KYBERNETIKA — VOLUME 7 (1971), NUMBER 6

The Kiefer-Wolfowitz Approximation Method in Controlled Markov Chains

PETR MANDL

A modification of the Kiefer - Wolfowitz stochastic approximation method is employed to maximize the mean reward per one step from a Markov chain depending on a regression parameter.

Consider a system $\boldsymbol{S}$ from which income is earned at times 1, 2, 3, ... Let S_n denote the state of $\boldsymbol{S}$ at time n. S_n is one of the numbers $1, 2, \ldots, r$. The law of motion of $\boldsymbol{S}$ is the following: For arbitrary $i \in \{1, 2, \ldots, r\} = I$, whenever $\boldsymbol{S}$ is in state i, the probability distribution of the next state is $(p_{i1}(x), \ldots, p_{ir}(x))$ where $x \in (-\infty, \infty)$ is a regression parameter. The income associated with a transition from i into j equals $v_{ij}(x)$. Thus, if X_m denotes the value of the regression parameter during the period $(m, m + 1)$, then the total income earned up to time $n = 1, 2, \ldots$ equals

$$V(n) = \sum_{m=1}^{n} v_{S_{m-1}S_m}(X_{m-1}), \quad V(0) = 0 .$$

The system is specified by matrices

$$P(x) = \|p_{ij}(x)\|_{i,j=1}^{r}, \quad \|v_{ij}(x)\|_{i,j=1}^{r}, \quad x \in (-\infty, \infty) .$$

For fixed regression parameter (i.e. $X_n = x$, $n = 0, 1, \ldots$), $\{S_n, n = 0, 1, \ldots\}$ is a homogeneous Markov chain with transition probability matrix $P(x)$. We introduce the n-step transition probabilities $P(x)^n = \|p_{ij}^{(n)}(x)\|_{i,j=1}^{r}$. The expectation of $V(n)$ for $S_0 = i$ is then given by

$$\mathsf{E}_i^x V(n) = \sum_{m=0}^{n-1} \sum_j \sum_k p_{ij}^{(m)}(x)\, p_{jk}(x)\, v_{jk}(x) .$$

Assumption 1.

1. $|v_{ij}(x)| \leqq K < \infty$, $x \in (-\infty, \infty)$, $i, j \in I$.

2. There exists a positive integer n_0, an $h \in I$ and a number $d > 0$ such that

$$p_{jh}^{(n_0)}(x) \geqq d\,, \quad j = 1, \ldots, r\,, \quad x \in (-\infty, \infty)\,.$$

Under Assumption 1, the limit

$$\Theta(x) = \lim_{n \to \infty} n^{-1} \mathsf{E}_i^x\, V(n)$$

is independent of i. $\Theta(x)$ is the mean income per one period corresponding to regression parameter x. It can also be expressed with aid of recurrence times. Denote by $N(n)$ the n-th recurrence time into h, i.e.

$$N(0) = \inf\{m : S_m = h,\ m \geqq 0\}\,,$$
$$N(n) = \inf\{m : S_m = h,\ m > N(n-1)\}\,, \quad n = 1, 2, \ldots$$

The pairs

$$[V(N(n+1)) - V(N(n)), N(n+1) - N(n)]\,, \quad n = 0, 1, \ldots,$$

are mutually independent, identically distributed as long as x is kept fixed. Using the strong law of large numbers it is not difficult to derive that

$$\Theta(x) = \mathsf{E}_i^x[V(N(n+1)) - V(N(n))]/\mathsf{E}_i^x[N(n+1) - N(n)]\,. \tag{1}$$

We place ourselves in the situation when the dependence of Θ on x is unknown to us and we are looking for a procedure to approximate the value $\hat{x}$ for which $\Theta(x)$ is maximal. (1) implies that we may consider this as a problem of maximizing the ratio of mean values by making independent observations on pairs of random variables. For the mean value of the ratio, i.e.

$$\mathsf{E}_i^x\{[V(N(n+1)) - V(N(n))]/[N(n+1) - N(n)]\}\,,$$

the Kiefer - Wolfowitz stochastic approximation method could be applied directly. Slight modification is necessary in the present case (see Theorem 1). We shall be basing on [1] and make therefore the following assumption:

Assumption 2. $\Theta(x)$ is increasing for $x < \hat{x}$ and decreasing for $x > \hat{x}$. The derivative $\Theta'(x)$ exists and is continuous. For $x \in (-\infty, \infty)$ holds

$$K_0|x - \hat{x}| \leqq \Theta'(x) \leqq K_1|x - \hat{x}| \quad \text{where} \quad 0 < K_0 < K_1 < \infty\,.$$

Description of the procedure. Let $\{a_n, n = 1, 2, \ldots\}$, $\{c_n, n = 1, 2, \ldots\}$ be sequences of positive numbers, $\{M_n, n = 1, 2, \ldots\}$ a sequence of positive integers. Let

$$c_n \to 0\,, \quad \sum_{n=1}^{\infty} a_n = \infty\,, \quad \sum_{n=1}^{\infty} a_n^2 < \infty\,, \quad \sum_{n=1}^{\infty} a_n c_n < \infty\,. \tag{2}$$

where

$$Y_{2n} = \frac{\eta^1_{2n,1} + \eta^1_{2n,2} + \ldots + \eta^1_{2n,M_n}}{\eta^2_{2n,1} + \eta^2_{2n,2} + \ldots + \eta^2_{2n,M_n}}, \quad Y_{2n-1} = \frac{\eta^1_{2n-1,1} + \ldots + \eta^1_{2n-1,M_n}}{\eta^2_{2n-1,1} + \ldots + \eta^2_{2n-1,M_n}},$$

and for given $\eta^1_{1,1}, \eta^2_{1,1}, \ldots, \eta^1_{2n-2,M_{n-1}}, \eta^2_{2n-2,M_{n-1}}$ *the vectors* $(\eta^1_{2n-1,i}, \eta^2_{2n-1,i})$, $(\eta^1_{2n,i}, \eta^2_{2n,i})$ $i = 1, 2, \ldots, M_n$ *are mutually independent with distribution function* $F(y^1, y^2 \mid x_n - c_n)$ *and* $F(y^1, y^2 \mid x_n + c_n)$, *respectively. Then*

$$\lim_{n \to \infty} \mathsf{E}(x_n - \hat{x})^2 = 0 .$$

The demonstration is obtained by inserting appropriate estimates in the proof of Theorem 1 in [1] and will not be given here. Under the assumption $m'''(x) \leqq Q < \infty$ for $x \in (-\infty, \infty)$, it can also be shown by the methods of [1] that for

$$a_n = an^{-1}, \quad c_n = cn^{-1/4}, \quad M_n = [dn^{3/4}] + 1, \quad n = 1, 2, \ldots,$$

where $a > \frac{1}{4}K_0$, $c > 0$, $d > 0$, we get

$$\mathsf{E}(x_n - \hat{x})^2 = O(R_n^{-4/7}) \quad \text{for} \quad n \to \infty .$$

$R_n = 2 \sum_1^n M_m$ is the number of observations employed. The corresponding estimate for the Kiefer - Wolfowitz method is

$$\mathsf{E}(x_n - \hat{x})^2 = O(n^{-2/3}) = O(R_n^{-2/3}) .$$

(Received June 3, 1971.)

REFERENCES

[1] Václav Dupač: O Kiefer - Wolfowitzově aproximační metodě. Časopis pro pěst. mat. *82* (1957), 1, 47—75. (Appeared in Selected Translations in Mathematical Statistics and Probability.)

[2] R. A. Howard: Dynamic Programming and Markov Processes. J. Wiley, New York 1960.

Kieferova - Wolfowitzova aproximační metoda v řízených Markovových řetězcích

PETR MANDL

V práci je modifikace Kieferovy - Wolfowitzovy stochastické aproximační metody použita k maximalizaci průměrného důchodu na jeden krok Markovova řetězce závislého na regresním parametru.

Dr. Petr Mandl, DrSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Vyšehradská 49, Praha 2.