Efficient Methods for Generating Approximately Equiprobable Random Bits

Boris Ryabko[1] and Elena Matchikina[1]
Sibirean State University of Telecomunications and Computer Science
Kirova str. 86, 630102, Novosibirsk, Russia
e-mail: {ryabko, e_match}@neic.nsk.su

The problem of random number generation is well-known in Information Theory and investegated by a number of authors due to its great importance for cryptology, Monte-Carlo methods and other applications. In [2] (see also [3]) the problem of generating almost equiprobable bits from arbitrary but known source was set up and investigated. But the proposed methods require exponentional growth of the necessary memory size. In this article we propose the methods for generation almost equiprobable random bits which are based on efficient algorithms of arithmetic coding for known and unknown sources. Complexity of the suggested methods is less in order of magnitude than that for known exact and approximate methods.

Let us stricly formulate the problem. Let there be given a source generating a sequence of letters $x_1x_2\dots x_N\dots$ over the alphabet $A=\{a_1,a_2,\dots,a_s\}$, $x_i\in A$. Denote by U^n the uniform distribution on the set $\{0,1\}^n$, $n\geq 1$, by $\tilde{U}^n$ the probability distribution in the output sequence of length n. We need to convert the input sequence $\{x_i\}$ into the output sequence $\{y_j\}$, $y_j\in\{0,1\}$, so that $\lim_{n\to\infty} d(\tilde{U}^n,U^n)\leq\rho$, where $d(\cdot,\cdot)$ is a measure of approximation, ρ is an a priori given deviation, $\rho\geq 0$. We shall assess the converting method complexity by memory size (M) required to the encoder, as well as by the average time (T) of encoding one symbol measured by the number of binary operations on single-bit words on a computer with random access memory. These two characteristics of complexity are considered as functions of deviation ρ. As it was suggested in [2] we use three measures of approximation of probability distributions: Kullbuck-Leibler's divergence, variational distance and Ornstein's distance.

Further we assume the source to be a Bernoulli source but all the results obtained for memoryless sources are extended to a Markov source of order μ, $\mu\geq 1$, with source alphabet A. For coding such a source the well-known in Information Theory technique of decomposing a Markov source into $|A|^\mu$ memoryless sources may be used. The main result is the following

Theorem 1 *There exists a method of converting the sequences of simbols which are generated by a Bernoulli source with known (or unknown) probabilities into sequences with distribution of probabilities which differs from uniform distribution not greater than ρ (in the sense one of the measures of approximation) with the following characteristics*

$$T=O\left(\log\frac{1}{\rho}\log\log\frac{1}{\rho}\log\log\log\frac{1}{\rho}\right),$$

$$M=O\left(\log\frac{1}{\rho}\right),$$

where T is time of calculation per symbol of source, M is required size of memory, $\rho\to 0$.

The idea of proof of Theorem 1 is the following. We use a well-known in Information Theory fact that source coding with arbitrarily small redundancy gives the output sequence of almost equiprobable bits (here the redundancy is defined as the difference between the mean codeword length and the source entropy). The next lemma states a relationship between the redundancy of coding and Kullbuck-Leibler' divergence (d_1) which is the measure of approximation between the coded sequence and the uniformly distributed sequence.

Lemma 1 *Let there be given a memoryless source with the Shannon entropy h_∞. Let coding with the redundancy r per symbol be applied for encoding symbols generated by the source. Denote by $\tilde{U}^n$ the distribution of probabilities on the set $\{0,1\}^n$, $n\geq 1$ in the coded sequence. Then*

$$\lim_{n\to\infty} d_1(\tilde{U}^n,U^n)=\frac{r}{h_\infty+r}.$$

Using the lemma we can assess the complexity of the method of coding as function of deviation ρ. For transforming input sequence we use the method of arithmetic coding from [1] when statistics of source are known. For unknown statistics of source we proposed the method which is based on arithmetic coding and sliding window. The sliding window is used for estimating source probabilities. The only disadvantage of the sliding window scheme is the necessity to store the entire window in the encoder memory. To remedy this disadvantage we use a type of the scheme called "imaginary sliding window" (the concept of which was presented by B. Ryabko in 1997) but such a scheme requires some random bits. A distinctive feature of this scheme is that storing the window is not needed.

References

[1] B. Ya. Ryabko, A. N. Fionov, "Fast and space-efficient adaptive arithmetic coding," // *Cryptography and Coding*, (Lecture Notes in Comput. Sci.; vol. 1746) Berlin: Springer, pp. 270–279, 1999.

[2] S. Vembu, S. Verdú, "Generating random bits from arbitrary source: fundamental limits" *IEEE Trans. Inform. Theory*, vol. 41, pp. 1322–1332, 1995.

[3] K. Visweswariah, S.R. Kulkarni, S. Verdú "Separation of random number generation and resolvability" *IEEE Trans. Inform. Theory*, vol 46, no. 6, pp. 2237–2242, 2000.

[1] This work was supported by INTAS (Grant 00-738).