Compression-based methods for density estimation for time series

Boris Ryabko
Institute of Computational Technologies of Siberian Branch of Russian Academy of Science and
Siberian State University of Telecommunications and Informatics,
Kirov Str. 86, 630102, Novosibirsk, Russia
Email: boris@ryabko.net

Abstract

We address the problem of nonparametric estimation of the density for time series. We show that any universal code (or a universal data compressor) can be used as a basis for constructing an asymptotically optimal method for this problem for a certain class of stationary and ergodic processes.

I. Introduction

Nowadays the most known practical application of universal codes is data compressors, which have shown their high efficiency as compressors of real data. It could be less known that the universal codes and archivers have begun to play an important role in hypothesis testing [28], [29] and prediction of time series [30]. Moreover, their applications in [6], [7] created a new and rapidly growing line of investigation in clustering and classification.

In this paper we show that universal codes can be applied to nonparametric density estimation and related problems for a certain class of stationary and ergodic time series. It is important to note that the problem of density estimation for different classes of time series has attracted attention of many researchers. Known density estimation methods include histogram, kernel, orthogonal series and many others; see [1], [4], [20], [22], [24], [34] and references there. In addition to the density estimation we consider the problem of prediction, because of its practical applications and importance for probability theory, information theory and statistics, see [1], [15], [19], [20], [24], [34].

We consider a stationary and ergodic source with unknown statistics which generates sequences $x_1x_2\cdots$ of letters from some set (or alphabet) A. Since universal codes are defined for sources generating letters from a finite alphabet we first consider this case. Informally, a universal code compresses a sequence generated by a stationary and ergodic source till the Shannon entropy (per letter), which is a lower bound for the compression ratio. We will show that universal codes can be directly applied to the density estimation, prediction and related problems for a certain class of a real-valued time series. It is worth noting that everyday methods of data compression (or archivers) like *zip*, *arj, rar,* etc., can be used as a tool for the density estimation and prediction, because the modern archivers are based on deep theoretical results of the universal code theory (see, for ex., [12], [16], [18], [24], [32]).

II. Definitions and Preliminaries

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

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

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

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 $u_1u_2...u_r$. Such codes are called uniquely decodable. It is well known that if a code φ is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft's inequality): $\sum_{u\in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, for ex., [13]. In what follows we call uniquely decodable codes just "codes".

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

$$\lim_{t\to\infty} |U(x_1 \dots x_t)|/t = H(P) \tag{2}$$

with probability 1, and

$$\lim_{t\to\infty} E(|U(x_1 \dots x_t)|)/t = H(P), \tag{3}$$

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

The well known Shannon-MacMillan-Breiman theorem states that for any stationary and ergodic source

$$\lim_{t\to\infty} -\log P(x_1 \dots x_t)/t = H(P) \tag{4}$$

with probability 1, see [5], [13]. This theorem plays a key role in our consideration, because we can see from (2) and (4) that

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

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

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

$$KL_t(P, \gamma) = \sum_{a\in A^t} P(v) \log \frac{P(v)}{\gamma(v)}. \tag{5}$$

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

$$\sum_{a\in A} P(a) \log \frac{P(a)}{Q(a)} \geq \frac{\log e}{2} ||P - Q||^2. \tag{6}$$

connects the KL divergence with a so-called variation distance

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

where P and Q are distributions over A, see [8]. It will be convenient to combine all properties of the probability estimators, which are based on universal codes.

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

$$\mu_U(u) = 2^{-|U(u)|}/\Sigma_{v\in A^{|u|}} 2^{-|U(v)|}. \tag{7}$$

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

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

with probability 1,

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

Now we briefly consider the problem of prediction. Let the (unknown) source P generate a message $x_1 \dots x_{t-1}x_t$, and the following letter x_{t+1} needs to be predicted. If one knows conditional probabilities $P(x_{t+1} = a|x_1 \dots x_t), a \in A$, he knows all about x_{t+1}, i.e. 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. That is why we will consider prediction as a set of estimations of unknown (conditional) probabilities. This approach to the problem of prediction was developed in [26] and now is often called on-line prediction or universal prediction [1], [20], [22]. At first glance it seems natural to estimate a precision of some prediction method γ by one of the two following values:

$$\log \frac{P(x_{t+1}|x_1 \dots x_t)}{\gamma(x_{t+1}|x_1 \dots x_t)},\ \sum_{a \in A} P(a|x_1 \dots x_t) \log \frac{P(a|x_1 \dots x_t)}{\gamma(a|x_1 \dots x_t)}, \tag{8}$$

where $\gamma(.|x_1 \dots x_t)$ is an estimation (a probability distribution) and $x_1 \dots x_t$ is a word generated by the unknown source. It turns out that this approach is contradictory. The point is that for any predictor γ there exists a stationary and ergodic source such that both values in (8) do not go to 0, when $t \to \infty$ (with probability 1). (The proof is given in [26]; see also [1], [20], where, in particular, it is mentioned that this fact was described in unpublished thesis [2].) On the other hand, it is proven in [26] that there exists a predictor R for which the following Cesaro averages go to 0 for any stationary and ergodic source: $t^{-1} \sum_{i=0}^{t-1} \log(P(x_{i+1}|x_1 \dots x_i)/R(x_{i+1}|x_1 \dots x_i)) = 0$ (with probability 1) and $t^{-1} \sum_{i=0}^{t-1} P(x_1 \dots x_{i+1}) \log(P(x_{i+1}|x_1 \dots x_i)/R(x_{i+1}|x_1 \dots x_i)) = 0$. Hence, for any predictor γ it is natural to estimate its error by values $t^{-1} \sum_{i=0}^{t-1} \log(P(x_{i+1}|x_1 \dots x_i)/\gamma(x_{i+1}|x_1 \dots x_i))$, (with probability 1) and $t^{-1} \sum_{i=0}^{t-1} P(x_1 \dots x_{i+1}) \log(P(x_{i+1}|x_1 \dots x_i)/\gamma(x_{i+1}|x_1 \dots x_i))$, which, in turn, are equal to the following expressions $t^{-1} \log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t))$, $t^{-1} P(x_1 \dots x_t) \log(P(x_1 \dots x_t)/\gamma(x_1 \dots x_t))$, correspondingly. So, if we take a universal code U and use it for prediction, the Theorem 1 will be true for the corresponding measure μ_U. In other words, from mathematical point of view the problems of probability estimation and prediction are completely the same and can be considered together.

III. Time series with a density

Here we will consider problems of the density estimation and prediction for a stationary process with densities. We have seen that Shannon-MacMillan-Breiman theorem played a key role in the case of finite-alphabet processes. In this part we will use its generalization to the processes with densities. This result was proved by Barron [3] and was an extension of the L^1 convergence obtained in [21], [23], [14]. First we describe considered processes with some properties needed for a fulfilment of the generalized Shannon-MacMillan-Breiman theorem.

Let (Ω, F, P) be a probability space and let $X_1, X_2, \dots$ be a stochastic process with each X_t taking values in a standard Borel space. As in [3], 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. Familiar cases for M_n are Lebesgue measure and counting 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 h defined by

$$h = \lim_{n \to \infty} -E(\log p(x_{n+1}|x_1 \dots x_n)), \tag{9}$$

where E denotes expectation with respect to P. The following generalization of the Shannon-MacMillan-Breiman theorem is obtained by Barron in [3]:

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

$$\lim_{n \to \infty} (-1/n) \log p(x_1 \dots x_n) = h \tag{10}$$

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 Ω that asymptotically generates the Borel sigma-field F and let $x^{[k]}$ denote the element of Π_k that contains the point x. (Informally, if Ω is an interval, $x^{[k]}$ is obtained by quantizing x to k bits of precision.) For integers s and n we define the following approximation of the density

$$p^s(x_1 \dots x_n) = P(x_1^{[s]} \dots x_n^{[s]})/M_n(x_1^{[s]} \dots x_n^{[s]}). \tag{11}$$

We also consider $h_s = \lim_{n\to\infty} -E(\log p^s(x_{n+1}|x_1 \dots x_n))$. Applying the claim to the density $p^s(x_1 \dots x_t)$, we obtain that a.s. $\lim_{t\to\infty} -\frac{1}{t}\log p^s(x_1 \dots x_t) = h_s$. Let U be a universal code, which is defined for any finite alphabet. In order to describe the density estimate we first define a probability distribution $\{\omega = \omega_1, \omega_2, ...\}$ on integers $\{1, 2, ...\}$ by

$$\omega_1 = 1 - 1/\log 3,\ \dots,\ \omega_i = 1/\log(i+1) - 1/\log(i+2),\ \dots . \tag{12}$$

(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:

$$r_U(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_i\, \mu_U(x_1^{[i]} \dots x_t^{[i]})/M_t(x_1^{[i]} \dots x_t^{[i]})\ , \tag{13}$$

where the measure μ_U is defined by (7). (It is supposed here that the code $U(x_1^{[i]} \dots x_t^{[i]})$ is defined for the alphabet, which contains $|\Pi_i|$ letters.)

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

Theorem 2 . *Let X_t be a stationary ergodic process with densities $p(x_1 \dots x_t) = dP_t/dM_t$ such that $\lim_{s\to\infty} h_s = h < \infty$. Then*

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

with probability 1 and

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

The following theorem is 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 $p(x|x_1 ... x_m)$.

Theorem 3. *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:*

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

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

Comment. In fact, the statements i) and ii) are equivalent, because one of them follows from the other. Our proof is similar to the method from [31], see Lemma 2 there.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] D. H. Bailey, *Sequential schemes for classifying and predicting ergodic processes* , PhD Dissertation, Stanford University, 1976.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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