Control
and Cybernetics
VOL. 14 (1985) No. 4

Iterative adaptive control of denumerable state average-cost Markov systems[1])

by
ROBERTO S. ACOSTA-ABREU
Departamento de Matematicas
E.S.F.M. del I.P.N.
Apartado Postal 75-702
Mexico, D.F.,C.P. 07300, Mexico

ONESIMO HERNANDEZ-LERMA[2])
Departamento de Matematicas
Centro de Investigacion del I.P.N.
Apartado Postal 14-740
Mexico, D.F.,C.P. 07000, Mexico

In this paper, we consider average-cost denumerable state Markov systems which depend on *unknown* parameters. A nonstationary value-iteration scheme is used to determine an optimal adaptive policy.

Key Words. Average-cost Markov decision processes with unknown parameters, Nonstationary value-iteration, Adaptive policies, Naive feedback controller.

1. Introduction

We consider in this paper the problem of determining optimal adaptive policies for average-cost, denumerable state, Markov decision processes which depend on unknown parameters. To solve this problem we use an iterative adaptive control scheme related to the *nonstationary value-iteration* (NVI) scheme of Federgruen and Schweitzer [4], and which is a variant of the usual method of succesive approximations; see, e.g., Bertsekas [2] or Ross [19]. The NVI scheme is also related to a recursive algorithm proposed by Baranov et al. [1], [20], and has been used by Hernández-Lerma and Marcus [6], [9] for the adaptive control of discounted-reward Markov and semi-

[1]) This research was supported in part by COFAA-IPN, and in part by the Consejo Nacional de Ciencia y Tecnología under Grant PCCBBNA 020630.

[2]) Author for correspondence.

-Markov processes. We briefly compare the NVI adaptive policy with Baranov's scheme and with the "naive feedback controller" studied by Kurano [15], Mandl [17], Kolonko [13], Georgin [5] and others. Before going any further let us introduce the decision models we are concerned with.

The decision model. Consider the Markov decision model (S, A, c, p), where S is the state space, and A is the action (or control) set. The system is observed at epochs $n = 0, 1, 2, \ldots$ We suppose that S is a denumerable set and A is a metric space. For each $x \in S$, let $A(x) \subset A$ denote the (measurable) set of admissible actions in state x, and let $K := \{(x, a): x \in S,\ a \in A(x)\}$. The measurable function $c: K \to R$ denotes the cost function and p the transition law; that is, if the system is in state x and action $a \in A(x)$ is chosen, an immediate (expected) cost $c(x, a)$ is incurred and the next state will be y with probability $p(y|x, a)$.

In our problem the functions c and p depend on an *unknown* parameter θ. No a *priori* information is given about θ, except that it belongs to a parameter set T. The following terminology and notation will help us to state the problem precisely; cf. [5, 6, 9, 14, 15].

Let T be a metric space. For each $\theta \in T$, consider the Markov decision model $(S, A, c(\theta), p(\theta))$ with transition probabilities $p(y|x, a, \theta)$, and cost function $c(x, a, \theta)$, where $(x, a) \in K$, $y \in S$. For each $n = 0, 1, \ldots$, let X_n and A_n be the state and the action at the n-th stage, respectively, and let

$$I_n = (X_0, A_0, \ldots, X_{n-1}, A_{n-1}, X_n)$$

be the history of the process -or information vector- up to time n. I_n is a random vector with values in H_n, where $H_0 = S$, and $H_{n+1} = K \times H_n$, $n = 0, 1, \ldots$. A policy is then defined as a equence $D = (D_n, n = 0, 1, \ldots)$ of measurable functions (possibly randomized [10, 22]) such that, for each n, D_n specifies which action to choose at the n-th decision epoch, given I_n. If the D_n are independent of the history of the system except for the present state X_n, the policy D is said to be *Markovian* or *memoryless*. A memoryless policy that regardless of time, always chooses the same action, say $f(x)$, whenever the system is in state x, is called *stationary*.

Throughout the following the cost function is assumed to be bounded:

ASSUMPTION 1.1. $\text{Sup}\{|c(x, a, \theta)|: (x, a) \in K, \theta \in T\} \leqslant M$, for some $M > 0$.

Now, for each policy D, $x \in S$ and $\theta \in T$ define

$$J_n(D, \theta, x) = E_x^{D,\theta} \sum_{k=0}^{n} c(X_k, A_k, \theta),\ n \leqslant 0, \tag{1}$$

and let [5, 14, 15, 17, ...]

$$J(D, \theta, x) = \lim_{n \to \infty} \inf \frac{1}{n+1} J_n(D, \theta, x) \tag{2}$$

be the *long-run expected average cost* per unit time when policy D is employed and the initial state is x, given that θ is the true parameter value. The limit in (2) exists, because of Assumption 1.1.

DEFINITION 1.2. A policy $\hat{D}$ is (average cost) *optimal* (when θ is the true parameter value) if

$$J(\hat{D}, \theta, x) = \inf_{D} J(D, \theta, x).$$

for each $x \in S$.

Statement of the problem. In this paper, the true parameter value is unknown and we can state our problem as follows: Given that the true parameter value, say $\theta \in T$, is assumed to be constant but unknown, find: (A) an average-cost optimal (adaptive) policy $\hat{D}$, and (B) an iterative procedure to determine $J(\hat{D}, \theta, x)$.

We give a solution to problems (A) and (B) in theorem 3.2 and corollary 3.3, respectively. Related results on the adaptive control of Markov and semi-Markov processes are briefly discussed in section 4. First we introduce in section 2 some preliminary notions.

2. Preliminaries

In addition to Assumption 1.1 we will assume throughout this paper the following conditions.

ASSUMPTION 2.1. (a) S is denumerable; A and T are metric spaces, and for each $x \in S$, $A(x)$ is compact.

(b) For all $x, y \in S$ the functions $(a, \theta) \to c(x, a, \theta)$ and $(a, \theta) \to p(y|x, a, \theta)$ satisfy that, if $(a', \theta') \to (a, \theta)$, then

$$\sup_{x \in S} |c(x, a', \theta') - c(x, a, \theta)| \to 0,$$

and

$$\sup_{x \in S} \sum_{y \in S} |p(y|x, a', \theta') - p(y|x, a, \theta)| \to 0.$$

Following a standard convention we identify the set of stationary policies with the set F of all functions $f: S \to A$ such that $f(x) \in A(x)$, for every $x \in S$, i.e., F is the product space $F = \prod_{x \in S} A(x)$, and often we refer to $D = (f, f, \ldots)$ as the stationary policy $f \in F$. For each stationary policy $f \in F$, and for each $\theta \in T$, let $P(f, \theta)$ be the stochastic matrix whose (x, y) element is $p(y|x, f(x), \theta)$. Under a stationary policy $f \in F$, the process $\{X_n, n \geq 0\}$ is a Markov chain with transition matrix $P(f, \theta)$.

In order to assure existence of solutions for the optimality equation ((3) below) we need the following (so-called strong scrambling) condition:

ASSUMPTION 2.2. There is a number $\varrho > 0$ such that

$$\sum_{y \in S} \min \left[p\left(y|x_1, f(x_1), \theta\right), p\left(y|x_2, f(x_2), \theta\right) \right] \geqslant \varrho$$

for all $x_1, x_2 \in S$, $\theta \in T$ and $f \in F$.

REMARKS 2.3. (a) Observe that under the above assumption any matrix $P(f, \theta)$, $f \in F$ is both aperiodic and indecomposable [3, 23].

(b) A sufficient condition for assumption 2.2 is clearly the following [23]

ASSUMPTION 2.2′. There is a state $s \in S$ and a number $\alpha > 0$ such that $p\left(s|x, f(x), \theta\right) \geqslant \alpha$, for all $x \in S$, $\theta \in T$ and $f \in F$.

(c) When S is a *finite* set, we can use Corollary 6.20 in [19] to see that another sufficient condition to have solutions of equation (3) is the following:

ASSUMPTION 2.2″. For each $\theta \in T$ and each stationary policy f, the matrix $P(f, \theta)$ is irreducible.

We have the following lemma [3, 19, 23].

LEMMA 2.4. *Under Assumptions* 2.2 (*or* 2.2′) *for each* $\theta \in T$ *there exist a constant* $k(\theta)$ *and a function* $v(x, \theta)$, $x \in S$, *such that*

$$k(\theta) + v(x, \theta) = \min_{a \in A(x)} \left\{ c(x, a, \theta) + \sum_{y \in S} p\left(y|x, a, \theta\right) v(y, \theta) \right\} \tag{3}$$

for each $x \in S$. *The function* $v(\cdot, \theta)$ *is bounded and* $k(\theta)$ *is uniquely determined by* (3). *Furthermore, for* $\theta \in T$, *let* $f^*(x, \theta)$ *be such that minimizes the right-hand side of* (3) *for each* $x \in S$. *Then* $D^* = \left(f^*(\cdot, \theta)\right)$ *satisfies that*

$$J(D^*, \theta, x) = \inf_D J(D, \theta, x)$$

for each $x \in S$, *and we also have that, independently of* $x \in S$,

$$k(\theta) = \inf_D J(D, \theta, x) = \lim_{n \to \infty} \frac{1}{n+1} J_n(D^*, \theta, x).$$

REMARK 2.5. Lemma 2.4 says that $D^* = \left(f^*(\cdot, \theta)\right)$ is an (average-cost) optimal policy, and we also have [3, 19] that under assumption 2.2 (or 2.2′), for every $f \in F$ and every $\theta \in T$

$$\lim_{n \to \infty} \frac{1}{n+1} J_n(f, \theta, x) = \sum_{y \in S} c\left(y, f(y), \theta\right) q_y(f, \theta), \; x \in S, \tag{4}$$

where $\{q_y(f, \theta), y \in S\}$ is the unique stationary probability of $P(f, \theta)$.

Equation (3) is called the (average cost) *optimality equation*, and for $\theta \in T$, if $k(\theta), v(x, \theta), x \in S$, satisfy (3), then they are called a solution of the optimality equation (3).

To prove the optimality of an adaptive policy, the parameter estimation scheme is required to be sufficiently "robust" — in the sense of the following assumption.

ASSUMPTION 2.6. For any $\theta \in T$, any $x \in S$ and any policy D, there exists a sequence $\{\hat{\theta}_n\}$ of measurable functions $\hat{\theta}_n: H_n \to T$ such that θ_n converges to θ $P_x^{D,\theta}$-almost surely as $n \to \infty$. The sequence $\{\hat{\theta}_n\}$ is said to be a sequence of *strongly consistent* (SC) estimators of θ.

REMARK 2.7. Examples of *SC* estimators are well known in the literature on adaptive control of Markov and semi-Markov processes [5, 13, 15, 17]. They have been obtained for quite general parameter spaces T using maximum likelihood or minimum contrast estimation, but in special situations — for instance, in the adaptive control of queues [7, 8, 12, 14, 22] — very often it is possible to get *SC* estimators in some elementary way, like (e.g.) moment estimation.

3. The NVI adaptive policy

In this section we consider the case in which the true parameter value, say θ, is unknown and proceed to solve the problems (A) and (B) stated in section 1.

Let us define the functions $v_n: S \times T \to R$ by

$$v_0(x, \theta) := \min_{a \in A(x)} c(x, a, \theta) \tag{5}$$

and for $n = 1, 2, \ldots$,

$$v_n(x, \theta) := Q_n v_{n-1}(x, \theta) \tag{6}$$

where

$$(Q_n u)(x) = \min_{a \in A(x)} \{c(x, a, \theta) + \sum_{y \in S} p(y|x, a, \theta) u(y)\} \tag{7}$$

for any bounded function u on S.

LEMMA 3.0. *Suppose that assumptions* 1.1 *and* 2.1 *hold. Then, for each* $x \in S$, *the functions* $v_n(x, \cdot): T \to R$, $n \geq 0$, *are continuous.*

Proof. The proof is easily obtained by induction, using lemma 4.2 (c) in [14], and lemma 2.3 in [3].

We now use (5)–17) to define the NVI adaptive policy. Let us define a sequence of functions $f_n: S \times T \to A$ by:

$$f_0(x, \theta) = \arg \min_{a \in A(x)} c(x, a, \theta) \tag{8}$$

and for $n = 1, 2, \ldots$

$$f_n(x, \theta) = \arg \min_{a \in A(x)} \{c(x, a, \theta) + \sum_{y \in S} p(y|x, a, \theta)\, v_{n-1}(y, \theta)\}. \tag{9}$$

We can use the conditions given in, e.g., the measurable selection Theorem 12.1 in [21] to assure the existence, for each n, of a measurable function $\theta \to f_n(x, \theta)$ on T, $x \in S$, that satisfies (9).

DEFINITION 3.1. Let $(\hat{\theta}_n)$ be a sequence of SC estimators of θ, and let f_n be the functions defined by (8), (9). The policy $\hat{D} = (\hat{D}_n)$ defined by $\hat{D}_n(I_n) = f_n(X_n, \hat{\theta}_n)$, $n = 0, 1, \ldots$ is called an *NVI adaptive policy*.

THEOREM 3.2. *If assumptions* 1.1, 2.1, 2.2 *and* 2.6 *hold and* $v(x, \theta)$ *is bounded on* $S \times T$, *then the NVI adaptive policy* $\hat{D}$ *is average-cost optimal.*

Proof. Fix $\theta \in T$. For each $(x, a) \in K$, let us define the function

$$\phi(x, a, \theta) = c(x, a, \theta) + \sum_{y \in S} p(y|x, a, \theta)\, v(y, \theta) - v(x, \theta) - k(\theta),$$

where $\{k(\theta), v(x, \theta), x \in S\}$ is a solution of the optimality equation (3). In view of Assumptions 1.1, 2.1, Lemma 2.4 and equation (4), the function $\phi(x, a, \theta)$ is bounded. Let us observe that $\phi(x, a, \theta)$ satisfies

$$\phi(X_k, A_k, \theta) = E_x^{D,\theta}[c(X_k, A_k, \theta) + v(X_{k+1}, \theta) - v(X_k, \theta)|I_k, A_k] - k(\theta)$$

for each x, θ, D and each $k = 0, 1, \ldots$. Therefore

$$\frac{1}{n+1} E_x^{D,\theta} \sum_{k=0}^{n} \phi(X_k, A_k, \theta) = \frac{1}{n+1} E_x^{D,\theta}\Big[\sum_{k=0}^{n} c(X_k, A_k, \theta)\Big] - k(\theta) + \frac{1}{n+1} E_x^{D,\theta}[v(X_n, \theta) - v(x, \theta)]$$

From this equation and Lemma 2.4 we see that in order to prove that $\hat{D}$ is average-cost optimal it is sufficient to prove that

$$\lim_{n \to \infty} \inf \frac{1}{n+1} E_x^{\hat{D},\theta} \sum_{k=0}^{n} \phi(X_k, A_k, \theta) = 0. \tag{10}$$

To do this, let (θ_n) be any sequence in T such that $\theta_n \to \theta$, and let $x_0 \in S$ be any fixed state. Using the results on page 365 of [3], there is a number $0 < \varrho < 1$ such that

$$|v_n(x, \theta) - nk(\theta) - v(x, \theta)| \leqslant (1 - \varrho)^n \cdot \text{constant}, \text{ for all } n \geqslant 1,\ x \in S,\ \theta \in T.$$

From this and Lemma 3.0, we then have that the sequences

$$w_n(x,\theta) := v_n(x,\theta) - v_n(x_0,\theta),\ x\in S,\ \theta\in T,\ n\geqslant 1,$$

and

$$k_n(\theta) := v_n(x_0,\theta) - v_{n-1}(x_0,\theta),\ \theta\in T,\ n\geqslant 1,$$

are continuous in θ and converge uniformly in x and θ to a solution of equation (3). Therefore we have, for all $x\in S$,

$$\lim_{n\to\infty} w_n(x,\theta_n) = v(x,\theta) - v(x_0,\theta) \tag{11}$$

and

$$\lim_{n\to\infty} k_n(\theta_n) = k(\theta). \tag{12}$$

Note that $v(x,\cdot)$ and $k(\cdot)$ are continuous on T. Now, making use of equations (5)–(9) we have.

$$\begin{aligned}\phi(x, f_n(x,\theta_n),\theta) = \phi(x, f_n(x,\theta_n),\theta) - w_n(x,\theta_n) + w_n(x,\theta_n) = \\ = c(x, f_n(x,\theta_n),\theta) - c(x, f_n(x,\theta_n),\theta_n) + \\ + \sum_{y\in S}\left[p(y|x, f_n(x,\theta_n),\theta)\,(v(y,\theta) - v(x_0,\theta)) - \right. \\ \left. - p(y|x, f_n(x,\theta_n),\theta_n)\, w_{n-1}(y,\theta_n)\right] + \\ + w_n(x,\theta_n) - (v(x,\theta) - v(x_0,\theta)) + k_n(\theta_n) - k(\theta).\end{aligned}$$

From this equation we see that, in view of Assumption 2.1 and equations (11), (12),

$$\sup_{x\in S} |\phi(x, f_n(x,\theta_n),\theta)| \to 0, \quad \text{as } n\to\infty,$$

for every sequence $\theta_n\to\theta$. Therefore, under Assumption 2.6

$$|\phi(X_n, A_n,\theta)| = |\phi(X_n, f_n(X_n,\hat{\theta}_n),\theta)| \to 0, \quad P_x^{\hat{D},\theta}\text{-almost surely, as } n\to\infty.$$

Finally, by the bounded convergence theorem, we see that for all $x\in S$, for all $x\in S$,

$$E_x^{\hat{D},\theta}\,\phi(X_n, A_n,\theta)\to 0 \quad \text{as } n\to\infty,$$

from which (10) follows immediately. This completes the proof of Theorem 3.2. ■

The NVI adaptive policy is average-cost optimal and gives us a solution to problem (A) in section 1. A solution to problem (B) in section 1 is now immediately obtained:

COROLLARY 3.3. *Suppose that the assumptions of Theorem* 3.3 *hold and let* $\hat{D}$ *be the (optimal) NVI policy. Let* $x_0\in S$ *be any fixed state and let* (θ_n) *be any sequence in* T *converging to* θ. *Then the sequence*

$$k_n(\theta_n) = v_n(x_0, \theta_n) - v_{n-1}(x_0, \theta_n), \quad n \geq 1,$$

satisfies

$$\lim_{n\to\infty} k_n(\theta_n) = k(\theta) = J(\hat{D}, \theta, x), \quad \text{for all } x \in S. \tag{13}$$

Proof. Since $\hat{D}$ is an optimal policy, it follows from Lemma 2.4, that $k(\theta) = J(\hat{D}, \theta, x)$ for each $x \in S$. Then the two equalities in (13) are a direct consequence of equations (12) and (10) in the proof of Theorem 3.2.

4. Related results and conclusions

Let us compare the NVI adaptive policy with the method which Mandl [18] called the "method of substituting the estimates into optimal stationary controls", and which — except for small variations — is also found in the stochastic control theory literature under the names of "certainty equivalence controller" or — "naive feedback controller (NFC)" (see e.g. [2]). The NFC adaptive policy is constructed as follows.

For each $\theta \in T$, determine an optimal stationary policy $g(\cdot, \theta) \in F$ (cf. Lemma 2.4), and let $(\hat{\theta}_n)$ be a sequence of SC estimators of θ. Then the policy $D' = (D'_n)$ defined by

$$D'_n(I_n) = g(X_n, \hat{\theta}_n), \quad n = 0, 1, \ldots,$$

is called an *NFC adaptive policy*.

The NFC policy D' was introduced, independently, by Kurano [15] and Mandl [17] for the case of finite-state Markov decision processes with average-cost criterion. It was generalized to Markov and semi-Markov decision processes with denumerable state space by Mandl [18] and Kolonko [14], respectively, and it has been applied to the optimal control of queues with unknown parameters by Kolonko [14], Kolonko and Schäl [12] and by Hernández-Lerma and Marcus [7, 8]. Schäl [22] has considered the NFC policy (called in [22] "principle of estimation and control") for *discounted* Markov decision processes. In comparison with the NVI — policy $\hat{D}$ (Definition 3.1), the main disadvantage of the NFC policy is that D' requires to know in advance the optimal policies $g(\cdot, \theta)$ for all values of θ, which is not the case for the NVI policy. That is, at the n-th decision epoch, the NVI controller observes the state x_n, computes the estimate θ_n and chooses the action $\hat{D}_n(I_n) = f_n(x_n, \theta_n)$ using (9). The NFC controller acts similarly: observes x_n, computes θ_n and chooses the action $D'_n(I_n) = g(x_n, \theta_n)$, but here he is assuming that he knows in advance the functions $g(\cdot, \theta) \in F$ for all $\theta \in T$.

The approach used by Baranov [1] and Salyga et al. [20] for *finite* state average-cost Markov chains is similar to our NVI, except that they use a particular parameter-estimation scheme — (related to Ljung's [16] prediction

error method) that seems to be more cumbersome than what is really needed, namely, *any* estimation scheme which gives SC estimates.

In summary, we have introduced in this paper an optimal adaptive policy for the control of Markov processes with unknown parameters that uses any estimation scheme which gives SC estimates. The NVI policy permits to control optimally the system (within the information at hand, as given by the estimators) as early as the second stage.

Acknowledgment. We would like to thank the referee for his many useful comments.

References

[1] BARANOV V. V. Recursive algorithms of adaptive control in stochastic systems. *Cybernetics*, **17** (1981) 815–824.

[2] BERTSEKAS D. Dynamic Programming and Stochastic Control. New York, Academic Press, 1976.

[3] FEDERGRUEN A., TIJMS H. C. The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms. *J. Appl. Probab.* **15** (1978) 356–373.

[4] FEDERGRUEN A., SCHWEITZER P. J. Nonstationary Markov decision problems with converging parameters. *J. Optim. Theory Appl.* **34** (1981) 207–241.

[5] GEORGIN J. P. Estimation et controle des chaines de Markov sur des espaces arbitraires. Lecture Notes Math. 636. Berlin, Springer-Verlag, 1978, 71–113.

[6] HERNÁNDEZ-LERMA O. Nonstationary value-iteration and adaptive control of discounted semi-Markov processes. *J. Math. Anal. Appl.*, 1985. To appear.

[7] HERNÁNDEZ-LERMA O., MARCUS S. I. Adaptive control of service in queueing systems. *Systems and Control Letters*, **3** (1983), 283–289.

[8] HERNÁNDEZ-LERMA O., MARCUS S. I. Optimal adaptive control of priority assignment in queueing systems. *Systems and Control Letters*, **4** (1984) 65–72.

[9] HERNÁNDEZ-LERMA O., MARCUS S. I. Adaptive control of discounted Markov decision chains. *J. Optim. Theory Appl.* To appear in Vol. 46, No. 2, June 1985.

[10] HINDERER K. Foundations of non-stationary dynamic programming with discrete time parameter. Lecture Notes in Operations Research, 33. New York, Springer-Verlag, 1970.

[11] HORDIJK A. Dynamic Programming and Potential Theory. Mathematical Centre Tract, 51. Amsterdam, Mathematisch Centrum, 1974.

[12] KOLONKO M., SCHAL M. Optimal control of semi-Markov chains under uncertainty with applications to queueing models. Proceedings in Operations Research, 9. Würzburg-Wien, Physica-Verlag, 1980, 430–435.

[13] KOLONKO M. Strongly consistent estimation in a controlled Markov renewal model. *J. Appl. Prob.* **19** (1982), 532–545.

[14] KOLONKO M. The average-optimal adaptive control of a Markov renewal model in presence of an unknown parameter. *Math. Operations. u. Statist.*, Serie Opt., **13** (1982), 567–591.

[15] KURANO M. Discrete-time markovian decision processes with an unknown parameter: average return criterion. *J. Oper. Res. Soc. Japan*, **15** (1972), 67–76.

[16] LJUNG L. Convergence analysis of parametric identification methods. *IEEE Trans. Automatic Cont.* **AC-23** (1978), 770–783.
[17] MANDL P. Estimation and control of Markov chains. *Adv. Appl. Prob.* **6** (1974), 40–60.
[18] MANDL P. On the adaptive control of countable Markov chains. In: Probability Theory, Banach Centre Publications, Warsaw, PWN-Polish Scientific Publishers, 1979, 159–173.
[19] ROSS S. M. Applied Probability Models with Optimization Applications, San Francisco, Holden-Day, 1970.
[20] SALYGA V. I., BARANOV V. V., SALYGA V. I., KUZMINA O.I. Methods of markovian decision processes in the problem of identification and adaptive control of stochastic systems. Proc. 6th IFAC Symp., Arlington, Virginia (1982), 566–569.
[21] SCHÄL M. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Zeit. für Wahrs. V. Geb.* **32** (1975), 179–196.
[22] SCHÄL M. Estimation and control in discounted stochastic dynamic programming. Univ. Bonn, Inst. Appl. Math., Preprint 428 (1981).
[23] THOMAS L. C. Connectedness conditions for denumerable state Markov decision processes, in: Recent Developments in Markov Decision Processes, edited by R. Hartley, L. C. Thomas and D. J. White. London, Academic Press, 1980, 181–204.

Received, October 1984

Iteracyjna procedura sterowania adaptacyjnego systemów Markowa z przeliczalną liczbą stanów i średnim kosztem

W pracy rozważono dyskretne systemy Markowa z przeliczalną liczbą stanów i nieskończonym horyzontem sterowania. Przyjęto, że koszt przejścia ze stanu do stanu i prawdopodobieństwo przejścia zależą od nieznanych parametrów. Jako kryteria sterowania przyjęto minimalizację uśrednionej po czasie sumy oczekiwanych kosztów. Jako rozwiązanie otrzymano optymalną strategię adaptacyjną w postaci niestacjonarnej procedury iteracyjnej.

Адаптативное, итерационное управление марковским процессом со средними издержками и счётным числом состоянии

В работе рассмотрено макровскую управлемую цепь со счётным числом состояний. Издержки за один шаг и вероятности перехода зависят от неизвестных параметров. Рассматривается бесконечный горизонт планирования. Надо минимизировать усредненную по времени сумму ожидаемых издержек. Для решения задачи применено алгоритм типа нестационарной итерации стратегии.