CYBERNETICS AND CONTROL THEORY

V. G. SRAGOVICH, Yu. A. FLEROV

1964

SovietRxiv

Abstract

Full Text

CYBERNETICS AND CONTROL THEORY

V. G. SRAGOVICH, Yu. A. FLEROV

CONSTRUCTION OF A CLASS OF OPTIMAL AUTOMATA

(Presented by Academician A. A. Dorodnitsyn, 7 IV 1964)

Below we propose a construction of a series of automata functioning in the following situation: at the input of the automaton, at moments of time $t = 0, 1, 2, ...$, numbers– "payoffs" $\pi_t \in \Pi$–arrive. This sequence is assumed to be random. In response to each payoff the automaton makes one of n ($\geqslant 2$) moves, and the probability distribution of the input sequence depends on the move it has made. The automaton is capable, without knowing in advance the magnitudes of the payoffs and the probabilities of their occurrence, of self-learning and developing the best behavior, i.e., such a rule for choosing moves that would maximize W, the mean payoff of the automaton. If the set Π (hereafter finite) and the probability distribution of the input quantities vary in time, the automaton must be able to change its behavior sufficiently quickly and economically. This means that the mean time required for it to reach the optimal strategy and the mean cost of learning must be small.

Let us denote the set of the automaton' s proper moves by $X_1, X_2, \dots, X_n$, the probabilities of choosing which at moment t form the vector $\mathbf{p}_t = (p_{1t}, \dots, p_{nt})$. Let W_i be the conditional mean payoff of the automaton when it chooses move X_i; then the total mean payoff will be

$$W(\mathbf{p}) = W(p_1, \dots, p_n) = \sum_{i=1}^{n} W_i p_i,$$

i.e., a multilinear function on the simplex $P = \{p_1, \dots, p_n : p_i \geqslant 0,$

$$\sum_{1}^{n} p_i = 1\}.$$

Consequently, its greatest value

$$\max_{(p)} W(\mathbf{p}) = \max_{(i)} W_i$$

is attained at one of the vertices of P, and if there are several maximal coefficients W_i, then on the corresponding face of P. For what follows it is important only that $W(\mathbf{p})$ is a monotone function. The moves corresponding to the maximal W_i are called optimal.

We now define the set of states of the automaton $S = \{s\}$. Each state s is a vector

$$s = \{\mathbf{p}; \mathbf{V}; \mathbf{N}\},$$

whose components, in turn, are vectors. The first of them, $\mathbf{p}$, denotes the strategy of the automaton; the set of its values is finite: $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(r)}$ $(r \geqslant n+1)$. A description of this set for arbitrary n is cumbersome. We characterize it on the simplest example $n = 2$, when the automaton makes two moves X_1 and X_2 with probabilities respectively p and $1 - p$. However, even in this case the values of p can be specified in many ways, for example $p = \delta, 1/6, 1/3, 1/2, 2/3, 5/6, 1-\delta$ $(0 < \delta < 1/6)$, or $p = \delta, 1/4, 1/2, 3/4, 1 - \delta$ $(0 < \delta < 1/4)$, and so on.

The vector $\mathbf{V} = \{v_{p(k)}\}$, of dimension not less than $n+1$, is formed by empirical mean payoffs of the automaton, which are calculated by it as the numbers π_t arrive at each step of operation. In addition to the mean payoff corresponding to the strategy being used by the automaton at the given moment, the mean payoffs corresponding to the remaining

possible strategies, except for the "extreme" ones (for $n = 2$, $p = \delta$ and $p = 1-\delta$ are excluded). Instead of them, average payoffs are computed corresponding to all pure strategies (for $n = 2$, $p = 0$ and $p = 1$). These procedures are carried out with the aid of a probabilistic mechanism, which selects from the input sequence of numbers $\{\pi_t\}$ subsequences corresponding to possible strategies. For $n = 2$ and $p = \delta$, $1/4$, $1/2$, $3/4$, $1 - \delta$, the vector $\mathbf{V}$ has the form $\mathbf{V} = (v_0, v_{1/4}, v_{1/2}, v_{3/4}, v_1)$. Thus, the ordered set $\{v_{p(k)}\}$ is a discrete statistical analogue of the average payoff $W(\mathbf{p})$. The numbers $v_{p(k)}$ are determined by the formula

$$v_{p(k)} = \frac{1}{N_{p(k)}} \sum_{i=1}^{N_{p(k)}} \pi_{p(k),i},$$

where $\{\pi_{p(k),i}\}$ is a sequence of payoffs of length $N_{p(k)}$ corresponding to the strategy $\mathbf{p}^{(k)}$. The vector $\mathbf{N}$ is defined as the ordered set $\{N_{p(k)}\}$ of the same dimension as $\mathbf{V}$.

The transition function $s_{t+1} = \sigma(s_t, \pi_t)$ has already been partially defined in the preceding paragraph. There it was indicated how the input signals $\{\pi_t\}$ transform the components of the state $\mathbf{V}_t$ and $\mathbf{N}_t$. It remains to specify the rule for changing the strategy $\mathbf{p}_t$. It may take different forms, being based each

time on a monotone dependence of the average payoff W on the strategy $\mathbf{p}$. For $n = 2$, for example, the following monotonicity criteria prove effective.

K_I. If the components of $\mathbf{V}$ strictly increase ($v_0 < v_{1/4} < v_{1/2} < v_{3/4} < v_1$), the probability p_t is shifted one step to the right up to $p = 1 - \delta$. If the components strictly decrease, p_t is shifted one step to the left up to $p = \delta$.

K_{II}. If $v_0 < v_1$ and $v_{1/4} < v_{3/4}$, then p_t is shifted one step to the right; and if $v_0 > v_1$ and $v_{1/4} > v_{3/4}$, then p_t is shifted one step to the left.

The output function of the automaton $X_t = \varphi(s_t)$ is defined, similarly to the transition function, by a probabilistic mechanism. According to the current value of $\mathbf{p}_t$, it selects one of the possible moves. Thus, the constructed automata, which we shall denote by $A(n, r, \delta, K)$, operate according to the scheme

$$\pi_t \to s_{t+1} = \sigma(s_t, \pi_t) \to X_{t+1} = \varphi(\mathbf{p}_{t+1}).$$

We shall say that the **external environment of the automaton is unchanged** if the set Π and the conditional mean payoffs W_i are constant. The automaton operates in a **δ-optimal regime** ($\delta > 0$) if the optimal (for an unchanged external environment) moves are chosen with probability $1 - \delta$, and all the others with total probability δ.

The main properties of the automata $A(n, r, \delta, K)$ are summarized in the following assertions:

Theorem 1. *The automaton $A(n, r, \delta, K)$ is probabilistic Markovian. Its set of states S, in an unchanged external environment, forms a homogeneous Markov chain with a countable set of states.*

Theorem 2. *In an unchanged external environment, for any $\delta > 0$ there exists an almost surely finite random variable τ such that, for $t > \tau$, the automaton $A(n, r, \delta, K)$ will be in a δ-optimal regime.*

The automaton $A(n, r, \delta, K)$ will adapt to changes in the external environment the better, the larger δ is and the more slowly the environment changes in comparison with the speed at which the automaton enters the optimal regime. Such "slow" environments do indeed arise in many problems.

Another approach to the construction of self-learning automata was carried out by M. L. Tsetlin ([1]).

Computing Center
Academy of Sciences of the USSR

Received
4 IV 1964

REFERENCES

1. M. L. Tsetlin, *DAN*, **139**, No. 4 (1961); *UMN*, **18**, 4 (112) (1963).

Note: Figure translations are in progress. See original paper for figures.