Adaptive Coding and Prediction of Sources with Large and Infinite Alphabets *

Boris Ryabko[†] and Jaakko Astola [‡]

[†] Siberian State University of Telecommunication and Computer Science, Russia
[‡] Technical University of Tampere, Finland

The problem of predicting a sequence $x_1, x_2, \cdots$ generated by a discrete source with unknown statistics is considered. This problem is of great importance for data compression, because of its use to estimate probability distributions for PPM algorithms and other adaptive codes. Laplace suggested the following prediction $P^*_L(a|x_1 \cdots x_t) = (\nu^t(a)+1)/(t+|A|)$, where $\nu^t(a)$ denote the count of letter a occurring in the word $x_1 \ldots x_{t-1}x_t$ and P^* is the estimation of the probability. It is known that the redundancy of the Laplace predictor is upper bounded by $(|A|-1)/(t+1)$, if the predictor is applied to an i.i.d. source. Krichevsky suggested the predictor $p^*_K(a|x_1 \cdots x_t) = (\nu^t(a)+1/2)/(t+|A|/2)$ and showed that its redundancy is asymptotically minimal. We suggest a scheme of adaptive coding (and prediction) for a case where a source generates letters from an alphabet with unknown or infinite size. This scheme can be applied along with Laplace, Krichevsky and any other predictors.

The following example is to explain the main idea of the scheme. Let $A = \{a_0, a_1, a_2\}$ and $t = 4$, $x_1x_2x_3x_4 = a_0a_2a_0a_1$. In this example we suggest the following grouping of letters into two subsets $A_0 = \{a_0, a_1\}, A_1 = \{a_3\}$ and carry out the prediction into two steps. First, $a_0a_2a_0a_1$ is represented as $A_0A_1A_0A_0$ and belonging to the subsets is predicted. Then, the sequence $A_0A_0A_0 = a_0a_0a_1$ is used for predicting conditional probabilities $p(x_5 = a_i/x_5 \in A_0)$, $i = 0, 1$. If we use the Laplace predictor, we obtain $p^*_L(x_5 = a_0) = (4/6)(3/5) = 2/5$, $p^*_L(x_5 = a_1) = (2/6)(2/5) = 2/15$, $p^*_L(x_5 = a_2) = 1/3$. The general case of the prediction, which is based on such a grouping, is considered and the estimates of the redundancy are given.

If the suggested scheme is applied to $s-$ letter source and s is unknown, the redundancy is asymptotically the same as if the predictor is applied to the $(s+1)-$ letter source and the alphabet size $(s+1)$ is known beforehand. When the suggested scheme is applied to an infinite alphabet, the redundancy of the code goes to 0, if, loosely speaking, the original representation of the alphabet letters has a finite average word length. It is shown that, in fact, this condition is necessary for existing of such predictors.

*The first author was supported by INTAS grant no. 00-738 and Russian Foundation for Basic Research under Grant no. 03-01-00495.