tag{1.12}$$

It is easy to see that $t^{-1} \sum_{x_1 \ldots x_t \in A^t} P(x_1 ... x_t) \log(P(x_1 ... x_t)) = -h_0(P)$ for the i.i.d. source. Hence, we can represent the average error $\bar{\rho}^t(P \| \gamma)$ in (1.9) as

$$\bar{\rho}^t(P \| \gamma) = t^{-1} \sum_{x_1 \ldots x_t \in A^t} P(x_1 ... x_t) \log(1/\gamma(x_1 ... x_t)) - h_0(P).$$

More formal and general consideration of universal codes will be given later, but here we briefly show how estimations and codes are connected. The point is that one can construct a code with codelength $\gamma_{code}(a|x_1 \cdots x_t) \approx -\log_2 \gamma(a|x_1 \cdots x_n)$ for any letter $a \in A$ (since Shannon's original research, it has been well known that, using block codes with large block length or more modern methods of arithmetic coding[31] , the approximation may be as accurate as you like). If one knows the real distribution P, one can base coding on the true distribution P and not on the prediction γ. The difference in performance measured by average code length is given by

$$\sum_{a \in A} P(a|x_1 \cdots x_t)(-\log_2 \gamma(a|x_1 \cdots x_t)) - \sum_{a \in A} P(a|x_1 \cdots x_t)(-\log_2 P(a|x_1 \cdots x_t))$$

$$= \sum_{a \in A} P(a|x_1 \cdots x_t) \log_2 \frac{P(a|x_1 \cdots x_t)}{\gamma(a|x_1 \cdots x_t)}.$$

Thus this excess is exactly the error defined above (1.6) . Analogously, if we encode the sequence $x_1 \ldots x_t$ based on a predictor γ the redundancy per letter is defined by (1.8) and (1.9). So, from mathematical point of view, the estimation of the limiting probabilities and universal coding are identical. But $-\log \gamma(x_1 ... x_t)$ and $-\log P(x_1 ... x_t)$ have a very natural interpretation. The first value is a code word length (in bits), if the "code" γ is applied for compressing the word $x_1 ... x_t$ and the second one is the minimal possible codeword length. The difference is the redundancy of the code and, at the same time, the error of the predictor. It is worth noting that there are many other deep interrelations between the universal coding, prediction and estimation.[32,35]

We can see from the claim and the Pinsker inequality (1.5) that the variation distance of the Laplace predictor and estimator goes to zero, too. Moreover, it can be easily shown that the error (1.6) (and the corresponding variation distance) goes to zero with probability 1, when t goes to infinity. (Informally, it means that the error (1.6) goes to zero for almost all sequences $x_1 \cdots x_t$ according to the measure P.) Obviously, such properties are very desirable for any predictor and for larger classes of sources, like Markov and stationary ergodic (they will be briefly defined in the next subsection). However, it is proven[35] that such predictors do not exist for the class of all stationary and ergodic sources (generating letters from a given finite alphabet). More precisely, if, for example, the alphabet has two letters, then for any predictor γ and for any $\delta > 0$ there exists a source P such that with probability 1 $\rho_{\gamma,P}(x_1 \cdots x_t) \geq 1/2 - \delta$ infinitely often when $t \to \infty$. In other words, 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 (1.6) and (1.7) to compare different predictors. On the other hand, it is shown[35] that there exists a predictor R, such that the following Cesaro average $t^{-1} \sum_{i=1}^{t} \rho_{R,P}(x_1 \cdots x_i)$ goes to 0 (with probability 1) for any stationary and ergodic source P, where t goes to infinity. (This predictor will be described in the next subsection.) That is why we will focus our attention on such averages. From the definitions (1.6), (1.7) and properties of the logarithm we can see that for any probability distribution γ

$$t^{-1} \sum_{i=1}^{t} \rho_{\gamma,P}(x_1 \cdots x_i) = t^{-1} \left(\log(P(x_1...x_t)/\gamma(x_1...x_t)\right),$$

$$t^{-1} \sum_{i=1}^{t} \rho^i(P\|\gamma) = t^{-1} \sum_{x_1...x_t \in A^t} P(x_1...x_t) \log(P(x_1...x_t)/\gamma(x_1...x_t)).$$

Taking into account these equations, we can see from the definitions (1.8) and (1.9) that the Chesaro averages of the prediction errors (1.6) and (1.7) are equal to the errors of estimation of limiting probabilities (1.8) and (1.9). That is why we will use values (1.8) and (1.9) as the main measures of the precision throughout the chapter.

A natural problem is to find a predictor and an estimator of the limiting probabilities whose average error (1.9) is minimal for the set of i.i.d. sources. This problem was considered and solved by Krichevsky.[23,24] He suggested the following predictor:

$$K_0(a|x_1 \cdots x_t) = (\nu_{x_1 \cdots x_t}(a) + 1/2)/(t + |A|/2), \tag{1.13}$$

where, as before, $\nu_{x_1 \cdots x_t}(a)$ is the number of occurrencies of the letter a in the word $x_1 \ldots x_t$. We can see that the Krychevsky predictor is quite close to the Laplace's one (1.35).

Example 1.3. Let $A = \{0, 1\}$, $x_1...x_5 = 01010$. 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 represented as follows:

$$K_0(x_1...x_t) = \prod_{i=1}^{t} \frac{\nu_{x_1...x_{i-1}}(x_i) + 1/2}{i - 1 + |A|/2} = \frac{\prod_{a \in A}(\prod_{j=1}^{\nu_{x_1...x_t}(a)}(j - 1/2))}{\prod_{i=0}^{t-1}(i + |A|/2)}. \tag{1.14}$$

It is known that

$$(r+1/2)((r+1)+1/2)...(s-1/2) = \frac{\Gamma(s+1/2)}{\Gamma(r+1/2)}, \tag{1.15}$$

where $\Gamma(\,)$ is the gamma function.[22] So, (1.14) can be presented as follows:

$$K_0(x_1...x_t) = \frac{\prod_{a\in A}(\Gamma(\nu_{x_1...x_t}(a)+1/2)\,/\Gamma(1/2)\,)}{\Gamma(t+|A|/2)\,/\Gamma(|A|/2)}\,. \tag{1.16}$$

The following claim shows that the error of the Krichevsky estimator is a half of the Laplace's one.

Claim 1.2. *For any i.i.d. source P generating letters from a finite alphabet A the average error (1.9) of the estimator K_0 is upper bounded as follows:*

$$\bar\rho_t(K_0,P) \equiv\; t^{-1}\sum_{x_1...x_t\in A^t} P(x_1...x_t)\log(P(x_1...x_t)/K_0(x_1...x_t)) \equiv$$

$$t^{-1}\sum_{x_1...x_t\in A^t} P(x_1...x_t)\log(1/K_0(x_1...x_t)) - h_0(p) \le ((|A|-1)\log t + C)/(2t), \tag{1.17}$$

where C is a constant.

Moreover, in a certain sense this average error is minimal: it is shown by Krichevsky[23] that for any predictor γ there exists such a source P^* that

$$\bar\rho_t(\gamma,P^*) \ge ((|A|-1)\log t + C')/(2t).$$

Hence, the bound $((|A|-1)\log t + C)/(2t)$ cannot be reduced and the Krichevsky estimator is the best (up to $O(1/t)$) if the error is measured by the KL divergence ρ.

1.2.2. *Consistent Estimations and On-line Predictors for Markov and Stationary Ergodic Processes*

Now we briefly describe consistent estimations of unknown probabilities and efficient on-line predictors for general stochastic processes (or sources of information).

First we give a formal definition of stationary ergodic processes. 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 $T^{-1}B = B$.[5,14]

We denote by $M_\infty(A)$ the set of all stationary and ergodic sources and let $M_0(A) \subset M_\infty(A)$ be the set of all i.i.d. processes. We denote by $M_m(A) \subset M_\infty(A)$ the set of Markov sources of order (or with memory, or connectivity) not larger than m, $m \ge 0$. By definition $\mu \in M_m(A)$ if

$$\mu(x_{t+1}=a_{i_1}|x_t=a_{i_2}, x_{t-1}=a_{i_3},\,\dots, x_{t-m+1}=a_{i_{m+1}},\dots) \tag{1.18}$$

$$= \mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, \dots, x_{t-m+1} = a_{i_{m+1}})$$

for all $t \geq m$ and $a_{i_1}, a_{i_2}, \ldots \in A$. 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.[44] 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 $p(a|x_t \cdots x_1) = p(a|x_t \cdots x_{t-m+1})$, 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 an i.i.d. 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 to 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 i.i.d. method γ corresponding to the x_1x_2- subsequence ($= OO$), then x_4 is predicted based on the i.i.d. 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 (1.16), we can present the Krichevsky predictors for $M_m(A)$ as follows:

$$K_m(x_1 \ldots x_t) = \begin{cases} \frac{1}{|A|^t}, & if \;\; t \leq m\,, \\ \\ \frac{1}{|A|^m} \prod_{v \in A^m} \frac{\prod_{a\in A} ((\Gamma(\nu_x(va)+1/2) \,/\, \Gamma(1/2))}{(\Gamma(\bar\nu_x(v)+|A|/2) \,/\, \Gamma(|A|/2))}, & if \;\; t > m\,, \end{cases} \tag{1.19}$$

where $\bar\nu_x(v) = \sum_{a\in A} \nu_x(va)$, $x = x_1 \ldots x_t$. It is worth noting that the representation (1.14) can be more convenient for carrying out calculations if t is small.

Example 1.4. For the word $OOIOIIOOIIIOIO$ considered in the previous example, we obtain $K_2(OOIOIIOOIIIOIO) = 2^{-2}\ \frac{1}{2}\frac{3}{4}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}\ \frac{1}{2}\frac{1}{4}\frac{1}{2}$. Here groups of multipliers correspond to subsequences II, $OIIO$, IOI, OIO.

In order to estimate the error of the Krichevsky predictor K_m we need a general definition of the Shannon entropy. Let P be a stationary and ergodic source generating letters from a finite alphabet A. The $m-$ order (conditional) Shannon entropy and the limiting Shannon entropy are defined as follows:

$$h_m(P) = \sum_{v \in A^m} P(v) \sum_{a \in A} P(a/v) \log P(a/v), \qquad h_\infty(\tau) = \lim_{m \to \infty} h_m(P). \tag{1.20}$$

(If $m = 0$ we obtain the definition (1.12).) It is also known that for any m

$$h_\infty(P) \le h_m(P) \tag{1.21}$$

.[5,14]

Claim 1.3. *For any stationary and ergodic source P generating letters from a finite alphabet A the average error of the Krichevsky predictor K_m is upper bounded as follows:*

$$-t^{-1} \sum_{x_1 \dots x_t \in A^t} P(x_1 ... x_t) \log(K_m(x_1 ... x_t)) - h_m(P) \le (|A|^m(|A|-1)\log t + C)/(2t), \tag{1.22}$$

where C is a constant.

The following so-called empirical Shannon entropy, which is an estimation of the entropy (1.20), will play a key role in the hypothesis testing. It will be convenient to consider its definition here, because this notation will be used in the proof of the next claims. Let $v = v_1 ... v_k$ and $x = x_1 x_2 \dots x_t$ be words from A^*. Denote the rate of a word v occurring in the sequence $x = x_1 x_2 \dots x_k$, $x_2 x_3 \dots x_{k+1}$, $x_3 x_4 \dots x_{k+2}$, $\dots$, $x_{t-k+1} \dots x_t$ as $\nu_x(v)$. For example, if $x = 000100$ and $v = 00$, then $\nu_x(00) = 3$. For any $0 \le k < t$ the empirical Shannon entropy of order k is defined as follows:

$$h_k^*(x) = -\sum_{v \in A^k} \frac{\bar\nu_x(v)}{(t-k)} \sum_{a \in A} \frac{\nu_x(va)}{\bar\nu_x(v)} \log \frac{\nu_x(va)}{\bar\nu_x(v)}\,, \tag{1.23}$$

where $x = x_1 \dots x_t$, $\bar\nu_x(v) = \sum_{a \in A} \nu_x(va)$. In particular, if $k = 0$, we obtain $h_0^*(x) = -t^{-1} \sum_{a \in A} \nu_x(a) \log(\nu_x(a)/t)$.

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

$$\omega_1 = 1 - 1/\log 3,\ \dots,\ \omega_i = 1/\log(i+1) - 1/\log(i+2),\ \dots\,. \tag{1.24}$$

(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:

$$R(x_1 ... x_t) = \sum_{i=0}^{\infty} \omega_{i+1} \, K_i(x_1 ... x_t). \tag{1.25}$$

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.

Example 1.5. Let us calculate $R(00), ..., R(11)$. From (1.14) and (1.19) 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 (1.24) and the measure R (1.25), we can calculate $R(z_1 z_2)$ 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, $R(01) = R(10) \approx 0.204, R(11) \approx 0.296$.

The main properties of the measure R are connected with the Shannon entropy (1.20).

Theorem 1.1 ([35]). *For any stationary and ergodic source P the following equalities are valid:*

$$i) \lim_{t \to \infty} \frac{1}{t} \log(1/R(x_1 \cdots x_t)) \; = \; h_\infty(P)$$

with probability 1,

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

So, if one uses the measure R for data compression in such a way that the codeword length of the sequence $x_1 \cdots x_t$ is (approximately) equal to $\log(1/R(x_1 \cdots x_t))$ bits, he/she obtains the best achievable data compression ratio $h_\infty(P)$ per letter. On the other hand, we know that the redundancy of a universal code and the error of corresponding predictor are equal. Hence, if one uses the measure R for estimation and/or prediction, the error (per letter) will go to zero.

1.2.3. *Hypothesis Testing*

Here we briefly describe the main notions of hypothesis testing and the two particular problems considered below. A statistical test is formulated to test a specific null hypothesis (H_0). Associated with this null hypothesis is the alternative hypothesis (H_1).[33] For example, we will consider the two following problems: goodness-of-fit testing (or identity testing) and testing of serial independence. Both problems are well known in mathematical statistics and there is an extensive literature dealing with their nonparametric testing.[2,8,9,12]

The goodness-of-fit testing is described as follows: a hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{id}. One particular case, mentioned in Introduction, is when the source alphabet A is $\{0, 1\}$ and the main hypothesis H_0^{id} is that a bit sequence is generated by the Bernoulli i.i.d. source with equal probabilities of 0's and 1's. In all cases, the testing should be based on a sample $x_1 \dots x_t$ generated by the source.

The second problem is as follows: the null hypothesis H_0^{SI} is that the source is Markovian of order not larger than m, $(m \geq 0)$, and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. In particular, if $m = 0$, this is the problem of testing for independence of time series.

For each applied test, a decision is derived that accepts or rejects the null hypothesis. During the test, a test statistic value is computed on the data (the sequence being tested). This test statistic value is compared to the critical value. If the test statistic value exceeds the critical value, the null hypothesis is rejected. Otherwise, the null hypothesis is accepted. So, statistical hypothesis testing is a conclusion-generation procedure that has two possible outcomes: either accept H_0 or accept H_1.

Errors of the two following types are possible: The Type I error occurs if H_0 is true but the test accepts H_1 and, vice versa, the Type II error occurs if H_1 is true, but the test accepts H_0. The probability of Type I error is often called the level of significance of the test. This probability can be set prior to the testing and is denoted α. For a test, α is the probability that the test will say that H_0 is not true when it really is true. Common values of α are about 0.01. The probabilities of Type I and Type II errors are related to each other and to the size n of the tested sequence in such a way that if two of them are specified, the third value is automatically determined. Practitioners usually select a sample size n and a value for the probability of the Type I error - the level of significance.[33]

1.2.4. *Codes*

We briefly describe the main definitions and properties (without proofs) of lossless codes, or methods of (lossless) data compression. 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. In what follows we call uniquely decodable codes just "codes". It is well known that if φ is a code then the lengths of the codewords satisfy the following inequality (Kraft's inequality)[14] : $\Sigma_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1$. It will be convenient to reformulate this property as follows:

Claim 1.4. *Let φ be a code over an alphabet A. Then for any integer n there exists a measure μ_φ on A^n such that*

$$-\log \mu_\varphi(u) \ \leq \ |\varphi(u)| \tag{1.26}$$

for any u from A^n .

(Obviously, this claim is true for the measure $\mu_\varphi(u) = 2^{-|\varphi(u)|} / \Sigma_{u \in A^n} 2^{-|\varphi(u)|}$).

It was mentioned above that, in a certain sense, the opposite claim is true, too. Namely, for any probability measure μ defined on $A^n, n \geq 1$, there exists a code φ_μ such that

$$|\varphi_\mu(u)| = -\log \mu(u). \tag{1.27}$$

(More precisely, for any $\varepsilon > 0$ one can construct such a code φ^*_μ, that $|\varphi^*_\mu(u)| < -\log \mu(u) + \varepsilon$ for any $u \in A^n$. Such a code can be constructed by applying a so-called arithmetic coding[31] .) For example, for the above described measure R we can construct a code R_{code} such that

$$|R_{code}(u)| = -\log R(u). \tag{1.28}$$

As we mentioned above there exist universal codes. For their description we recall that sequences $x_1 \ldots x_t$, generated by a source P, can be "compressed" to the length $-\log P(x_1 ... x_t)$ bits (see (1.27)) and, on the other hand, for any source P there is no code ψ for which the average codeword length ($\Sigma_{u \in A^t} P(u) |\psi(u)|$) is less than $-\Sigma_{u \in A^t} P(u) \log P(u)$. Universal codes can reach the lower bound $-\log P(x_1 ... x_t)$ asymptotically for any stationary and ergodic source P in average and 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:

$$\lim_{t \to \infty} |U(x_1 \ldots x_t)|/t = h_\infty(P) \tag{1.29}$$

with probability 1, and

$$\lim_{t \to \infty} E(|U(x_1 \ldots x_t)|)/t = h_\infty(P), \tag{1.30}$$

where $E(f)$ is the expected value of f, $h_\infty(P)$ is the Shannon entropy of P, see (1.21). So, informally speaking, a universal code estimates the probability characteristics of a source and uses them for efficient "compression".

In this chapter we mainly consider finite-alphabet and real-valued sources, but sources with countable alphabet also were considered by many authors.[4,16,18,39,40] In particular, it is shown that, for infinite alphabet, without any condition on the source distribution it is impossible to have universal source code and/or universal predictor, i.e. such a predictor whose average error goes to zero, when the length of a sequence goes to infinity. On the other hand, there are some necessary and sufficient conditions for existence of universal codes and predictors.[4,18,39]

1.3. Finite Alphabet Processes

1.3.1. *The Estimation of (Limiting) Probabilities*

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

Theorem 1.2. *Let U be a universal code and*

$$\mu_U(u) = 2^{-|U(u)|}/\Sigma_{v\in A^{|u|}}\, 2^{-|U(v)|}. \tag{1.31}$$

Then, for any stationary and ergodic source P the following equalities are valid:

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

with probability 1,

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

The informal outline of the proof ia as follows: $\frac{1}{t}(-\log P(x_1\cdots x_t)$ and $\frac{1}{t}(-\log \mu_U(x_1\cdots x_t))$ goes to the Shannon entropy $h_\infty(P)$, that is why the difference is 0.

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

Nowadays there are many efficient universal codes (and universal predictors connected with them), which can be applied to estimation. For example, the above described measure R is based on a universal code[34,35] and can be applied for probability estimation. More precisely, Theorem 1.2 (and the following theorems) are true for R, if we replace μ_U by R.

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

Theorem 1.3. *([34,35]) 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:

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

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

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

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:

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

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

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

1.3.2. *Prediction*

As we mentioned above, any universal code U can be applied for prediction. Namely, the measure μ_U (1.31) can be used for prediction as the following conditional probability:

$$\mu_U(x_{t+1}|x_1...x_t) = \mu_U(x_1...x_t x_{t+1})/\mu_U(x_1...x_t). \tag{1.32}$$

The following theorem shows that such a predictor is quite reasonable. Moreover, it gives a possibility to apply practically used data compressors for prediction of real data (like EUR/USD rate) and obtain quite precise estimation[41] .

Theorem 1.4. *Let U be a universal code and P be any stationary and ergodic process. Then*

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

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

and

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

An informal outline of the proof is as follows:

$$\frac{1}{t}\ \{E(\log \frac{P(x_1)}{\mu_U(x_1)}) + E(\log \frac{P(x_2|x_1)}{\mu_U(x_2|x_1)}) + \ldots + E(\log \frac{P(x_t|x_1...x_{t-1})}{\mu_U(x_t|x_1...x_{t-1})})\}$$

is equal to $\frac{1}{t}E(\log \frac{P(x_1...x_t)}{\mu_U(x_1...x_t)})$. Taking into account Theorem 1.2, we obtain the first statement of the theorem.

Comment. 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 $R(x_{t+1}|x_1...x_t) = R(x_1...x_t x_{t+1})/R(x_1...x_t)$.[36]

Comment. It is known[45] that, in fact, the statements ii) and iii) are equivalent.

1.3.3. *Problems with Side Information*

Now we consider the 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 $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and 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 $P(x_t|\ (x_1, y_1), \ldots, (x_{t-1}, y_{t-1}), y_t)$ 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 previous subsection. More precisely, for any universal code U and the corresponding measure μ_U (1.31) we define the following estimate for the problem with side information:

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

The following theorem shows that this estimate is quite reasonable.

Theorem 1.5. *Let U be a universal code and let P be any stationary and ergodic process. Then*

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

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

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

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

and

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

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

The proof is very close to the proof of the previous theorem.

1.3.4. *The Case of Several Independent Samples*

In this part we consider a situation which is important for practical applications, but needs cumbersome notations. Namely, we extend our consideration to the case where the sample is presented as several independent samples $x^1 = x^1_1 \dots x^1_{t_1}$, $x^2 = x^2_1 \dots x^2_{t_2}, ..., x^r = x^r_1 \dots x^r_{t_r}$ generated by a source. More precisely, we will suppose that all sequences were independently created by one stationary and ergodic source. (The point is that it is impossible just to combine all samples into one, if the source is not i.i.d.) We denote them by $x^1 \diamond x^2 \diamond \dots \diamond x^r$ and define $\nu_{x^1\diamond x^2\diamond\dots\diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. 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:

$$K_m(x^1\diamond x^2\diamond\dots\diamond x^r) = \tag{1.33}$$

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

whereas the definition of R is the same (see (1.25)). (Here, as before, $\bar\nu_{x^1\diamond x^2\diamond\dots\diamond x^r}(v) = \sum_{a\in A}\nu_{x^1\diamond x^2\diamond\dots\diamond x^r}(va)$. Note, that $\bar\nu_{x^1\diamond x^2\diamond\dots\diamond x^r}(\,) = \sum_{i=1}^r t_i$ if $m=0$.)

The following example is intended to show the difference between the case of many samples and one.

Example 1.6. 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 $P(z_1z_2), z_1, z_2 \in \{0,1\}$ (here $z_1z_2\dots$ can be considered as an independent sequence, generated by the source) and predict x_4x_5 (i.e. estimate conditional probability $P(x_4x_5|x_1\dots x_3 = 101, y_1\dots y_4 = 0101)$. For solving both problems we will use the measure R (see (1.25)). First we consider the case where $P(z_1z_2)$ is to be estimated without knowledge of sequences x and y. Those probabilities were calculated in the example 1.5 and we obtained: $R(00) \approx 0.296$, $R(01) = R(10) \approx 0.204$, $R(11) \approx 0.296$. Let us now estimate the probability $P(z_1z_2)$ taking into account that there are two independent samples $y = y_1 \dots y_4 = 0101$ and $x = x_1 \dots x_3 = 101$. First of all we note that such estimates are based on the formula for conditional probabilities:

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

Then 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 $P(z_1 z_2 = 01)$. Carrying out similar calculations, we obtain $R(0101 \diamond 101 \diamond 01) \approx 0.00292$, $R(z_1 z_2 = 01 | y_1 \ldots y_4 = 0101, x_1 \ldots x_3 = 101) = R(0101 \diamond 101 \diamond 01)/R(\,0101 \diamond 101) \approx 0.32812$. 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 to 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 describe properties of the universal code R, but 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 1.5. *Let $x^1, x^2, ..., x^r$ be independent sequences generated by a stationary and ergodic source and let t be a total length of these sequences ($t = \sum_{i=1}^r |x^i|$). Then, if $t \to \infty$, (and r is fixed) the statements of the Theorems 1.2 - 1.5 are valid, when applied to $x^1 \diamond x^2 \diamond ... \diamond x^r$ instead of $x_1 \ldots x_t$. (In theorems 1.2 - 1.5 μ_U should be changed to R.)*

The proofs are completely analogous to the proofs of the Theorems 1.2—1.5.

Now we can extend the definition of the empirical Shannon entropy (1.23) to the case of several words $x^1 = x^1_1 \ldots x^1_{t_1}$, $x^2 = x^2_1 \ldots x^2_{t_2}, ...,$ $x^r = x^r_1 \ldots x^r_{t_r}$. We 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 $\nu_{x^1 \diamond x^2}(00) = 1$. Analogously to (1.23),

$$h^*_k(x^1 \diamond x^2 \diamond ... \diamond x^r) = -\sum_{v \in A^k} \frac{\bar{\nu}_{x^1 \diamond \ldots \diamond x^r}(v)}{(t - kr)} \sum_{a \in A} \frac{\nu_{x^1 \diamond \ldots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \ldots \diamond x^r}(v)} \log \frac{\nu_{x^1 \diamond \ldots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \ldots \diamond x^r}(v)}\,, \tag{1.34}$$

where $\bar{\nu}_{x^1\diamond\ldots\diamond x^r}(v) = \sum_{a\in A}\nu_{x^1\diamond\ldots\diamond x^r}(va)$.

For any sequence of words $x^1 = x^1_1\ldots x^1_{t_1}$, $x^2 = x^2_1\ldots x^2_{t_2}$, ..., $x^r = x^r_1\ldots x^r_{t_r}$ from A^* and any measure θ we define $\theta(x^1\diamond x^2\diamond\ldots\diamond x^r) = \prod_{i=1}^r \theta(x^i)$. The following lemma gives an upper bound for unknown probabilities.

Lemma 1.1. *Let θ be a measure from $M_m(A), m\geq 0$, and $x^1,\ldots,x^r$ be words from A^*, whose lengths are not less than m. Then*

$$\theta(x^1\diamond\ldots\diamond x^r) \leq 2^{-(t-rm)\,h^*_m(x^1\diamond\ldots\diamond x^t)}, \tag{1.35}$$

where $\theta(x^1\diamond\ldots\diamond x^r) = \prod_{i=1}^r\theta(x^i)$.

1.4. Hypothesis Testing

1.4.1. *Goodness-of-Fit or Identity Testing*

Now we consider the problem of testing H_0^{id} against H_1^{id}. Let us recall that the hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{id}. Let the required level of significance (or the Type I error) be α, $\alpha\in(0,1)$. We describe a statistical test which can be constructed based on any code φ.

The main idea of the suggested test is quite natural: compress a sample sequence $x_1...x_t$ by a code φ. If the length of the codeword ($|\varphi(x_1...x_t)|$) is significantly less than the value $-\log\pi(x_1...x_t)$, then H_0^{id} should be rejected. The key observation is that the probability of all rejected sequences is quite small for any φ, that is why the Type I error can be made small. The precise description of the test is as follows: *The hypothesis H_0^{id} is accepted if*

$$-\log\pi(x_1...x_t) - |\varphi(x_1...x_t)| \leq -\log\alpha. \tag{1.36}$$

Otherwise, H_0^{id} is rejected. We denote this test by $T_\varphi^{\,id}(A,\alpha)$.

Theorem 1.6. *i) For each distribution $\pi,\alpha\in(0,1)$ and a code φ, the Type I error of the described test $T_\varphi^{\,id}(A,\alpha)$ is not larger than α and ii) if, in addition, π is a finite-order stationary and ergodic process over A^∞ (i.e. $\pi\in M^*(A)$) and φ is a universal code, then the Type II error of the test $T_\varphi^{\,id}(A,\alpha)$ goes to 0, when t tends to infinity.*

1.4.2. *Testing for Serial Independence*

Let us recall that the null hypothesis H_0^{SI} is that the source is Markovian of order not larger than m, ($m\geq 0$), and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. In particular, if $m=0$, this is the problem of testing for independence of time series.

Let there be given a sample $x_1...x_t$ generated by an (unknown) source π. The main hypothesis H_0^{SI} is that the source π is Markovian whose order is not greater than m, $(m \geq 0)$, and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. The described test is as follows.

Let φ be any code. By definition, the hypothesis H_0^{SI} is accepted if

$$(t-m)\, h_m^*(x_1...x_t) - |\varphi(x_1...x_t)| \leq \log(1/\alpha)\,, \tag{1.37}$$

where $\alpha \in (0,1)$. Otherwise, H_0^{SI} is rejected. We denote this test by $T_\varphi^{SI}(A,\alpha)$.

Theorem 1.7. *i) For any code φ the Type I error of the test $T_\varphi^{SI}(A,\alpha)$ is less than or equal to $\alpha, \alpha \in (0,1)$ and, ii) if, in addition, φ is a universal code, then the Type II error of the test $T_\varphi^{SI}(A,\alpha)$ goes to 0, when t tends to infinity.*

1.5. Real-Valued Time Series

1.5.1. *Density Estimation and Its Application*

Here we address the problem of nonparametric estimation of the density for time series. Let X_t be a time series and the probability distribution of X_t is unknown, but it is known that the time series is stationary and ergodic. We have seen that Shannon-MacMillan-Breiman theorem played a key role in the case of finite-alphabet processes. In this part we will use its generalization to the processes with densities, which was established by Barron.[3] First we describe considered processes with some properties needed for the generalized Shannon-MacMillan-Breiman theorem to hold. In what follows, we restrict our attention to processes that take bounded real valued. However, the main results may be extended to processes taking values in a compact subset of a separable metric space.

Let B denote the Borel subsets of R, and B^k denote the Borel subsets of R^k, where R is the set of real numbers. Let R^∞ be the set of all infinite sequences $x = x_1, x_2 \dots$ with $x_i \in \mathsf{R}$, and let B^∞ denote the usual product sigma field on R^∞, generated by the finite dimensional cylinder sets $\{A_1, \dots A_k, \mathsf{R}, \mathsf{R}, \dots\}$, where $A_i \in B, i = 1, \dots, k$. Each stochastic process $X_1, X_2, \dots, X_i \in \mathsf{R}$, is defined by a probability distribution on $(\mathsf{R}^\infty, B^\infty)$. Suppose that the joint distribution P_n for $(X_1, X_2, \dots, X_n)$ has a probability density function $p(x_1 x_2 \dots x_n)$ with respect to a sigma-finite measure M_n. Assume that the sequence of dominating measures M_n is Markov of order $m \geq 0$ with a stationary transition measure. A familiar case for M_n is Lebesgue measure. 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 $\tilde{h}$ defined by

$$\tilde{h} = \lim_{n \to \infty} -E(\log p(x_{n+1}|x_1 \dots x_n)), \tag{1.38}$$

where E denotes expectation with respect to P. We will use the following generalization of the Shannon-MacMillan-Breiman theorem:

Claim 1.6 ([3]). *If $\{X_n\}$ is a $P-$stationary ergodic process with density $p(x_1 \dots x_n) = dP_n/dM_n$ and $\tilde{h}_n < \infty$ for some $n \geq m$, the sequence of relative entropy densities $-(1/n) \log p(x_1 \dots x_n)$ convergence almost surely to the relative entropy rate, i.e.,*

$$\lim_{n\to\infty} (-1/n) \log p(x_1 \dots x_n) = \tilde{h} \tag{1.39}$$

with probability 1 (according to P).

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

$$p^s(x_1 \dots x_n) = P(x_1^{[s]} \dots x_n^{[s]})/M_n(x_1^{[s]} \dots x_n^{[s]}). \tag{1.40}$$

We also consider

$$\tilde{h}_s = \lim_{n\to\infty} -E(\log p^s(x_{n+1}|x_1 \dots x_n)). \tag{1.41}$$

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

$$\lim_{t\to\infty} -\frac{1}{t} \log p^s(x_1 \dots x_t) = \tilde{h}_s. \tag{1.42}$$

Let U be a universal code, which is defined for any finite alphabet. In order to describe a density estimate we will use the probability distribution $\omega_i, i = 1, 2, \dots$, see (1.24) (In what follows we will use this distribution, but results described below are obviously true for any distribution with nonzero probabilities.) Now we can define the density estimate r_U as follows:

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i \, \mu_U(x_1^{[i]} \dots x_t^{[i]})/M_t(x_1^{[i]} \dots x_t^{[i]}) \,, \tag{1.43}$$

where the measure μ_U is defined by (1.31). (It is assumed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters.)

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

$$\lim_{s\to\infty} \tilde{h}_s = \tilde{h} < \infty, \tag{1.44}$$

where $\tilde{h}$ and $\tilde{h}_s$ are relative entropy rates, see (1.38), (1.41). Then

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 ... x_t)}{r_U(x_1 ... x_t)} = 0 \tag{1.45}$$

with probability 1 and

$$\lim_{t\to\infty} \frac{1}{t} E(\log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)}) \;=\; 0\,. \tag{1.46}$$

We have seen that the requirement (1.44) plays an important role in the proof. The natural question is whether there exist processes for which (1.44) is valid. The answer is positive. For example, let a process possess values in the interval $[-1,1]$, M_n be Lebesgue measure and the 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

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

In words, the density depends on a sign of the previous value. If the value is positive, then the density is more than 1/2, otherwise it is less than 1/2. It is easy to see that (1.44) is true for any $\alpha \in (0, 1)$.

The following two theorems are devoted to the conditional probability $r_U(x|x_1...x_m) = r_U(x_1...x_m x)/r_U(x_1...x_m)$ 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 the unknown density $p(x|x_1...x_m)$.

Theorem 1.9. *Let $B_1, B_2, \dots$ be a sequence of measurable sets. Then the following equalities are true:*

$$i) \; \lim_{t\to\infty} \; E(\frac{1}{t} \sum_{m=0}^{t-1} (P(x_{m+1} \in B_{m+1}|x_1...x_m) - R_U(x_{m+1} \in B_{m+1}|x_1...x_m))^2) = 0\,, \tag{1.47}$$

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

where $R_U(x_{m+1} \in B_{m+1}|x_1...x_m) = \int_{B_{m+1}} r_U(x|x_1...x_m) dM_{1/m}$

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

Theorem 1.10. Let f be an integrable function, whose absolute value is bounded by a certain constant $\bar{M}$ and all conditions of the theorem 2 are true. Then the following equality is valid:

$$i)\ \lim_{t\to\infty}\frac{1}{t}E(\sum_{m=0}^{t-1}(\int f(x)\,p(x|x_1...x_m)dM_m-\int f(x)\,r_U(x|x_1...x_m)dM_m)^2)=0, \tag{1.48}$$

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

It is worth noting that this approach was used for prediction of real processes.[41]

1.5.2. *Hypothesis Testing*

In this subsection we consider a case where the source alphabet A is infinite, say, a part of R^n. Our strategy is to use finite partitions of A and to consider hypotheses corresponding to the partitions. This approach can be directly applied to the goodness-of-fit testing, but it cannot be applied to the serial independence testing. The point is that if someone combines letters (or states) of a Markov chain, the chain order (or memory) can increase. For example, if the alphabet contains three letters, there exists a Markov chain of order one, such that combining two letters into one transforms the chain into a process with infinite memory. That is why in this part we will consider the independence testing for i.i.d. processes only (i.e. processes from $M_0(A)$).

In order to avoid repetitions, we will consider a general scheme, which can be applied to both tests using notations $H_0^\aleph, H_1^\aleph$ and $T_\varphi^\aleph(A,\alpha)$, where $\aleph$ is an abbreviation of one of the described tests (i.e. *id* and *SI*.)

Let us give some definitions. Let $\Lambda=\lambda_1,...,\lambda_s$ be a finite (measurable) partition of A and let $\Lambda(x)$ be an element of the partition Λ which contains $x\in A$. For any process π we define a process π_Λ over a new alphabet Λ by the equation

$$\pi_\Lambda(\lambda_{i_1}...\lambda_{i_k})=\pi(x_1\in\lambda_{i_1},...,x_k\in\lambda_{i_k}),$$

where $x_1...x_k\in A^k$.

We will consider an infinite sequence of partitions $\hat{\Lambda}=\Lambda_1,\Lambda_2,....$ and say that such a sequence discriminates between a pair of hypotheses $H_0^\aleph(A), H_1^\aleph(A)$ about processes, if for each process ϱ, for which $H_1^\aleph(A)$ is true, there exists a partition Λ_j for which $H_1^\aleph(\Lambda_j)$ is true for the process ϱ_{Λ_j}.

Let $H_0^\aleph(A), H_1(A)^\aleph$ be a pair of hypotheses, $\hat{\Lambda}=\Lambda_1,\Lambda_2,...$ be a sequence of partitions, α be from $(0,1)$ and φ be a code. The scheme for both tests is as follows:

The hypothesis $H_0^\aleph(A)$ is accepted if for all $i=1,2,3,...$ the test $T_\varphi^\aleph(\Lambda_i,(\alpha\omega_i))$ accepts the hypothesis $H_0^\aleph(\Lambda_i)$. Otherwise, $H_0^\aleph$ is rejected. We denote this test $\mathbf{T}^\aleph_{\alpha,\varphi}(\hat{\Lambda})$.

Comment. It is important to note that one does not need to check an infinite number of inequalities when applying this test. The point is that the hypothesis

$H_0^{\aleph}(A)$ has to be accepted if the left part in (1.36) or (1.37) is less than $-\log(\alpha\omega_i)$. Obviously, $-\log(\alpha\omega_i)$ goes to infinity if i increases. That is why there are many cases, where it is enough to check a finite number of hypotheses $H_0^{\aleph}(\Lambda_i)$.

Theorem 1.11. *i) For each $\alpha \in (0,1)$, sequence of partitions $\hat{\Lambda}$ and a code φ, the Type I error of the described test $\boldsymbol{T}^{\aleph}_{\alpha,\varphi}(\hat{\Lambda})$ is not larger than α, and ii) if, in addition, φ is a universal code and $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A), H_1(A)^{\aleph}$, then the Type II error of the test $\boldsymbol{T}^{\aleph}_{\alpha,\varphi}(\hat{\Lambda})$ goes to 0, when the sample size tends to infinity.*

1.6. Conclusion

Time series is a popular model of real stochastic processes which has a lot of applications in industry, economy, meteorology and many other fields. Despite this, there are many practically important problems of statistical analysis of time series which are still open. Among them we can name the problem of estimation of the limiting probabilities and densities, on-line prediction, regression, classification and some problems of hypothesis testing (goodness-of-fit testing and testing of serial independence). This chapter describes a new approach to all the problems mentioned above, which, on the one hand, gives a possibility to solve the problems in the framework of the classical mathematical statistics and, on the other hand, allows to apply methods of real data compression to solve these problems in practise. Such applications to randomness testing[42] and prediction of currency exchange rates[41] showed high efficiency, that is why the suggested methods look very promising for practical applications. Of course, problems like prediction of price of oil, gold, etc. and testing of different random number generators can be used as case studies for students.

1.7. Problems for Chapter

Problem 1.1. *Suppose* 010101 *is a sequence generated by a source whose alphabet is $\{0,1\}$. Calculate the probabilities $L_0(01010)$ and $K_0(01010)$. Predict the next symbol by the predictors L_0 and K_0 (i.e. calculate the conditional probabilities $L_0(0|01010), L_0(1|01010)$ and $K_0(0|01010), K_0(1|01010)$).*

Problem 1.2. *Suppose the sequence* 010101 *is generated by the first order Markov chain and the alphabet is $\{0,1\}$ (i.e. the source belongs to $M_1(\{0,1\})$).*

i) Represent this sequence by two ones generated by i.i.d. sources.

ii) Repeat all calculations from Problem 1.1 for L_1 and K_1. Compare results of Problems 1.1 and 1.2 . Explain the difference.

Problem 1.3. *For the sequence* 001100110011 *calculate the following empirical Shannon entropies: h_0^*, h_1^* and h_2^*.*

Problem 1.4. *Repeat all calculations from Problem 1.1 for the measure R. Compare obtained results with solutions of Problems 1.1 and 1.2 .*

Problem 1.5. *Let $\varphi(a) = 000, \varphi(b) = 01, \varphi(c) = 001, \varphi(d) = 1$ be a code over the alphabet $\{a, b, c, d\}$. Calculate the corresponding measure μ_φ.*

Problem 1.6. (problems with side information) *Let alphabets X and Y be as follows: $A = \{0, 1\}$, $Y = \{a, b, c\}$, correspondingly. There is a sequence $(x_1, y_1), \ldots, (x_4, y_4) = (0, a), (1, c), (1, b), (0, a)$ and it is known that $y_5 = a$. Predict x_5 based on the measure R, i.e. estimate the following conditional probabilities:*

$$R(x_5 = 0|(x_1, y_1), \ldots, (x_4, y_4) = (0, a), (1, c), (1, b), (0, a), y_5 = a),$$

$$R(x_5 = 1|(x_1, y_1), \ldots, (x_4, y_4) = (0, a), (1, c), (1, b), (0, a), y_5 = a).$$

Problem 1.7. (several independent samples) *Let there be two independent samples $\bar{x}^1 = x^1_1 ... x^1_5 = 10101$ and $\bar{x}^1 = x^2_1 ... x^2_6 = 010101$, generated by a stationary and ergodic source with the alphabet 0, 1. Based on the measure R estimate the (limiting) probability $P(x_1 x_2 x_3 = 010|\bar{x}^1 \diamond \bar{x}^2 = 10101 \diamond 010101)$ and predict x^2_7 (i.e. calculate conditional probability $R(x^2_7 = 0|\bar{x}^1 \diamond \bar{x}^2 = 10101 \diamond 010101)$ and $R(x^2_7 = 1|\bar{x}^1 \diamond \bar{x}^2 = 10101 \diamond 010101)$.*

Problem 1.8. (several independent samples) *For the sequences* 010101 *and* 010 *calculate the following empirical Shannon entropies: $h^*_0(010101 \diamond 010)$, $h^*_1(010101 \diamond 010)$ and $h^*_2(010101 \diamond 010)$.*

Problem 1.9. (hypothesis testing) *Let H^{id}_0 be a hypothesis that a source π is i.i.d. and generates letters from the alphabet $\{0, 1\}$ with equal probabilities, i.e. $\pi(0) = \pi(1) = 0.5$. The hypothesis H^{id}_1 is that the sequence is generated by a stationary and ergodic source which differs from the source under H^{id}_0 and the level of significance (α) is* 0.01. *There is a sample sequence* 0101010101. *The problem is to test H^{id}_0 based on the two following codes:*

i) the Laplace code $L_{0\,code}$ whose codeword length is given by $L_{0\,code}(u) = -\log L_0(u)$ and

ii) the R_{code} whose codeword length is given by $R_{code}(u) = -\log R(u)$

Problem 1.10. (hypothesis testing) *Let H^{SI}_0 be a hypothesis that is that the source is Markovian of order not larger than 1, and the alternative hypothesis H^{SI}_1 is that the source is stationary and ergodic which differs from the source under H^{SI}_0. There is a sequence* 001001001001 *and let the level of significance be* 0.01. *The problem is to test H^{SI}_0 against H^{SI}_1 based on the code R_{code}.*

Problem 1.11. (hypothesis testing) *Use the same sequence and the same α as in the previous problem for testing H^{SI}_0 that the source is Markovian of order not larger than 2, and the alternative hypothesis H^{SI}_1 is that the source is stationary and ergodic which differs from the source under H^{SI}_0. Compare results of two last problems and explain the difference.*

1.8. Appendix

Proof. [Claim 1.1] We employ the general inequality

$$D(\mu\|\eta) \leq \log e\ (-1 + \sum_{a\in A} \mu(a)^2/\eta(a)\ ),$$

valid for any distributions μ and η over A (follows from the elementary inequality for natural logarithm $\ln x \leq x - 1$), and find:

$$\rho^t(P\|L_0) = \sum_{x_1\cdots x_t\in A^t} P(x_1\cdots x_t) \sum_{a\in A} P(a|x_1\cdots x_t) \log \frac{P(a|x_1\cdots x_t)}{\gamma(a|x_1\cdots x_t)}$$

$$= \log e\,(\sum_{x_1\cdots x_t\in A^t} P(x_1\cdots x_t) \sum_{a\in A} P(a|x_1\cdots x_t) \ln \frac{P(a|x_1\cdots x_t)}{\gamma(a|x_1\cdots x_t)})$$

$$\leq \log e\,(-1 + \sum_{x_1\cdots x_t\in A^t} P(x_1\cdots x_t) \sum_{a\in A} \frac{P(a)^2(t+|A|)}{\nu_{x_1\cdots x_t}(a)+1}$$

Applying the well-known Bernoulli formula, we obtain

$$\rho^t(P\|L_0) = \log e\,(-1 + \sum_{a\in A}\sum_{i=0}^{t} \frac{P(a)^2(t+|A|)}{i+1} \binom{t}{i} P(a)^i(1-P(a))^{t-i})$$

$$= \log e\,(-1 + \frac{t+|A|}{t+1}\sum_{a\in A} P(a) \sum_{i=0}^{t} \binom{t+1}{i+1} P(a)^{i+1}(1-P(a))^{t-i})$$

$$\leq \log e\,(-1 + \frac{t+|A|}{t+1}\sum_{a\in A} P(a) \sum_{j=0}^{t+1} \binom{t+1}{j} P(a)^{j}(1-P(a))^{t+1-j}).$$

Again, using the Bernoulli formula, we finish the proof

$$\rho^t(P\|L_0) = \log e\,\frac{|A|-1}{t+1}.$$

The second statement of the claim follows from the well-known asymptotic equality

$$1 + 1/2 + 1/3 + ... + 1/t = \ln t + O(1),$$

the obvious presentation

$$\bar\rho^t(P\|L_0) = t^{-1}(\rho^0(P\|L_0) + \rho^1(P\|L_0) + ... + \rho^{t-1}(P\|L_0))$$

and (1.10). □

Proof. [Claim 1.2] The first equality follows from the definition (1.9), whereas the second from the definition (1.12). From (1.16) we obtain:

$$-\log K_0(x_1...x_t) = -\log(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \frac{\prod_{a\in A}\Gamma(\nu^t(a)+1/2)}{\Gamma((t+|A|/2)})$$

$$= c_1 + c_2|A| + \log\Gamma(t+|A|/2) - \sum_{a\in A}\Gamma(\nu^t(a)+1/2),$$

where c_1, c_2 are constants. Now we use the well known Stirling formula

$$\ln\Gamma(s) = \ln\sqrt{2\pi} + (s-1/2)\ln s - s + \theta/12,$$

where $\theta \in (0,1)^{22}$. Using this formula we rewrite the previous equality as

$$-\log K_0(x_1...x_t) = -\sum_{a\in A}\nu^t(a)\log(\nu^t(a)/t) + (|A|-1)\log t/2 + \bar{c}_1 + \bar{c}_2|A|,$$

where $\bar{c}_1, \bar{c}_2$ are constants. Hence,

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\log(K_0(x_1...x_t)))$$

$$\leq t(\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\sum_{a\in A}\nu^t(a)\log(\nu^t(a)/t)) + (|A|-1)\log t/2 + c|A|.$$

Applying the well known Jensen inequality for the concave function $-x\log x$ we obtain the following inequality:

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\log(K_0(x_1...x_t)) \leq$$

$$-t(\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)((\nu^t(a)/t))$$

$$\log\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(\nu^t(a)/t) + (|A|-1)\log t/2 + c|A|.$$

The source P is i.i.d., that is why the average frequency $\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)\nu^t(a)$ is equal to $P(a)$ for any $a \in A$ and we obtain from two last formulas the following inequality:

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\log(K_0(x_1...x_t))$$

$$\leq t(-\sum_{a\in A}P(a)\log P(a)) + (|A|-1)\log t/2 + c|A| \qquad (1.49)$$

On the other hand,

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(\log P(x_1\ldots x_t)) = \sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)\sum_{i=1}^{t}\log P(x_i)$$

$$= t(\sum_{a\in A} P(a)\log P(a)). \tag{1.50}$$

From (1.49) and (1.50) we can see that

$$t^{-1} \sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)\log\frac{P(x_1\ldots x_t)}{(K_0(x_1...x_t)} \le ((|A|-1)\log t/2 + c)/t.$$

$\square$

Proof. [Claim 1.3] First we consider the case where $m = 0$. The proof for this case is very close to the proof of the previous claim. Namely, from (1.16) we obtain:

$$-\log K_0(x_1...x_t) = -\log(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}}\,\frac{\prod_{a\in A}\Gamma(\nu^t(a)+1/2)}{\Gamma((t+|A|/2)})$$

$$= c_1 + c_2|A| + \log\Gamma(t+|A|/2) - \sum_{a\in A}\Gamma(\nu^t(a)+1/2),$$

where c_1, c_2 are constants. Now we use the well known Stirling formula

$$\ln\Gamma(s) = \ln\sqrt{2\pi} + (s-1/2)\ln s - s + \theta/12,$$

where $\theta \in (0,1)^{22}$. Using this formula we rewrite the previous equality as

$$-\log K_0(x_1...x_t) = -\sum_{a\in A}\nu^t(a)\log(\nu^t(a)/t) + (|A|-1)\log t/2 + \bar c_1 + \bar c_2|A|,$$

where $\bar c_1, \bar c_2$ are constants. Having taken into account the definition of the empirical entropy (1.23), we obtain

$$-\log K_0(x_1...x_t) \le t h_0^*(x_1\ldots x_t) + (|A|-1)\log t/2 + c|A|.$$

Hence,

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\log(K_0(x_1...x_t)))$$

$$\le t(\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t) h_0^*(x_1\ldots x_t) + (|A|-1)\log t/2 + c|A|.$$

Having taken into account the definition of the empirical entropy (1.23), we apply the well known Jensen inequality for the concave function $-x\log x$ and obtain the following inequality:

$$\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(-\log(K_0(x_1...x_t)) \le +c|A| -$$

$$t(\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)((\nu^t(a)/t))\log\sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)(\nu^t(a)/t) + (|A|-1)\log t/2.$$

P is stationary and ergodic, that is why the average frequency $\sum_{x_1 \ldots x_t \in A^t} P(x_1 \ldots x_t) \nu^t(a)$ is equal to $P(a)$ for any $a \in A$ and we obtain from two last formulas the following inequality:

$$\sum_{x_1 \ldots x_t \in A^t} P(x_1 \ldots x_t)(-\log(K_0(x_1 ... x_t)) \leq t\, h_0(P) + (|A|-1)\log t/2 + c|A|,$$

where $h_0(P)$ is the first order Shannon entropy, see (1.12).

We have seen that any source from $M_m(A)$ can be presented as a "sum" of $|A|^m$ i.i.d. sources. From this we can easily see that the error of a predictor for the source from $M_m(A)$ can be upper bounded by the error of i.i.d. source multiplied by $|A|^m$. In particular, we obtain from the last inequality and the definition of the Shannon entropy (1.20) the upper bound (1.22). □

Proof. [Theorem 1.1] We can see from the definition (1.25) of R and the Claim 1.19 that the average error is upper bounded as follows:

$$-\,t^{-1} \sum_{x_1 \ldots x_t \in A^t} P(x_1 ... x_t) \log(R(x_1 ... x_t)) - h_k(P)$$

$$\leq (|A|^k(|A|-1)\log t + \log(1/\omega_i) + C)/(2t),$$

for any $k = 0, 1, 2, ...$ Taking into account that for any $P \in M_\infty(A)$ $\lim_{k\to\infty} h_k(P) = h_\infty(P)$, we can see that

$$(\lim_{t\to\infty} t^{-1} \sum_{x_1 \ldots x_t \in A^t} P(x_1 ... x_t) \log(R(x_1 ... x_t)) - h_\infty(P)) = 0.$$

The second statement of the theorem is proven. The first one can be easily derived from the ergodicity of P[5,14] . □

Proof. [Theorem 1.2] The proof is based on the Shannon-MacMillan-Breiman theorem which states that for any stationary and ergodic source P

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

with probability 1[5,14] . From this equality and (1.29) we obtain the statement i). The second statement follows from the definition of the Shannon entropy (1.21) and (1.30). □

Proof. [Theorem 1.4] i) immediately follows from the second statement of the theorem 1.2 and properties of log. The statement ii) can be proven as follows:

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

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

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

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

Here the first inequality is obvious, the second follows from the Pinsker's inequality (1.5), the others from properties of expectation and log . iii) can be derived from ii) and the Jensen inequality for the function x^2. $\square$

Proof. [Theorem 1.5] The following inequality follows from the nonnegativity of the KL divergency (see (1.5)), whereas the 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 the 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. $\square$

Proof. [Lemma 1.1] . First we show that for any source $\theta^* \in M_0(A)$ and any words $x^1 = x^1_1 \ldots x^1_{t_1}$, ..., $x^r = x^r_1 \ldots x^r_{t_r}$,

$$\theta^*(x^1 \diamond \ldots \diamond x^r) = \prod_{a\in A} (\theta^*(a))^{\nu_{x^1\diamond\ldots\diamond x^r}(a)}$$

$$\le \prod_{a\in A} (\nu_{x^1\diamond\ldots\diamond x^r}(a)/t)^{\nu_{x^1\diamond\ldots\diamond x^r}(a)}, \tag{1.51}$$

where $t = \sum_{i=1}^r t_i$. Here the equality holds, because $\theta^* \in M_0(A)$. The inequality follows from Claim 1. Indeed, if $p(a) = \nu_{x^1\diamond\ldots\diamond x^r}(a)/t$ and $q(a) = \theta^*(a)$, then

$$\sum_{a\in A} \frac{\nu_{x^1\diamond\ldots\diamond x^r}(a)}{t} \log \frac{(\nu_{x^1\diamond\ldots\diamond x^r}(a)/t)}{\theta^*(a)} \ge 0.$$

From the latter inequality we obtain (1.51). Taking into account the definition (1.34) and (1.51), we can see that the statement of Lemma is true for this particular case.

For any $\theta \in M_m(A)$ and $x = x_1 \dots x_s$, $s > m$, we present $\theta(x_1 \dots x_s)$ as $\theta(x_1 \dots x_s) = \theta(x_1 \dots x_m) \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)}$, where $\theta(x_1 \dots x_m)$ is the limiting probability of the word $x_1 \dots x_m$. Hence, $\theta(x_1 \dots x_s) \le \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)}$. Taking into account the inequality (1.51), we obtain $\prod_{a \in A} \theta(a/u)^{\nu_x(ua)} \le \prod_{a \in A} (\nu_x(ua)/\bar{\nu}_x(u))^{\nu_x(ua)}$ for any word u. Hence,

$$\theta(x_1 \dots x_s) \le \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_x(ua)}$$

$$\le \prod_{u \in A^m} \prod_{a \in A} (\nu_x(ua)/\bar{\nu}_x(u))^{\nu_x(ua)}.$$

If we apply those inequalities to $\theta(x^1 \diamond ... \diamond x^r)$, we immediately obtain the following inequalities

$$\theta(x^1 \diamond ... \diamond x^r) \le \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu_{x^1 \diamond \dots \diamond x^r}(ua)} \le$$

$$\prod_{u \in A^m} \prod_{a \in A} (\nu_{x^1 \diamond \dots \diamond x^r}(ua)/\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(u))^{\nu_{x^1 \diamond \dots \diamond x^r}(ua)}.$$

Now the statement of the Lemma follows from the definition (1.34). □

Proof. [Theorem 1.6] Let C_α be a critical set of the test $T_\varphi^{id}(A, \alpha)$, i.e., by definition, $C_\alpha = \{u : u \in A^t \ \& \ -\log \pi(u) - |\varphi(u)| > -\log \alpha\}$. Let μ_φ be a measure for which the claim 2 is true. We define an auxiliary set $\hat{C}_\alpha = \{u : -\log \pi(u) - (-\log \mu_\varphi(u)) > -\log \alpha\}$. We have $1 \ge \sum_{u \in \hat{C}_\alpha} \mu_\varphi(u) \ge \sum_{u \in \hat{C}_\alpha} \pi(u)/\alpha$ $= (1/\alpha)\pi(\hat{C}_\alpha)$. (Here the second inequality follows from the definition of $\hat{C}_\alpha$, whereas all others are obvious.) So, we obtain that $\pi(\hat{C}_\alpha) \le \alpha$. From definitions of $C_\alpha, \hat{C}_\alpha$ and (1.26) we immediately obtain that $\hat{C}_\alpha \supset C_\alpha$. Thus, $\pi(C_\alpha) \le \alpha$. By definition, $\pi(C_\alpha)$ is the value of the Type I error. The first statement of the theorem is proven.

Let us prove the second statement of the theorem. Suppose that the hypothesis $H_1^{id}(A)$ is true. That is, the sequence $x_1 \dots x_t$ is generated by some stationary and ergodic source τ and $\tau \ne \pi$. Our strategy is to show that

$$\lim_{t \to \infty} -\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = \infty \qquad (1.52)$$

with probability 1 (according to the measure τ). First we represent (1.52) as

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|$$

$$= t(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + \frac{1}{t}(-\log \tau(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|)).$$

From this equality and the property of a universal code (1.29) we obtain

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = t\,(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + o(1)). \qquad (1.53)$$

From (1.29) and (1.21) we can see that

$$\lim_{t\to\infty} - \log \tau(x_1 \dots x_t)/t \le h_k(\tau) \tag{1.54}$$

for any $k \ge 0$ (with probability 1). It is supposed that the process π has a finite memory, i.e. belongs to $M_s(A)$ for some s. Having taken into account the definition of $M_s(A)$ (1.18), we obtain the following representation:

$$- \log \pi(x_1 \dots x_t)/t = -t^{-1} \sum_{i=1}^{t} \log \pi(x_i/x_1 \dots x_{i-1})$$

$$= -t^{-1}(\sum_{i=1}^{k} \log \pi(x_i/x_1 \dots x_{i-1}) + \sum_{i=k+1}^{t} \log \pi(x_i/x_{i-k} \dots x_{i-1}))$$

for any $k \ge s$. According to the ergodic theorem there exists a limit

$$\lim_{t\to\infty} t^{-1} \sum_{i=k+1}^{t} \log \pi(x_i/x_{i-k} \dots x_{i-1}),$$

which is equal to $h_k(\tau)$[5,14] . So, from the two last equalities we can see that

$$\lim_{t\to\infty} (- \log \pi(x_1 \dots x_t))/t = - \sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log \pi(a/v).$$

Taking into account this equality, (1.54) and (1.53), we can see that

$$- \log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \ge t\,(\sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log(\tau(a/v)/\pi(a/v))) + o(t)$$

for any $k \ge s$. From this inequality and Claim 1.1 we can obtain that $- \log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \ge c\,t + o(t)$, where c is a positive constant, $t \to \infty$. Hence, (1.52) is true and the theorem is proven. $\square$

Proof. [Theorem 1.7] Let us denote the critical set of the test $T_\varphi^{SI}(A, \alpha)$ as C_α, i.e., by definition, $C_\alpha = \{x_1 \dots x_t :\ (t-m)\, h_m^*(x_1 \dots x_t) - |\varphi(x_1 ... x_t)|) > \log(1/\alpha)\}$. From Claim 1.2 we can see that there exists such a measure μ_φ that $- \log \mu_\varphi(x_1 ... x_t)$ $\le |\varphi(x_1 ... x_t)|$. We also define

$$\hat{C}_\alpha = \{x_1 \dots x_t :\ (t - m)\, h_m^*(x_1 \dots x_t) - (- \log \mu_\varphi(x_1 ... x_t))\,) > \log(1/\alpha)\}. \tag{1.55}$$

Obviously, $\hat{C}_\alpha \supset C_\alpha$. Let θ be any source from $M_m(A)$. The following chain of equalities and inequalities is true:

$$1 \ge \mu_\varphi(\hat{C}_\alpha) = \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \mu_\varphi(x_1 \dots x_t)$$

$$\ge \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} 2^{(t-m) h_m^*(x_1 \dots x_t)} \ge \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \theta(x_1 \dots x_t) = \theta(\hat{C}_\alpha).$$

(Here both equalities and the first inequality are obvious, the second and the third inequalities follow from (1.55) and the Lemma, correspondingly.) So, we obtain that $\theta(\hat{C}_\alpha) \le \alpha$ for any source $\theta \in M_m(A)$. Taking into account that $\hat{C}_\alpha \supset C_\alpha$, where C_α is the critical set of the test, we can see that the probability of the Type I error is not greater than α. The first statement of the theorem is proven.

The proof of the second statement will be based on some results of Information Theory. We obtain from (1.29) that for any stationary and ergodic p

$$\lim_{t\to\infty} t^{-1}|\varphi(x_1...x_t)| = h_\infty(p) \tag{1.56}$$

with probability 1. It can be seen from (1.23) that h^*_m is an estimate for the $m-$order Shannon entropy (1.20). Applying the ergodic theorem we obtain $\lim_{t\to\infty} h^*_m(x_1 \ldots x_t) = h_m(p)$ with probability 1[5,14] . It is known in Information Theory that $h_m(\varrho) - h_\infty(\varrho) > 0$, if ϱ belongs to $M_\infty(A) \setminus M_m(A)$[5,14] . It is supposed that H_1^{SI} is true, i.e. the considered process belongs to $M_\infty(A) \setminus M_m(A)$. So, from (1.56) and the last equality we obtain that $\lim_{t\to\infty}((t-m)\, h^*_m(x_1 \ldots x_t) - |\varphi(x_1...x_t)|) = \infty$. This proves the second statement of the theorem. □

Proof. [Theorem 1.8] First we prove that with probability 1 there exists the following limit $\lim_{t\to\infty} \frac{1}{t}\log(p(x_1 \ldots x_t)/r_U(x_1 \ldots x_t))$ and this limit is finite and nonnegative. Let $A_n = \{x_1, \ldots, x_n : p(x_1, \ldots, x_n) \neq 0\}$. Define

$$z_n(x_1 \ldots x_n) = r_U(x_1 \ldots x_n)/p(x_1 \ldots x_n) \tag{1.57}$$

for $(x_1, \ldots, x_n) \in A$ and $z_n = 0$ elsewhere.

Since

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

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

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

$$= z_{n-1}\int_A r_U(x_n|x_1 \ldots x_{n-1})dM_n(x_n|x_1 \ldots x_{n-1}) \le z_{n-1}$$

the stochastic sequence (z_n, B^n) is, by definition, a non-negative supermartingale with respect to P, with $E(z_n) \le 1$,[49] . Hence, Doob's submartingale convergence theorem implies that the limit z_n exists and is finite with $P-$probability 1 (see [49, Theorem 7.4.1]). Since all terms are nonnegative so is the limit. Using the definition (1.57) with P-probability 1 we have

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

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

and

$$\lim_{n\to\infty} n^{-1} \log(p(x_1 \dots x_n)/r_U(x_1 \dots x_n)) \geq 0. \quad (1.58)$$

Now we note that for any integer s the following obvious equality is true: $r_U(x_1 \dots x_t) = \omega_s \mu_U(x_1^{[s]} \dots x_t^{[s]})/M_t(x_1^{[s]} \dots x_t^{[s]})\,(1+\delta)$ for some $\delta > 0$. From this equality, (1.31) and (1.43) we immediately obtain that a.s.

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{r_U(x_1 \dots x_t)} \leq \lim_{t\to\infty} \frac{-\log \omega_t}{t}$$

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

$$\leq \lim_{t\to\infty} \frac{1}{t} \log \frac{p(x_1 \dots x_t)}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}/M_t(x_1^{[s]} \dots x_t^{[s]})}. \quad (1.59)$$

The right part can be presented as follows:

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

$$= \lim_{t\to\infty} \frac{1}{t} \log \frac{p^s(x_1 \dots x_t)\, M_t(x_1^{[s]} \dots x_t^{[s]})}{2^{-|U(x_1^{[s]} \dots x_t^{[s]})|}} \quad (1.60)$$

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

Having taken into account that U is a universal code, (1.40) and the theorem 1.2, we can see that the first term is equal to zero. From (1.39) and (1.42) we can see that a.s. the second term is equal to $\tilde{h}_s - \tilde{h}$. This equality is valid for any integer s and, according to (1.44), the second term equals zero, too, and we obtain that

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

Having taken into account (1.58), we can see that the first statement is proven.

From (1.59) and (1.60) we can can see that

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

$$+ E \log \frac{p(x_1 \dots x_t)}{p^s(x_1, \dots, x_t)}. \quad (1.61)$$

The first term is the average redundancy of the universal code for a finite- alphabet source, hence, according to the theorem 1.2, it tends to 0. The second term tends to $\tilde{h}_s - \tilde{h}$ for any s and from (1.44) we can see that it is equals to zero. The second statement is proven. $\square$

Proof. [Theorem 1.9] Obviously,

$$E(\frac{1}{t}\sum_{m=0}^{t-1}(P(x_{m+1}\in B_{m+1}|x_1...x_m)-R_U(x_{m+1}\in B_{m+1}|x_1...x_m))^2)\le \tag{1.62}$$

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

$$|P(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)-R_U(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)|)^2.$$

From the Pinsker inequality (1.5) and convexity of the KL divergence (1.6) we obtain the following inequalities

$$\frac{1}{t}\sum_{m=0}^{t-1}E(|P(x_{m+1}\in B_{m+1}|x_1...x_m)-R_U(x_{m+1}\in B_{m+1}|x_1...x_m)|+ \tag{1.63}$$

$$|P(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)-R_U(x_{m+1}\in \bar{B}_{m+1}|x_1...x_m)|)^2\le$$

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

$$\frac{const}{t}\sum_{m=0}^{t-1}(\int p(x_1...x_m)(\int p(x_{m+1}|x_1...x_m))\log\frac{p(x_{m+1}|x_1...x_m)}{r_U(x_{m+1}|x_1...x_m)}dM)dM_m).$$

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 (1.62), (1.63) and (1.46) we obtain (1.47). ii) can be derived from i) and the Jensen inequality for the function x^2. □

Proof. [Theorem 1.10] The last inequality of the following chain follows from the Pinsker's one, whereas all others are obvious.

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

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

$$\le \bar{M}^2(\int(p(x|x_1...x_m)-r_U(x|x_1...x_m))\,dM_m)^2$$

$$\le \bar{M}^2(\int|p(x|x_1...x_m)-r_U(x|x_1...x_m)|dM_m)^2$$

$$\le const\int p(x|x_1...x_m)\,\log\frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)}dM_m.$$

From these inequalities we obtain:

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

$$\int f(x)\,r_U(x|x_1...x_m)\,dM_m)^2) \leq \tag{1.64}$$

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

The last term can be presented as follows:

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

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

$$\int p(x|x_1...x_m)\log\frac{p(x|x_1...x_m)}{r_U(x|x_1...x_m)}dM_{1/m}dM_m$$

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

From this equality, (1.64) and Corollary 1 we obtain (1.48). ii) can be derived from (1.64) and the Iensen inequality for the function x^2. □

Proof. [Theorem 1.11] The following chain proves the first statement of the theorem:

$$P\{H_0^{\aleph}(A)\quad is\ rejected\ /H_0 is true\} = P\{\bigcup_{i=1}^{\infty}\{H_0^{\aleph}(\Lambda_i)\quad is\ rejected\ /H_0 is true\}\} \leq$$

$$\sum_{i=1}^{\infty} P\{H_0^{\aleph}(\Lambda_i)\ /H_0\ is\ true\}\ \leq \sum_{i=1}^{\infty}(\alpha\omega_i)\ = \alpha.$$

(Here both inequalities follow from the description of the test, whereas the last equality follows from (1.24).)

The second statement also follows from the description of the test. Indeed, let a sample is created by a source ϱ, for which $H_1(A)^{\aleph}$ is true. It is supposed that the sequence of partitions $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A), H_1^{\aleph}(A)$. By definition, it means that there exists j for which $H_1^{\aleph}(\Lambda_j)$ is true for the process ϱ_{Λ_j}. It immediately follows from Theorem 1.1 - 1.4 that the Type II error of the test $T_{\varphi}^{\aleph}(\Lambda_j, \alpha\omega_j)$ goes to 0, when the sample size tends to infinity. □

1.9. A list of terminologies/keywords

Empirical Shannon entropy of order k.

$$h_k^*(x) = -\sum_{v \in A^k} \frac{\bar{\nu}_x(v)}{(t-k)} \sum_{a \in A} \frac{\nu_x(va)}{\bar{\nu}_x(v)} \log \frac{\nu_x(va)}{\bar{\nu}_x(v)}$$

where $x = x_1 \dots x_t$, $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va)$. In particular, if $k = 0$, we obtain $h_0^*(x) = -t^{-1} \sum_{a \in A} \nu_x(a) \log(\nu_x(a)/t)$.

Goodness-of-fit testing. The hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{id}.

i.i.d. sources generates independent and identically distributed random variables.

Krichevsky predictor for i.i.d. sources. $L_0(a|x_1 \cdots x_t) = (\nu_{x_1 \cdots x_t}(a) + 1/2)/(t + |A|/2)$, where $\nu_{x_1 \cdots x_t}(a)$ denotes the count of letter a occurring in the word $x_1 \dots x_{t-1} x_t$.

Kullback-Leibler (KL) divergence.

$$D(P,Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} ,$$

where $P(a)$ and $Q(a)$ are probability distributions over an alphabet A (here and below $\log \equiv \log_2$ and $0 \log 0 = 0$).

Laplace predictor for i.i.d. sources. $L_0(a|x_1 \cdots x_t) = (\nu_{x_1 \cdots x_t}(a) + 1)/(t + |A|)$, where $\nu_{x_1 \cdots x_t}(a)$ denotes the count of letter a occurring in the word $x_1 \dots x_{t-1} x_t$.

Markov sources of order (or with memory) m, $m \geq 0$.

$$\mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, \dots, x_{t-m+1} = a_{i_{m+1}}, \dots)$$

$$= \mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, \dots, x_{t-m+1} = a_{i_{m+1}})$$

for all $t \geq m$ and $a_{i_1}, a_{i_2}, \dots \in A$.

Measure R.

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

where K_i is the Krichevsky predictors for the set of i-memory Markov sources, $\{\omega = \omega_1, \omega_2, \dots\}$ is the following probability distribution on integers: $\omega_1 = 1 - 1/\log 3$, ..., $\omega_i = 1/\log(i+1) - 1/\log(i+2)$, ...

Pinsker inequality.

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

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

Serial independence test. The null hypothesis H_0^{SI} is that the source is Markovian of order not larger than m, $(m \geq 0)$, and the alternative hypothesis H_1^{SI} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. In particular, if $m = 0$, this is the problem of testing for independence of time series.

Shannon entropy for the i.i.d. source P is
$h_0(P) = -\sum_{a \in A} P(a) \log P(a)$.

The $m-$ order (conditional) Shannon entropy and the limiting Shannon entropy are defined as follows:

$$h_m(P) = \sum_{v \in A^m} P(v) \sum_{a \in A} P(a/v) \log P(a/v), \qquad h_\infty(\tau) = \lim_{m \to \infty} h_m(P).$$

Stationary ergodic processes. 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 $T^{-1}B = B$.

Universal code. U is a universal code if for any stationary and ergodic source P the following equalities are valid:

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

with probability 1, and

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

where $E(f)$ is the expected value of f, $h_\infty(P)$ is the Shannon entropy of P, see (1.21).

References

1. P. Algoet, Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information, *IEEE Trans. Inform. Theory,* **45**, 1165-1185, (1999).
2. G. J. Babu, A. Boyarsky, Y. P. Chaubey, P. Gora, New statistical method for filtering and entropy estimation of a chaotic map from noisy data, *International Journal of Bifurcation and Chaos,* **14** (11), 3989-3994, (2004).
3. A.R. Barron, The strong ergodic theorem for dencities: generalized Shannon-McMillan-Breiman theorem, *The annals of Probability*, **13** (4), 1292–1303, 1985.
4. L.Györfi, I.Páli and E.C. van der Meulen, There is no universal code for infinite alphabet, *IEEE Trans. Inform. Theory,* **40**, 267–271, 1994.
5. P. Billingsley, *Ergodic theory and information.* (John Wiley & Sons, 1965).
6. R. Cilibrasi and P. M.B. Vitanyi, Clustering by Compression, *IEEE Transactions on Information Theory*, **51** (4), (2005).
7. R. Cilibrasi, R. de Wolf and P. M.B. Vitanyi, Algorithmic Clustering of Music, *Computer Music Journal*, **28** (4) 49–67, (2004).
8. I. Csiszár and P. Shields, *Notes on information theory and statistics.* (Foundations and Trends in Communications and Information Theory, 2004).

9. I. Csiszár and P. Shields, The consistency of the BIC Markov order estimation. *Annals of Statistics,* **6**, 1601-1619, 2000.
10. M. Effros, K. Visweswariah, S. R.Kulkarni and S. Verdu, Universal lossless source coding with the Burrows Wheeler transform, *IEEE Trans. Inform. Theory,* **45**, 1315–1321, (1999).
11. W. Feller, *An Introduction to Probabability Theory and Its Applications*, vol.1. (John Wiley & Sons, New York, 1970).
12. L. Finesso, C. Liu, and P. Narayan, The optimal error exponent for Markov order estimation, *IEEE Trans. Inf. Theory,* **42**, (1996).
13. B. M. Fitingof, Optimal encoding for unknown and changing statistica of messages, *Problems of Information Transmission,* **2** (2), 3–11, (1966).
14. R. G. Gallager, *Information Theory and Reliable Communication.* (John Wiley & Sons, New York, 1968).
15. E. N. Gilbert, Codes Based on Inaccurate Source Probabilities, *IEEE Trans. Inform. Theory,* **17**, (1971).
16. N.Jevtic, A.Orlitsky and N.P.Santhanam. A lower bound on compression of unknown alphabets. *Theoretical Computer Science,* **332**, 293–311, 2004.
17. J. L. Kelly, A new interpretation of information rate, *Bell System Tech. J.*, **35**, 917–926, (1956).
18. J. Kieffer. A unified approach to weak universal source coding . *IEEE Trans. Inform. Theory,* **24**, 674–682, 1978.
19. J. Kieffer, Prediction and Information Theory, *Preprint*, (available at ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/), 1998.
20. J. C. Kieffer and En-Hui Yang, Grammar-based codes: a new class of universal lossless source codes. *IEEE Transactions on Information Theory,* **46** (3), 737–754, (2000).
21. A. N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission*, **1**, 3–11, (1965).
22. D. E. Knuth *The art of computer programming.* Vol.2. (Addison Wesley, 1981).
23. R. Krichevsky, A relation between the plausibility of information about a source and encoding redundancy, *Problems Inform. Transmission,* **4**(3), 48–57, (1968).
24. R. Krichevsky, *Universal Compression and Retrival*, (Kluver Academic Publishers, 1993).
25. S. Kullback, *Information Theory and Statistics.* (Wiley, New York, 1959).
26. U. Maurer, Information-Theoretic Cryptography, In: *Advances in Cryptology - CRYPTO '99, Lecture Notes in Computer Science,* Springer-Verlag, vol. 1666, pp. 47–64, (1999).
27. D. S. Modha and E. Masry, Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory,* **44**(1), 117–133, (1998).
28. A. B. Nobel, On optimal sequential prediction, *IEEE Trans. Inform. Theory*, **49**(1), 83–98, (2003).
29. A. Orlitsky, N. P. Santhanam, and J. Zhang, Always Good Turing: Asymptotically Optimal Probability Estimation, *Science*, **302**, (2003).
30. Zh. Reznikova, *Animal Intelligence. From individual to social cognition.* (CUP, 2007).
31. J. Rissanen, Generalized Kraft inequality and arithmetic coding, *IBM J. Res. Dev.*, **20** (5), 198–203, (1976).
32. J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory,* 30(4), 629–636, (1984).
33. A. Rukhin 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

34. B. Ya. Ryabko, Twice-universal coding, *Problems of Information Transmission,* **20**(3), 173–177, (1984).
35. Ryabko B.Ya., Prediction of random sequences and universal coding. *Problems of Inform. Transmission*, **24**(2) 87-96, (1988).
36. B. Ya. Ryabko, A fast adaptive coding algorithm, *Problems of Inform. Transmission,* **26**(4), 305–317, (1990).
37. B. Ya. Ryabko, The complexity and effectiveness of prediction algorithms, *J. Complexity,* **10**(3), 281–295, (1994).
38. B. Ryabko, J. Astola and A. Gammerman, Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series, *Theoretical Computer Science,* **359**, 440-448, (2006).
39. B. Ryabko, J. Astola and A. Gammerman, Adaptive Coding and Prediction of Sources with Large and Infinite Alphabets, *IEEE Transactions on Information Theory*, **54**(8), (2008).
40. B. Ryabko, J.Astola and K. Egiazarian, Fast Codes for Large Alphabets, *Communications in Information and Systems*, **3** (2), 139–152, (2003).
41. B. Ryabko and V. Monarev, Experimental Investigation of Forecasting Methods Based on Data Compression Algorithms. *Problems of Information Transmission,* **41**, (1), 65-69, (2005).
42. B. Ryabko and V. Monarev, Using Information Theory Approach to Randomness Testing, *Journal of Statistical Planning and Inference,* **133**(1), 95–110, (2005).
43. B. Ryabko and Zh. Reznikova, Using Shannon Entropy and Kolmogorov Complexity To Study the Communicative System and Cognitive Capacities in Ants, *Complexity,* **2** (2), 37–42, (1996).
44. B. Ryabko and F. Topsoe, On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources, *Journal of Complexity,* **18**(1), 224–241, (2002).
45. D. Ryabko and M. Hutter, 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
46. S. A. Savari, A probabilistic approach to some asymptotics in noiseless communication, *IEEE Transactions on Information Theory*, **46**(4), 1246–1262, (2000).
47. C. E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.* , **27**, 379–423, 623–656, (1948).
48. C. E. Shannon, Communication theory of secrecy systems, *Bell Sys. Tech. J.*, **28**, 656–715, (1948).
49. A.N. Shiryaev, *Probability*, (second edition), Springer, 1995.