Twice-universal coding

Boris Ryabko

1984

Abstract

Assume that A is a finite alphabet; Ω_i is a set of Markov sources of connectedness i that generate letters from A ($i = 1, 2, \ldots$) and Ω_0 is a set of Bernoulli sources. A code is proposed whose redundancy as a function of the block length on each Ω_i is asymptotically as small as that of the universal code that is optimal on Ω_i ($i = 1, 2, \ldots$). A generalization of this problem to the case of an arbitrary countable family of sets of stationary ergodic sources is considered.

Problems of Information Transmission, 1984, 20:3, 173–177

are new and improve the available lower bounds for $d_{max}(n; k)$. The parameters of our codes that improve the values of the table of binary linear codes with $n \leqslant 127$ and $k \leqslant 127$ from [1] are given in the accompanying table. Linear [55, 7, 25], [56, 7, 26], and [57, 7, 26] codes are mentioned in [2], (p. 657).

LITERATURE CITED

1. H. J. Helgart and R. D. Stinaff, "Minimum-distance bounds for binary linear codes," IEEE Trans. Inf. Theory, 19, No. 3, 344-356 (1973).
2. F. J. McWilliams and N. J. Sloane, Theory of Error-Correcting Codes [Russian translation], Svyaz', Moscow (1979).
3. G. P. Gavrilov and A. A. Sapozhenko, Collection of Problems in Discrete Mathematics [in Russian], Nauka, Moscow (1977).
4. P. Delsarte, "Four fundamental parameters of a code and their combinatorial significance," Inf. Control, 23, No. 5, 407-438 (1973).
5. L. A. Bassalygo and V. A. Zinov'ev, "Remark on uniformly packed codes," Probl. Peredachi Inf., 8, No. 3, 22-25 (1977).

TWICE-UNIVERSAL CODING

B. Ya. Ryabko

UDC 621.391.15

Assume that A is a finite alphabet; Ω_i is a set of Markov sources of connectedness i that generate letters from A($i = 1, 2, \ldots$); and Ω_0 is a set of Bernoulli sources. A code is proposed whose redundancy as a function of the block length on each Ω_i is asymptotically as small as that of the universal code that is optimal on Ω_i ($i = 0, 1, 2, \ldots$). A generalization of this problem to the case of an arbitrary countable family of sets of stationary ergodic sources is considered.

Assume that Ω_∞ is the set of all stationary ergodic sources that generate letters from some finite alphabet A. Noiseless block code $\mathfrak{L}$ is called universal [1] or strongly universal [2] on the set of sources $\Omega \subset \Omega_\infty$ if, with increasing block length n, the redundancy, i.e., the difference between the mean length of a code word $(c_n(\mathfrak{L}, \omega))$ and the entropy $(H(\omega))$ for each $\omega \in \Omega$, converges to 0, this convergence being uniform with respect to $\omega \in \Omega$. If, however, the convergence is not uniform, the code is called weakly universal [2]. Assume that $\{\Omega_\lambda\}$, $\lambda \in \Lambda$, is a finite or countable family of sets of sources, for each of which there exists a strongly universal (SU) code. By $\mathfrak{L}^\lambda$ we denote the SU code that is optimal on Ω_λ, i.e., a code for which $R_n(\mathfrak{L}^\lambda, \Omega_\lambda) = \sup\{c_n(\mathfrak{L}^\lambda, \omega) - H(\omega), \omega \in \Omega_\lambda\}$ is minimal for block lengths $n = 1, 2, \ldots$.

In this paper we propose a code $W(\Lambda)$ such that for all $\lambda \in \Lambda$ and $n \geqslant 1$ we have the inequality

$$R_n(W(\Lambda), \Omega_\lambda) \leqslant R_n(\mathfrak{L}^\lambda, \Omega_\lambda) + c(\lambda)/n,$$

where $c(\lambda)$ is independent of n. Thus, $W(\Lambda)$ on each Ω_λ, $\lambda \in \Lambda$, is asymptotically just as efficient as the code that is optimal on Ω_λ. Code $W(\Lambda)$ possesses two additional properties: 1) if for some $\Omega \subset \Omega_\infty$ there exists an SU code, then $W(\Lambda)$ is an SU code; on all Ω_∞ code $W(\Lambda)$ is weakly universal. (Note that no SU code exists on Ω_∞ [3].)

Consider the following example. Assume that $\tilde{\Omega}_0 \subset \Omega_\infty$ is a set of Bernoulli sources, while $\tilde{\Omega}_i \subset \Omega_\infty (i > 0)$ is a set of Markov sources of connectedness (or memory) i. Asymptotically optimal SU codes for $\tilde{\Omega}_0$ [1, 4] and $\tilde{\Omega}_i (i > 0)$ [5, 6] are known. The redundancy of these codes, as a function of the block length n, is

Translated from Problemy Peredachi Informatsii, Vol. 20, No. 3, pp. 24-28, July-September, 1984. Original article submitted October 19, 1982; revision submitted July 25, 1983.

0032-9460/84/2003-0173$08.50

$$(|A|-1)|A|^i \log n/2n+O(1/n) \tag{1}$$

as $n \to \infty$, $i = 0, 1, \ldots$. (Here and henceforth, $|A|$ is the power of A, $\log x \equiv \log_2 x$.) A weakly universal code on Ω_∞ was first constructed in [7]; its redundancy on Ω_i is

$$(|A|-1)|A|^i(\log n)\varphi(n)/n+O(1/n) \tag{2}$$

where $n \to \infty$, $i = 0, 1, \ldots$, $\varphi(n)$ increases without limit as $n \to \infty$. This is also true for other known codes that we weakly universal on Ω_∞ and were given in [3, 8]. As can be seen from (1), the redundancy of a universal code depends significantly on the source memory. However, information on the source memory may not necessarily be known exactly, but may be given, e.g., in the form of the inequality $0 \leqslant i < \gamma$. For $\gamma < \infty$ we can employ a code that is optimal on $\tilde{\Omega}_\gamma$, but on $\tilde{\Omega}_0$ the redundancy may exceed the minimum value by a factor of $|A|^\gamma$. If, however, $\gamma = \infty$, i.e., it is known only that the source memory is finite, then only codes that are weakly universal on Ω_∞ can be employed for data compression. Then, as can be seen from (2) and (1), the redundancy as $n \to \infty$ will have a higher order than that of the optimal code for $\tilde{\Omega}_i$. The redundancy of the code proposed in this paper, as constructed for the family $\{\tilde{\Omega}_i\}$, $i = 0, 1, \ldots$, is not greater than

$$(|A|-1)|A|^i \log n/2n+c(i)/n$$

on each $\tilde{\Omega}_i$, $i = 0, 1, 2, \ldots$; here $c(i)$ is independent of n. Thus, the redundancy of this code coincides asymptotically with (1) — i.e., the redundancy of the optimal code on Ω_i — for all $i = 0, 1, \ldots$. The proposed code is twice universal: it is efficient not only for unknown probability characteristics of the source but also for unknown source memory.

Let us give some exact definitions. For integer $n \geqslant 1$ and source $\omega \in \Omega_\infty$ we denote by $H_n(\omega)$ the n-th approximation to the entropy, and by $H(\omega)$ the entropy of ω. For alphabet A and integer $n > 0$ we denote by A^n the set of all words of length n over alphabet A; we set $A^* = \bigcup_{n=1}^{\infty} A^n$. Assume that Φ_n is the set of all mappings φ: $A^n \to \{0, 1\}^*$ such that $\varphi(x) \neq \varphi(y)$ for $x \neq y$ and $\varphi(A^n)$ is a decodable (separable) code. The sequence of mappings $\{\varphi_n\}$, $n=1, 2, \ldots$, will be called a code if $\varphi_n \in \Phi_n$ for $n = 1, 2, \ldots$. When no confusion results, we will also call φ_n a code. For $\omega \in \Omega_\infty$ and $x \in A^n (n>0)$ we denote by $p_\omega(x)$ the probability with which ω generates x. The cost and redundancy of code $\varphi=\{\varphi_n\}$, $n=1, 2, \ldots$ what we call the quantities

$$c_n(\varphi, \omega) = n^{-1} \sum_{x \in A^n} p_\omega(x) |\varphi_n(x)|, \; r_n(\varphi, \omega) = c_n(\varphi, \omega) - H(\omega) \tag{3}$$

for $n = 1, 2, \ldots$. (For word x, we denote its length by $|x|$.) The redundancy of φ on the set of sources $\Omega \subset \Omega_\infty$ is the quantity

$$R_n(\varphi_n, \Omega) = \sup\{r_n(\varphi, \omega), \omega \in \Omega\} \tag{4}$$

for $n = 1, 2, \ldots$. We also define

$$R_n(\Omega) = \inf\{R_n(\varphi_n, \Omega), \varphi_n \in \Phi_n\}. \tag{5}$$

Code $\varphi=\{\varphi_n\}$, $n=1, 2, \ldots$ will be called optimal on $\Omega \subset \Omega_\infty$ if for all n we have $R_n(\varphi_n, \Omega) = R_n(\Omega)$. We note immediately that an optimal code exists for any $\Omega \subset \Omega_\infty$. By ψ_k we denote the optimal code for set $\tilde{\Omega}_k$ of Markov sources of connectedness k, $k = 0, 1, 2, \ldots$ ($\tilde{\Omega}_0$ are Bernoulli sources). It is known that for any $\omega \in \Omega_\infty$ and any $k = 0, 1, 2, \ldots$ we have the inequality

$$c_n(\psi_k, \omega) \leqslant H_k(\omega) + (|A|-1)|A|^k \log n/n + O(1/n) \tag{6}$$

as $n \to \infty$ [5, 6]. In what follows we will employ the decodable mapping v of the set of natural numbers onto $\{0, 1\}^*$, constructed in [9]. For it we have the following inequality as $i \to \infty$:

$$|v(i)|\leqslant \log i+O(\log\log i). \tag{7}$$

The basic result of this paper is the following theorem.

THEOREM. Assume that A is a finite alphabet; $\{\Omega_i\}$, i = 1, 2, ..., is a finite or countable family of sets of stationary ergodic sources that generate letters from A, such that a strongly universal code exists for every Ω_i. Then there exists a code W with the following properties:

1) on each Ω_i the redundancy of W asymptotically coincides with that of the optimal code on Ω_i, i.e., for all i = 1, 2, ...

$$R_n(W_n, \Omega_i)\leqslant R(\Omega_i)+c(i)/n \tag{8}$$

for n = 1, 2, ..., where c(i) is independent of n;

2) if for some family of sources $\Omega\subset\Omega_\infty$ there exists a strongly universal code, then W is also strongly universal on Ω;

3) code W is weakly universal on Ω_∞.

Remark. In the case of a countable alphabet A, when the remaining conditions of the theorem are satisfied, there exists code W that possesses property 1.

Proof. First let us consider the case of a countable family $\{\Omega_i\}$, i = 1, 2, ... We number the elements of this family by odd numbers, beginning with 1. Assume that, as before, $\tilde{\Omega}_j$, j = 1, 2, ..., is the set of all Markov sources of memory j; $\tilde{\Omega}_0$ is the set of Bernoulli sources. We number them using even numbers such that $\tilde{\Omega}_j$ corresponds to 2j(j = 0, 1, ...). Assume that $\{\Sigma_i\}$ i = 0, 1, ... is the union of families of sources $\{\Omega_i\}$ and $\{\tilde{\Omega}_i\}$ with the above numbering, i.e.,

$$\Sigma_i=\begin{cases}\Omega_{(i+1)/2}, & i=1,3,5,\ldots,\\ \tilde{\Omega}_{i/2}, & i=0,2,4,\ldots\end{cases} \tag{9}$$

Assume that $\mathfrak{L}^i=\{l_n^i\}$, $n=1, 2,\ldots$, is the optimal code on Σ_i for i = 0, 1, 2, ... For each integer $n \geqslant 1$ and $x \in A^n$ we define

$$m(x)=\min\{|l_n^i(x)|+|v(i)|,\quad i=0, 1,\ldots\}, \tag{10}$$

$$k(x)=\min\{j;\ |l_n^j(x)|+|v(j)|=m(x)\}. \tag{11}$$

We define the mapping W_n: $A^n \to \{0, 1\}^*$ as follows:

$$W_n(x)=v(k(x))\,l_n^{k(x)}(x) \tag{12}$$

for all $x \in A^n$. (The right side of (12) contains a concatenation of two code words.) We note immediately that W_n is a decodable mapping, since ν and $\mathfrak{L}_n^{k(x)}$ are decodable. The sequence of mappings $\{W_n\}$, n = 1, 2, ..., will be code W.

Let us prove the first assertion of the theorem. We fix arbitrary $n\geqslant 1$, $i\geqslant 1$ and $x \in A^n$. We have the following chain of inequalities, from which expression (8) follows:

$$R_n(W, \Omega_i)=R_n(W, \Sigma_{2i-1})=\sup\left\{n^{-1}\sum_{x\in A^n} p_\omega(x)\,|W_n(x)|-H(\omega), \omega\in\Sigma_{2i-1}\right\}\leqslant$$

$$\leqslant\sup\left\{n^{-1}\left(\sum_{x\in A^n} p_\omega(x)\,(|l_n^{2i-1}(x)|+|v(2i-1)|)\right)-H(\omega), \omega\in\Sigma_{2i-1}\right\}\leqslant$$

$$\leqslant\sup\left\{n^{-1}\left(\sum_{x\in A^n} p_\omega(x)\,|l_n^{2i-1}(x)|\right)-H(\omega), \omega\in\Sigma_{2i-1}\right\}+$$

$$+|v(2i-1)|\,n^{-1}=R_n(\Sigma_{2i-1})+|v(2i-1)|\,n^{-1}=R_n(\Omega_i)+|v(2i-1)|\,n^{-1}.$$

The first equality follows from (9), the second from definitions (3)-(5); the first inequality follows from (10)-(12), while the second inequality is obvious; and the last two equalities follow from (3)-(5) and (9) respectively.

Let us now prove the second assertion of the theorem. For this we employ the existence criterion for an SU code from [3]: for a set of sources $\Omega\subset\Omega_\infty$ an SU code exists if and only if $H_n(\omega)$ converges to $H(\omega)$ (as $n \to \infty$) uniformly with respect to $\omega\in\Omega$. Assume that $\Omega\subset\Omega_\infty$ and an SU code exists for Ω. We take an arbitrary $\varepsilon > 0$. In view of the existence criterion for an SU code, there exists a k such that for $n > k$

$$\sup\{H_n(\omega)-H(\omega),\ \omega\in\Omega\}<\varepsilon/2. \tag{13}$$

The following chain of inequalities is valid:

$$\begin{aligned}
&R_n(W,\Omega)=\sup\left\{n^{-1}\sum_{x\in A^n}p_\omega(x)\,|w_n(x)|-H(\omega),\ \omega\in\Omega\right\}\leqslant\\
&\leqslant\sup\left\{n^{-1}\sum_{x\in A^n}p_\omega(x)\,(|l_n{}^k(x)|+|v(2k)|)-H_k(\omega),\ \omega\in\Omega\right\}+\\
&+\sup\{H_k(\omega)-H(\omega),\ w\in\Omega\}\leqslant(|A|-1)\,|A|^k\log n/n+\\
&+O(1/n)+|v(2k)|\,n^{-1}+\varepsilon/2\leqslant\varepsilon/2+O(\log n/n).
\end{aligned}$$

The equality follows from from definitions (3)-(5); the first inequality follows from (10)-(12) and a well-known property of the upper bound; the second inequality follows from the fact that $\{l_n{}^k\}$, $n = 1, 2, \ldots$, is the optimal code on $\Sigma_{2i} = \Omega_i$ from (6) and (13); and the last inequality derives from (7). Thus, there exists an $n(\varepsilon)$ such that $R_n(W, \Omega) < \varepsilon$ for $n > n(\varepsilon)$. Since ε is arbitrary, the second property of code W has been proved. The proof of the third assertion is similar to the one just considered (in effect the third assertion is a particular case of the second one for $|\Omega| = 1$).

The case of a finite family can readily be reduced to one considered by adding a countable family of sources to $\{\Omega_i\}$ (say $\{\tilde{\Omega}_i\}$). The theorem has thus been proved.

In concluding we should note that the problem of constructing a twice-universal code, i.e., one that is asymptotically optimal on a countable class of sources, has not been considered earlier. However, the method used above to solve this problem is a familiar one in information theory. It was evidently first employed in [10]. Similar methods have also been employed to construct universal codes: the combinatorial method [7] and the maximum-likelihood method [5]. The latter method was employed in [11] to obtain a code similar to W for the adaptive (with respect to complexity) coding problem. Moreover, a code analogous to W can be constructed by the method of averaging over families of sources [12].

LITERATURE CITED

1. R. E. Krichevskii, "Relationship between redundancy of coding and reliability of information on the source," Probl. Peredachi Inf., 4, No. 3, 48-57 (1968).
2. L. D. Davisson, "Universal noiseless coding," IEEE Trans. Inf. Theory, 19, No. 6, 783-795 (1973).
3. R. E. Krichevskii, "Universal coding and Kolmogorov complexity," in: Proc. Fifth International Symposium on Information Theory, Part 2: Abstracts of Reports [in Russian], Moscow–Tbilisi (1979), pp. 22-25.
4. R. E. Krichevskii, Lectures on Information Theory [in Russian], NGU, Novosibirsk (1970).
5. Yu. M. Shtar'kov, "Coding of messages of finite length at the output of a source with unknown statistics," in: Proc. Fifth All-Union Conference on Coding Theory and Information Transmission, Part 1: Abstracts of Reports [in Russian], Moscow–Gor'kii (1972), pp. 147-152.
6. V. K. Trofimov, "Redundancy of universal coding of arbitrary Markov sources," Probl. Peredachi Inf., 10, No. 4, 16-24 (1974).
7. Yu. M. Shtarkov and V. F. Babkin, "Combinatorial encoding for discrete stationary sources," in: Proc. Second Internat. Symp. Inform. Theory, Tsachkadsor, Armenia, USSR (1971); Budapest: Acad. Kiado (1973), pp. 249-257.
8. L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," IEEE Trans. Inf. Theory, 27, No. 3, 269-279 (1981).
9. V. I. Levenshtein, "Redundancy and delay of divisible coding of natural numbers," in: Problems of Cybernetics [in Russian], No. 20, Nauka, Moscow (1968), pp. 173-179.
10. A. N. Kolmogorov, "Three approaches to the definition of 'quantity of information'," Probl. Peredachi Inf., 1, No. 1, 3-7 (1965).

11. Yu. M. Shtar'kov, "Adaptive coding algorithm for discrete sources," in: Proc. Sixth All-Union Conference on Coding Theory and Information Transmission, Part 1: Abstracts of Reports [in Russian], Moscow–Tomsk (1975), pp. 204-209.
12. V. K. Trofimov, Universal Coding of Markov Sources: dissertation for the degree of Candidate of Physical and Mathematical Sciences [in Russian], In-t Matematiki Sib. Otd. Akad. Nauk SSSR, Novosibirsk (1977).

SIGNAL PROCESSING BY THE NONPARAMETRIC MAXIMUM-LIKELIHOOD METHOD

A. S. Nemirovskii, B. T. Polyak,
and A. B. Tsybakov

UDC 621.391.1:519.27

The authors propose a class of estimates that are a generalization of maximum-likelihood and M-estimates in the nonparamagnetic regression problem. Existence conditions and calculation methods for such estimates are considered, and it is shown that they are consistent.

1. INTRODUCTION

The maximum-likelihood method is usually regarded as a method of estimating finite-dimensional parameters. When generalized to nonparametric (infinite-dimensional) problems, a primary question concerns the form of the maximum-likelihood estimates (MLE), and how to calculate them. It is also of interest to determine the conditions under which nonparamagnetic MLE are consistent, and to what extent they are efficient as compared to other estimates. In this paper, we will consider these issues in relation to the problem of nonparamagnetic regression, i.e., reconstruction of a function on the basis of its values observed with errors. The observations have the form

$$y_i=f^*(x_i)+\xi_i, \quad i=1,\ldots,n, \tag{1.1}$$

where x_i are observation points belonging to some set X; f* is an unknown function, $X \to R^1$, $\xi_i \in R^1$ are independent identically distributed (IID) random errors that have distribution function G. Generally speaking, the x_i are random quantities. Regarding f* it is known only that it belongs to a specified class $\mathcal{F}$.

Consider the following estimate of f*:

$$f_n=\arg\min_{f\in\mathcal{F}}\sum_{i=1}^{n}F(y_i-f(x_i)). \tag{1.2}$$

Here $F: R^1 \to [0, +\infty]$ is some estimation function and $\arg\min_{f\in\mathcal{F}} J(f)$ denotes the solution of the extremal problem

$$J(f)=\sum_{i=1}^{n}F(y_i-f(x_i))\to\min_{f\in\mathcal{F}}. \tag{1.3}$$

If F(y) = −ln(dG(y)/dy), then estimate (1.2) is an MLE. Estimate (1.2) can be interpreted as a generalization of Huber's M-estimates [1].

Let us give some examples of classes $\mathcal{F}$.

1. $\mathcal{F}$ is a parametric family, $\mathcal{F}=\{f: f(x)=g(x, \theta), \theta\in\Theta\}$, where $g(\cdot, \cdot)$ is a known function and $\Theta\subseteq R^N$ is the set of parametric values. Then estimate (1.2) has the form $f_n(x) = g(x, \theta_n)$, where the quantity

Translated from Problemy Peredachi Informatsii, Vol. 20, No. 3, pp. 29-46, July-September, 1984. Original article submitted June 16, 1983; revision submitted January 21, 1984.

0032-9460/84/2003-0177$08.50