# On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Sources with Unknown Memory

Boris Ryabko[1]
Siberian State University of
Telecom. and Inform. Sci.
Kirov St. 86, Novosibirsk 630102
Russia
e-mail: ryabko@neic.nsk.su

Flemming Topsøe
Department of Mathematics,
University of Copenhagen
Universitetsparken 5, Copenhagen
DK-2100 Denmark
e-mail: topsoe@math.ku.dk

*Abstract* — **The asymptotically optimal methods of prediction for Markov sources with unknown memory are suggested. The methods are based on modified twice universal scheme.**

## I. Introduction

The problem of prediction and the closely related problem of adaptive coding of time series is well known in Information Theory, Probability Theory and Statistics [1].

We consider a source with unknown statistics which generates sequences $x_1x_2\ldots$ of letters from a finite alphabet $A = \{a_1, \ldots, a_n\}$. We imagine that we have at our disposal a computer for solving the prediction problem. As input we consider any finite string $x_1x_2\ldots x_t$ of letters from $A$ and as output we receive at each time instant $t$ non-negative numbers $p^*(a_1|x_1\ldots x_t), \ldots, p^*(a_n|x_1\ldots x_t)$ which are estimates of the unknown conditional probabilities $p(a_1|x_1\ldots x_t), \ldots, p(a_n|x_1\ldots x_t)$, i.e., of the probabilities $p(x_{t+1} = a_i|x_1\ldots x_t)$; $i = 1, \ldots, n$. The set $p^*(a_i|x_1\ldots x_t)$; $i \le n$ is called the *prediction.*

The *precision* of a prediction method is measured by the divergence between $p$ and $p^*$ and the *complexity* of a method is characterized by two numbers: the *average time* of calculation at each time instant in bit operations and the *memory size* in bits of the program defining the method. Let us denote the set of Markov sources of memory (or connectivity) $k$ as $M_k(A)$ and let $M_0(A)$ be the set of all Bernoulli sources.

In this report we consider the prediction problem for Markov sources with unknown statistics and memory.

## II. The Main Results

We will use two asymptotically optimal prediction methods for $M_i(A), i = 0, 1, \ldots$, which were suggested in [2]. The method $\alpha_i$ is asymptotically optimal in average and $\beta_i$ with probability one.

According to twice universal scheme, at each time instant $t$ a computer compares the average precision of all methods $\beta_0, \beta_1, \ldots, \beta_N$ on the interval $t = 1, 2, \ldots, T-1$ and finds $j_0$ for which $\beta_{j_0}$ gives the best precision on the interval $t = 1, 2, \ldots, T-1$. Then the computer uses $\beta_{j_0}$ in order to predict for the next moment $T$. (It looks like the likelihood principle).

It is clear that the computer should calculate $(N+1)$ prediction sets (for $\beta_0, \beta_1, \ldots, \beta_N$) instead of one set as it does in case of known memory of the source. So the time of calculation increases $(N+1)$ times. Similarly, the memory space of the computer should be divided into $(N+1)$ parts in order to store statistics for $\beta_0, \beta_1, \ldots, \beta_N$.

The new methods are based on a simplified twice universal scheme (STUS). According to STUS, a computer which is used for the implementation of the suggested method compares two methods $\beta_{i_1}$ and $\beta_{i_2}$ at each time instant $t$. First, at $t = 1, 2, \ldots, T$ the computer compares $\beta_0$ and $\beta_1$ which are optimal for $M_0(A)$ and $M_1(A)$ ( $T$ is a parameter of the method). Then the computer removes the worst method and includes $\beta_2$ instead of it. After that both methods are compared during the period of $[T+1, \ldots, 2T]$, the worst of them is removed and so on. At each time instant $t$ the computer uses the best method $\beta_{i_j}$ for prediction. (At the first interval $[1, \ldots, T]$ $\beta_0$ is used). At the moment $(N+1)T+1$ the computer again includes $\beta_0$ instead of removed $\beta_{i_j}$. And so on. It is quite obvious that the computer will find the best $\beta_i$ and will use it almost all time for prediction if $T$ is quite large. On the other hand, this universal scheme is fast and space-efficient because at every moment only two methods are compared instead of $N$ in the "conventional" twice universal scheme. We designate this method as $\beta^1_{stu}$ and describe two other modifications.

The $\beta^1_{stu}$ is effective with probability 1. We obtain the method $\beta^2_{stu}$ which is simpler if the computer stops to look for the best method $\beta^1_{stu}$ after the moment $(N+1)T$ and uses for prediction at the moments $(N+1)T+1, (N+1)T2, \ldots$ the $\beta_{i_j}$ which was the best during $[NT+1, \ldots, (N+1)T]$. The new method $\beta^2_{stu}$ is effective in average only. (For simplification of the method it is possible to use optimal in average $\alpha_{i_j}$ instead of $\beta_{i_j}$). The last modification $\beta^3_{stu}$ may be used when $N$ is infinite or when it is known only that a source is ergodic. The method $\beta^3_{stu}$ looks like $\beta^2_{stu}$ but the computer includes randomly chosen method $\beta_i$ from the $\beta_0, \beta_1, \ldots$ (Recall, that $\beta_i$ is included instead of the worst method $\beta_{i_j}$ at the moments $T+1, 2T+1, 3T+1, \ldots$).

The main property of the suggested STUS may be formulated as follows: if $\beta^1_{stu}$ is used with $T(r) = \left\lceil \left(\log \frac{1}{r}\right)^2 \right\rceil$, where $r$ is the precision, then for every $M_i(A)$ its precision is asymptotically equal to the precision of the method which is optimal for $M_i(A)$, when $r$ goes to 0.

## References

[1] P. Algoet, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Trans. Inform. Theory,* vol. 45, no. 4, pp. 1165–1185, 1999.

[2] B. Ryabko and F. Topsoe, "On asymptotically optimal methods of prediction and adaptive coding," in *Proc. IEEE Int. Symp. Inform. Theory,* Cambridge, MA, August 1998, p. 316.

[1] This work was supported by RFBR Grant 99-01-00586.

0-7803-5857-0/00/$10.00 ©2000 IEEE.