Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

SERDICA — Bulgaricae
mathematicae publicationes
20 (1994) 5-16

AN APPROACH TO DECISION MARKOV CHAINS

B. D. DOYTCHINOV, T. I. DOITCHINOVA

ABSTRACT. Decision Markov Chains with total sum criterion are explored. A new approach is proposed. It is proved that under certain conditions stationary Markov strategies are sufficient for the optimal control of the process. The existence of an optimal strategy is proved in a problem of constrained extremum.

Introduction. This paper deals with Decision Markov Chains with total sum criterion. We show that stationary Markov strategies are sufficient for the optimal control of the process under certain conditions. To prove that, we use a new and simple approach, similar to the method used by N.V.Krylov [1] for controlled diffusion processes.

In [1] N. V. Krylov considers a diffusion process described by the stochastic differential equation

$$dx_t = \sigma(x_t, \alpha_t)dw_t + b(x_t, \alpha_t)dt, \quad t \geq 0, \; x_0 = x$$

where w_t is a multidimensional Brownian Motion and α_t is a stochastic process which takes values in a set of actions A and which is used to implement the control. Under appropriate regularity conditions, N. V. Krylov proves the existence of an optimal Markov process. The proof is very elegant and simple, because it is based on using the Green measure of the process and does not use the technique of dynamic programming at all. Theorem 1.2 of [1] is fundamental for the method and appears somewhat unexpected; it says that for every strategy (i.e. possibly non-Markov) one can find more or less explicitly a Markov strategy with the same Green's measure.

We were inspired for our research by the note in [1] that the same construction should work in the case of Decision Markov Chains. The statement of our Theorem 1 could be considered as an analog of Krylov's Theorem 1.2 in [1]. We give a direct proof for Decision Markov Chains, not trying to emulate Krylov's proof, although the reader could find a parallel between our Lemma 1 and Green's measure.

In the literature on Decision Markov Chains other problems are often considered as well: optimal control with average criterion, the optimal stopping time problem, problems with finite horizon, etc. We do not deal with these questions in our paper at all. The reader is referred to the monographs [2] and [3], in which similar and other problems are considered. They contain a very good overview of the subject as well as an extensive bibliography.

Upon examination of the existing results, one should notice that the major part of proofs of existence of optimal Markov strategies relies heavily on Bellman's principle. We believe that our approach is an interesting alternative, because it makes it possible to prove similar results without using Bellman's principle, and, in fact, without using any deep technique of stochastic processes.

The paper consists of four parts. The first part contains some definitions and descriptions. The main result is given in the second part. Some straightforward corollaries (both new results and simpler proofs for known facts) are deduced in the third part. In the last part the main result is applied to prove the existence of an optimal strategy for a problem of constrained extremum.

1. Some preliminaries. We consider a discrete time decision Markov process, having a countable set of states X, a countable set of actions A, and transition probabilities $p_a(i,j)$, $p_\alpha(i,j) \geq 0$, $\sum\limits_j p_a(i,j) = 1$, $a \in A$, $i,j \in X$.

Denote $H_n = (X \times A)^n \times X$, $n = 0,1,\ldots$. A strategy π is a sequence $(\pi_0, \pi_1, \ldots, \pi_n, \ldots)$ of functions $\pi_n : A \times H_n \to [0,1]$, such that $\forall n$ and $\forall h_n \in H_n$ we have $\sum\limits_a \pi_n(a|h_n) = 1$.

Informally, the dynamic of a Decision Markov Chain can be described as follows.

A function $r : X \times A \to R$ is given, called the return function.

At each moment n, the Chain is at a state $x_n \in X$. We observe this state and, knowing the history $h_n = (x_0, a_0, \ldots, a_{n-1}, x_n)$ of the process up to this moment, we choose an action $a_n \in A$. As a result, two things happen. First, we immediately gain a (possibly negative) return $r(x_n, a_n)$. Second, the Chain moves to another state $x_{n+1} \in X$, with probability $\rho_{a_n}(x_n, x_{n+1})$.

To use a strategy $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots)$ means that at every moment n we choose to apply the action $a \in A$ with probability $\pi(a|h_n)$, where $h_n = (x_0, a_0, \ldots, a_{n-1}, x_n)$ is the history of the process up to moment n.

We can choose which strategy to use and the goal is to find the strategy which maximizes the expectation of the total income

$$\sum_{n=0}^{\infty} r(x_n, a_n).$$

A slight generalization is to try to maximize the expectation of the expression

$$\sum_{n=0}^{\tau} \beta^n r(x_n, a_n),$$

where τ is a stopping time and $0 \leq \beta \leq 1$. The number β is called a discount factor.

To make things precise, let us introduce some definitions and notations.

Let $\Omega = (X \times A)^\infty$ and let $\mathcal{F}$ be the σ-algebra in Ω generated by the rectangular subsets of Ω, i.e. the subsets of the form $\Pi_{n=1}^{\infty}(X_n \times A_n)$, where $\forall n : X_n \subseteq X$, $A_n \subseteq A$ and $\exists n_0 \geq n \; \forall n > n_0 : X_n = X$, $A_n = A$.

Given a strategy $\pi = (\pi_0, \pi_1, \ldots, \pi_n, \ldots)$ we can define for every $n \geq 0$ a transition probability $\mathrm{P}^{\pi,n+1}$ by the formula

$$\mathrm{P}^{\pi,n+1}(C|h_n) = \sum_{(a,x)\in C} \pi(a|h_n) p_a(x, x)$$

for all $C \subseteq A \times X$, $h_n = (x_0, \ldots, a_{n-1}, x_n) \in H_n$. Then, by the Theorem of Ionescu-Tulcea (see e.g. [4]), for every initial state $x \in A$ there exists a unique probability P_x^π on $(\Omega, \mathcal{F})$ whose value for every rectangular set $\prod_{n=1}^{\infty}(X_n \times A_n)$ is given by

$$\mathrm{P}_x^\pi \left(\prod_{n=0}^{\infty} (X_n \times A_n) \right)$$

$$= \sum \delta_{xx_0} \pi_0(a_0|x_0) \rho_{a_0}(x_0, x_1) \pi_1(a_1|x_0, a_0, x_1) \rho_{a_1}(x_1, x_2) \ldots \pi_{n_0}(a_1|x_0, a_0, \ldots, x_{n_0}),$$

where δ_{ij} is the Kronecker's δ-symbol, n_0 is large enough so that $\forall n > n_0 : X_n = X$, $A_n = A$, and the summation is over all $a_0 \in A_0$, $a_1 \in A_1, \ldots, a_{n_0} \in A_{n_0}$, and all $x_0 \in X_0$, $x_1 \in X_n, \ldots, x_{n_0} \in X_{n_0}$.

The expectation associated with the probability measure P_x^π on $(\Omega, \mathcal{F})$ will be denoted by E_x^π.

Now let us introduce some classes of strategies which we will consider in this paper.

We will denote the set of all possible strategies by Π.

A strategy π is said to be a randomized Markov strategy, if

$$\forall n \forall h_n = (x_0, \ldots, a_{n-1}, x_n) \in H : \pi_n(a|h_n) = \pi(a|x_n).$$

Two narrower classes of strategies can be defined as follows.

A randomized Markov strategy π is called to be stationary if $\forall n \forall a \forall x : \pi_n(a|x) = \pi(a|x)$, i.e. it does not depend on n. We will denote the class of stationary Markov strategies by $\mathcal{RM}$.

Another subclass of the randomized Markov strategies is the class of all pure (non-randomized, deterministic) Markov strategies. Formally, a randomized Markov strategy is pure if $\forall n \forall x \in X \exists a \in A : \pi_n(a|x) = 1$. We will denote the set of all pure Markov strategies by $\mathcal{M}$.

Let D be a subset of X, and let τ be the first moment for the process to exit D, i.e.

$$\tau = \begin{cases} \inf\{n > 0 : x_n \notin D\} & \text{if } \{n > 0 : x_n \notin D\} \neq \emptyset \\ \infty & \text{if } \{n > 0 : x \notin D\} = \emptyset. \end{cases} \tag{1}$$

Throughout this paper we assume that

$$T = \sup_{\substack{x \in X \\ \pi \in \Pi}} \mathrm{E}_x^\pi \tau < \infty. \tag{2}$$

For every bounded function $r(x, a)$ and every number $\beta \in [0, 1]$ we denote $\mathrm{E}_x^\pi \sum_{n=0}^{\tau-1} \beta^n r(x_n, a_n)$ by $(R_\beta^\pi(D)r)(x)$ or simply by $R_\beta^\pi(D)r(x)$.

We will use the letter χ to denote the indicator functions, e.g.:

$$\chi_{i,b}(x, a) = \begin{cases} 1, & \text{if } x = i, a = b \\ 0, & \text{otherwise} \end{cases}, \quad \chi_i(x) = \begin{cases} 1, & \text{if } x = i \\ 0, & \text{otherwise.} \end{cases}$$

2. The main result.

Theorem 1. *Let a subset $D \subseteq X$ and an initial state $x_0 \in D$ be given. Let τ be the exit time defined by (1) and assume that (2) is satisfied. Then for each strategy $\pi \in \Pi$ there exists a $\widetilde{\pi} \in \mathcal{RM}$, such that for every bounded function $r(x, a)$*

$$R_1^\pi(D)r(x_0) = R_1^{\widetilde{\pi}}(D)r(x_0).$$

The strategy $\widetilde{\pi}$ can be determined by the formula:

$$\widetilde{\pi}(a|i) = (R_1^\pi(D)\chi_{i,a})(x_0) \cdot ((R_1^\pi(D)\chi_i)(x_0))^{-1}.$$

To prove this theorem we need the following two lemmas.

Lemma 1. *For all bounded functions $f(i,l)$, $i \in X$, $l \in A$ consider the transformation $\tilde{f}(i) = \sum_{l \in A} f(i,l)\tilde{\pi}(l|i)$, where $\tilde{\pi}$ is defined by the expression given in Theorem 1. Then for all $x_0 \in X$*

$$R_1^{\pi}(D)f(x_0) = R_1^{\pi}(D)\tilde{f}(x_0).$$

Proof. If $x_0 \notin D$, then $R_1^{\pi}(D)f(x_0) = R_1^{\pi}(D)\tilde{f}(x_0) = 0$. If $x_0 \in D$, then

$$R_1^{\pi}(D)f(x_0) = \mathrm{E}_{x_0}^{\pi} \sum_{n=0}^{\infty} \sum_{i \in D} \sum_{l \in A} f(i,l)\chi\{x_n = i, a_n = l, n < \tau\}$$

$$= \sum_{i \in D} \sum_{l \in A} f(i,l) R_1^{\pi}(D)\chi_{i,a}(x_0) = \sum_{i \in D} \sum_{l \in A} f(i,l)\tilde{\pi}(l|i) R_1^{\pi}(D)\chi(x_0)$$

$$= R_1^{\pi}(D)\tilde{f}(x_0).$$

Thus, the proof of Lemma 1 is completed.

Lemma 2. *Let $u(x)$ be a bounded function such that $u(x) = 0$ for $x \notin D$. Then $u(x_0) = -R_1^{\pi}(D)\omega(x_0)$, where*

$$\omega(x,l) = \sum_{j \in D} u(j)\rho_l(x,j) - u(x).$$

Proof. Note that $u(x_0) = -\mathrm{E}_{x_0}^{\pi} \sum_{n=0}^{\tau-1} (u(x_{n+1}) - u(x_n))$. Indeed,

$$-\mathrm{E}_{x_0}^{\pi} \sum_{n=0}^{\tau-1} (u(x_{n+1} - u(x_n)) = \mathrm{E}_{x_0}^{\pi} \left(\sum_{n=0}^{\infty} u(x_n)\chi\{n < \tau\} - \sum_{n=1}^{\infty} u(x_n)\chi\{n \leq \tau\} \right)$$

$$= \mathrm{E}_{x_0}^{\pi} \left(u(x_0)\chi\{0 < \tau\} - \sum_{n=1}^{\infty} u(x_n)\chi\{n = t\} \right) = u(x_0),$$

since $u(x) = 0$ for $x \notin D$ and $\chi\{0 < t\} = 1$. Further,

$$\mathbf{E}^{\pi}_{x_0} u(x_{n+1})\chi\{n < \tau\} = \sum_{j\in D} u(j)\mathbf{P}^{\pi}_{x_0}\{x_{n+1} = j, n < \tau\}$$

$$= \sum_{j\in D} u(j) \sum_{l\in A}\sum_{i\in D} \mathbf{P}^{\pi}_{x_0}\{x_{n+1} = j, a_n = l, x_n = i, n < \tau\}$$

$$= \sum_{j\in D}\sum_{i\in D}\sum_{l\in A} u(j)\rho_l(i,j)\mathbf{E}^{\pi}_{x_0}\chi\{a_n = l, x_n = i, n < \tau\}$$

$$= \mathbf{E}^{\pi}_{x_0} \sum_{j\in D} u(j)\rho_{a_n}(x,j)\chi\{n < \tau\}.$$

Therefore,

$$u(x_0) = -\mathbf{E}^{\pi}_{x_0} \sum_{n=0}^{\tau-1} (u(x_{n+1}) - u(x_n))$$

$$= -\mathbf{E}^{\pi}_{x_0} \sum_{n=0}^{\tau-1} \left(\sum_{j\in D} u(j)\rho_{a_n}(x_n, j) - u(x_n) \right) = -R^{\pi}_1(D)\omega(x_0).$$

All the calculations with the series are legitimate because of the condition (2). Thus,the lemma is proved.

P r o o f of Theorem 1. Take an arbitrary bounded function $r(x, a)$ and consider $u(x) = R^{\tilde{\pi}}_1(D)r(x)$. Note that for $x \notin D$, $u(x) = 0$. If $x \in D$, we have

$$u(x) = E^{\tilde{\pi}}_x \sum_{n=0}^{\infty} r(x_n, a_n)\chi\{n < \tau\}$$

$$= \sum_{l\in A} r(x,l)\tilde{\pi}(l|a) + \sum_{l\in A}\sum_{j\in D} u(j)\tilde{\pi}(l|x)\rho_l(x,j).$$

Hence, for $x \in D$:

$$\sum_{l\in A} r(x,l)\tilde{\pi}(l|x) = u(x) - \sum_{l\in A}\sum_{j\in D} u(j)\tilde{\pi}(l|x)p_l(x,j).$$

Now, for $x \in D$ we have

$$\tilde{r}(x) = \sum_{l\in A} \tilde{\pi}(l|x)u(x) - \sum_{l\in A}\sum_{j\in D} u(j)\tilde{\pi}(l|x)\rho_l(x,j) = -\sum_{l\in A} \tilde{\pi}(l|x)\omega(x,l) = -\tilde{\omega}(x).$$

Next, by using lemmas 2 and 1, we obtain:

$$u(x_0) = -R_1^{\pi}(D)\omega(x_0) = -R_1^{\pi}(D)\omega(x_0) = R_1^{\pi}(D)\tilde{r}(x_0) = R_1^{\pi}(D)r(x_0).$$

On the other hand, by definition, $u(x_0) = R_1^{\tilde{\pi}}(D)\tilde{r}(x_0)$, which implies that

$$R_1^{\pi}(D)r(x_0) = R_1^{\tilde{\pi}}(D)r(x_0).$$

Thus Theorem 1 is proved.

3. Applications. In this section some applications of Theorem 1 are given.

Corollary 1. *Let $0 < \beta < 1$. Fix $x_0 \in X$. Then, for each strategy π there exists $\tilde{\pi} \in \mathcal{RM}$ such that for every bounded function $r(x, a)$:*

$$R_{\beta}^{\pi}(X)r(x_0) = R_{\beta}^{\tilde{\pi}}(X)r(x_0).$$

Proof. To establish this, we use Theorem 1, incorporating the discount coefficient β into the transition probabilities and by adding an additional absorption state x^* as follows.

Denote $\widehat{X} = X \cup \{x^*\}$, $x^* \notin X$, $\widehat{A} = A$, $\widehat{r}(i,a) = r(i,a)$, $\widehat{r}(x^*,a) = 0$, $\widehat{\rho}_a(i,j) = \beta\rho_a(i,j)$, $\widehat{\rho}_a(i,x^*) = 1-\beta$, $\widehat{\rho}_a(x^*,x^*) = 1$ for all $i,j \in X$, $a \in A$. Further, denote $\tau = \inf\{n > 0 : x_n = x^*\} = \inf\{n > 0 : x_n \notin X\}$. It is easy to check by induction that

$$\forall \pi \forall i,a,n : \widehat{\mathrm{P}}_{x_0}^{\pi}\{x_n = i, a_n = a\} = \beta^n \mathrm{P}_{x_0}^{\pi}\{x_n = i, a_n = a\}.$$

Hence, for every strategy π we have

$$\widehat{R}_1^{\pi}(X)r(x_0) = R_{\beta}^{\pi}(X)r(x_0).$$

Now, using theorem 1, we find a strategy $\tilde{\pi} \in \mathcal{RM}$ such that

$$\widehat{R}_1^{\pi}(X)r(x_0) = \widehat{R}_1^{\tilde{\pi}}(X)r(x_0).$$

This means that

$$R_{\beta}^{\pi}(X)r(x_0) = R_{\beta}^{\tilde{\pi}}(X)r(x_0).$$

Thus the proof is completed.

Corollary 2. *Under the assumptions of Theorem 1, for each strategy π there exists $\tilde{\pi} \in \mathcal{RM}$ such that for every bounded function $r(x, a)$:*

$$\mathrm{E}_{x_0}^{\pi} \sum_{n=0}^{\tau} r(x_n, a_n) = \mathrm{E}_{x_0}^{\tilde{\pi}} \sum_{n=0}^{\tau} r(x_n, a_n)$$

Proof. Let us add one more state, x^*, such that $\rho_a(x^*, x^*) = 1$ for all $a \in A$ and $\rho_a(x, x^*) = 1$ for all $x \in X \setminus D$ and for all $a \in A$. Denote by σ the first moment for the process to exit X (the old state space), i.e. the first moment to reach x^*. Apply Theorem 1, taking into account that $\sigma - 1 = \tau$. This completes the proof.

Next follows a new, straightforward proof of a known fact. In the Introduction of [3] it is proved by induction while here it is derived immediately from Corollary 1:

Corollary 3. *Fix $x_0 \in X$. Then for each strategy $\pi \in \Pi$ there exists a randomized time-depending Markov strategy $\tilde{\pi}$, such that for every $n \in \mathbf{N}$, $y \in X$, $b \in A$ the following equality holds:*

$$\mathrm{P}^{\pi}_{x_0}\{x_n = y, a_n = b\} = \mathrm{P}^{\tilde{\pi}}_{x_0}\{x_n = y, a_n = b\}.$$

The strategy $\tilde{\pi}$ can be determined as follows:

$$\tilde{\pi}_n(a|x) = \mathrm{P}^{\pi}_{x_0}\{x_n = x, a_n = a\} \cdot \left(\mathrm{P}^{\pi}_{x_0}\{x_n = x\}\right)^{-1}.$$

Proof. Along with the process $x_0, a_0, x_1, a_1, \ldots$ consider the "extended" process $\hat{x}_0, a_0, \hat{x}_1, a_1, \ldots$, where $\hat{x}_n = (x_n, n) \in X \times (\mathbf{N} \cup \{0\})$, $a_n \in A$. This new process can be regarded as a decision Markov chain with a state set $\hat{X} = X \times (\mathbf{N} \cup \{0\})$, with an action set A, and with transition probabilities $\hat{\rho}_a((x, m), (y, n)) = \rho_a(x, y)\delta_{m+1,n}$, where $\delta_{i,j}$ stands for the Kronecker's δ-symbol. Moreover, there is an obvious and natural interconnection between the strategies π for the original process and the strategies π of the "extended" one:

$$\hat{\pi}_n(a|(x_0, 0), a_0, \ldots, (x_n, n)) = \pi_n(a|x_0, a_0, \ldots, x_n).$$

Note that for every n, x and a:

$$\mathrm{P}^{\pi}_{x_0}\{x_n = x, a_n = a\} = \mathrm{P}^{\hat{\pi}}_{x_0}\{\hat{x}_n = (x, n), a_n = a\}.$$

Note also that to a stationary Markov strategy $\hat{\pi}$, there is a corresponding Markov time-depending strategy π. To finish the proof, apply Corollary 1, with an arbitrary fixed $\beta \in (0, 1)$, taking for r the function $r(\hat{x}, a) = r((x, k), a) = \chi\{x = y, k = n, a = b\}$, where the integer n, $y \in X$, and $b \in A$ can be chosen arbitrarily.

The next corollary presents an illustration of the usefulness of Theorem 1. Usually results of this type are obtained by using Bellman's principle (see, e.g. [2], [3]). Here we give a very simple proof. Moreover, in this form it is a new result and in our

view, it will be shorter and more convenient to give a straightforward proof rather than to try to deduce it from a similar fact.

Corollary 4. *Under the assumptions of Theorem 1, for every bounded function $r(x,a)$ the following relation holds:*

$$\sup_{\pi\in\Pi} R_1^\pi(D)r(x_0) = \sup_{\pi\in\mathcal{M}} R_1^\pi(D)r(x_0),$$

where Pi is the class of all strategies, and $\mathcal{M}$ is the class of all pure (i.e. non-randomized) Markov strategies.

Proof. Fix $x_0 \in D$. From Theorem 1 we have

$$\forall\pi \in \Pi \;\; \exists\widetilde{\pi} \in \mathcal{RM} \text{ such that } R_1^\pi(D)r(x_0) = R_1^{\widetilde{\pi}}(D)r(x_0).$$

Now it is enough to show that $\forall\pi \in \mathcal{RM}\ \exists\sigma \in \mathcal{M}$ such that $u(\pi,x_0) \le u(\sigma,x_0)$, where $u(\pi,x_0) = R_1^\pi(D)r(x_0)$.

For each $x \in D$ we have $u(\pi,x) = \sum_{a\in A} \pi(a|x)g(x,a)$, where $g(x,a) = r(x,a) + \sum_{y\in D} \rho_a(x,y)u(\pi,y)$. Since for every fixed x the probabilities $\pi(a|x)$ determine a distribution on A, then for every $x \in X$ an action $a_x \in A$ can be found, such that $g(x,a_x) \ge u(\pi,x)$. Let us determine the strategy σ as follows:

$$\sigma(a|x) = \begin{cases} 1, & a = a_x \\ 0, & a \ne a_x \end{cases}$$

Then,

$$u(\sigma\circ\pi,x) = r(x,a_x) + \sum_{y\in D} \rho_{a_x}(x,y)u(\pi,y) = g(x,a_x)$$

So, $u(\sigma\circ\pi,x) \ge u(\pi,x)$. Here $\sigma\circ\pi$ denotes a strategy which coincides with σ at the zero moment and with π further on. More generally, we use the notation $\sigma^n\circ\pi$ to denote the strategy, which prescribes applying σ in the first n steps, and π further on after that, i.e.

$$(\sigma^n\circ\pi)_k(a|x) = \begin{cases} \sigma(a|x), & \text{if } k < n \\ \pi(a|x), & \text{if } k \ge n. \end{cases}$$

From the last inequality we conclude by induction that for every n, $u(\sigma^n\circ\pi,x) \ge u(\sigma^{n-1}\circ\pi,x)$, and hence $u(\sigma^n\circ\pi,x) \ge u(\pi,x)$. Further,

$$u(\sigma^n\circ\pi,x) = \mathbf{E}_x^\sigma \sum_{k=0}^{n-1} \chi\{k<\tau\}r(x_k,a_k) + \mathbf{E}_x^\sigma \chi\{n<\tau\}u(\pi,x_n).$$

The function r is bounded, $\sup|r(x,a)| = C < \infty$. Therefore, for every x:

$$|u(\pi,x)| \le \mathbf{E}_x^\pi \sum_{k=0}^{\infty} \chi\{k<\tau\} C \le CT < \infty.$$

This implies $|\mathbf{E}_x^\sigma \chi\{n<\tau\} u(\pi,x_n)| \le C\cdot T\cdot \mathbf{E}_x^\sigma \chi\{n<\tau\}$. The last expectation tends to zero as $n\to\infty$, by the Tchebyshev's inequality and (2). Consequently, $u(\sigma^n\circ\pi,x)\to u(\sigma,x)$ as $n\to\infty$. Thus we obtain $u(\pi,x)\le u(\sigma,x)$, and in particular $u(\pi,x_0)\le u(\sigma,x_0)$ which yields the proof.

4. Optimal Control in a Problem with Constraints.

Theorem 2. *Let the state set X and the action set A be finite. Let a subset $D\subseteq X$ and an initial state $x_0\in D$ be given. Let τ be the exit time defined by (1) and assume that (2) is satisfied. Let $f: X\times A\to \mathbf{R}^n$ be an arbitrary bounded vector-valued function, and let F be a closed subset of $\mathbf{R}^n$. Denote by $\mathcal{G}$ the class of strategies π such that $\mathbf{E}_{x_0}^\pi \sum_{n=0}^{\tau} f(x_n,a_n)\in F$, and by $\mathcal{GM}$ the subclass of all Markov strategies in $\mathcal{G}$. Suppose $\mathcal{G}\neq\emptyset$. Then for every bounded function $r(x,a)$:*

$$\sup_{\pi\in\mathcal{G}} \mathbf{E}_x^\pi \sum_{n=0}^{\tau} r(x_n,a_n) = \sup_{\widetilde{\pi}\in\mathcal{GM}} \mathbf{E}_{x_0}^{\widetilde{\pi}} \sum_{n=0}^{\tau} r(x_n,a_n)$$

Moreover, a strategy $\widetilde{\pi}\in\mathcal{GM}$ can be found for which the supremum is attained.

Proof. From Corollary 2 it follows immediately that both suprema are equal. We just need to show that this supremum is attained for some $\widetilde{\pi}\in\mathcal{GM}$.

It will be convenient for the purposes of our proof to consider the strategies in $\mathcal{RM}$ as points of the finite dimensional space $\mathbf{R}^{|X|\cdot|A|}$. Then the class $\mathcal{RM}$ can be regarded as a subset of $\mathbf{R}^{|X|\cdot|A|}$ described by the constraints:

1) $\forall a\forall x: 0\le \widetilde{\pi}(a|x)\le 1$.

2) $\forall x: \sum_{a\in A}\widetilde{\pi}(a|x)=1$.

The first family of inequalities shows that this subset is contained in the cube $[0,1]^{|X|\cdot|A|}$, and is therefore bounded. Further, since all inequalities are not strict (i.e. they admit equality) this subset is closed.

Let us prove first that for every bounded function g and for every $x\in D$ the expression $\mathbf{E}_x^{\widetilde{\pi}} \sum_{n=0}^{\tau} g(x_n,a_n)$ is continuous with respect to $\widetilde{\pi}\in\mathcal{RM}$ (if the strategies of $\mathcal{RM}$ are regarded as points in $\mathbf{R}^{|X|\cdot|A|}$). This statement is clear for expressions of the

type $\mathbf{E}_x^{\tilde{\pi}} \sum_{n=0}^{N} \chi\{n \le \tau\} g(x_n, a_n)$. For $\mathbf{E}_x^{\tilde{\pi}} \sum_{n=0}^{\tau} g(x_n, a_n) = \sum_{n=0}^{\infty} \mathbf{E}_x^{\tilde{\pi}} \chi\{n \le \tau\} g(x_n, a_n)$ the continuity follows from the uniform convergence of the series $\sum_{n=0}^{\infty} \mathbf{E}_x^{\tilde{\pi}} \chi\{n \le \tau\} g(x_n, a_n)$, which can be estimated as follows. From Tchebyshev's inequality $\forall \tilde{\pi} \in \mathcal{RM} : \mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2T\} \le 0.5$. Further,

$$\mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2nT\}$$

$$= \sum_{y \in D} \mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2nT | x_{(n-1)2T} = y, \tau \ge 2(n-1)T\} \cdot \mathbf{P}_x^{\tilde{\pi}}\{x_{(n-1)2T} = y, \tau \ge 2(n-1)T\}$$

$$\le 0.5 \sum_{y \in D} \mathbf{P}_x^{\tilde{\pi}}\{x_{(n-1)2T} = y, \tau \ge 2(n-1)T\} = 0.5 \mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2(n-1)T\}.$$

From here we can conclude that $\mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2nT\} \le (0.5)^n$. Using this estimation we get

$$\sum_{n=1}^{\infty} \mathbf{P}_x^{\tilde{\pi}}\{\tau \ge n\} = \sum_{m=0}^{\infty} \sum_{k=1}^{2T} \mathbf{P}_x^{\tilde{\pi}}\{\tau \ge 2mT + k\} \le \sum_{m=0}^{\infty} \sum_{k=1}^{2T} (0.5)^m = 4T.$$

Since the absolute value of the function g is bounded by some constant C, we obtain

$$\sum_{n=0}^{\infty} |\mathbf{E}_x^{\tilde{\pi}} \chi\{n \le \tau\} g(x_n, a_n)| \le C \sum_{n=0}^{\infty} \mathbf{E}_x^{\tilde{\pi}} \chi\{n \le \tau\} \le C \cdot \left(1 + \sum_{m=0}^{\infty} \sum_{k=1}^{2T} (0.5)^m\right) < \infty.$$

Thus, we proved that $\mathbf{E}_{x_0}^{\tilde{\pi}} \sum_{n=0}^{\tau} g(x_n, a_n)$ is continuous with respect to $\tilde{\pi} \in \mathcal{RM}$ for each $x \in D$. Therefore, both $\mathbf{E}_{x_0}^{\tilde{\pi}} \sum_{n=0}^{\tau} f(x_n, a_n)$ and $\mathbf{E}_{x_0}^{\tilde{\pi}} \sum_{n=0}^{\tau} r(x_n, a_n)$ are continuous with respect to $\tilde{\pi} \in \mathcal{RM}$.

It is well known that a continuous function on a compact attains its supremum. Therefore it is enough to prove that $\mathcal{GM}$ is a compact. As we already mentioned, in the space $\mathbf{R}^{|X| \cdot |A|}$ the class of strategies $\mathcal{RM}$ forms a closed and bounded subset. Since X and A are finite, it is compact. Further, since $\mathbf{E}_{x_0}^{\tilde{\pi}} \sum_{n=0}^{\tau} f(x_n, a_n)$ is continuous with respect to $\tilde{\pi} \in \mathcal{RM}$, and F is closed, then $\mathcal{GM}$ is closed as the pre-image of a closed set under a continuous mapping. Since $\mathcal{GM} \subseteq \mathcal{RM}$, and $\mathcal{RM}$ is compact, we conclude that $\mathcal{GM}$ is compact. Thus, the theorem is proved.

Corollary. *Let the state set X and the action set A be finite, and let $0 < \beta < 1$. Fix a state $x_0 \in X$. Let $f : X \times A \to \mathbf{R}^n$ be an arbitrary bounded vector-valued*

function, and let F be a closed subset of $\mathbf{R}^m$. Denote by $\mathcal{G}$ the class of strategies π such that $R^\pi_\beta(X)f(x_0) \in F$, and by $\mathcal{GM}$ the subclass of all Markov strategies in $\mathcal{G}$. Suppose $\mathcal{G} \neq \emptyset$. Then for every bounded function $r(x,a)$:

$$\sup_{\pi \in \mathcal{G}} R^\pi_\beta(X) r(x_0) = \sup_{\widetilde{\pi} \in \mathcal{GM}} R^{\widetilde{\pi}}_\beta(X) r(x_0).$$

Moreover, a strategy $\widetilde{\pi} \in \mathcal{GM}$ can be found for which the supremum is attained.

Proof. This fact is derived from Theorem 2 in the same way as Corollary 1 is derived from Theorem 1.

REFERENCES

[1] Н. В. Крылов. Об одном подходе к управляемым диффузионным процессам. *Теория Вероятн. и ее прим.*, **XXXI** (4) (1986), 685-709.

[2] D. P. Bertsekas, S. E. Shreve. Stochastic Optimal Control. The Discrete Time Case. Academic Press, New York,San Francisco, London, 1978.

[3] J. Van der Wal. Stochastic Dynamic Programming, Successive Approximations and Nearly Optimal Strategies for Markov Decision Processes and Markov Games. Amsterdam, Mathematisch Centrum, 1981.

[4] J. Neveu. Mathematical Foundations of the Calculus of Probability, Holden-Day, San Francisco, London, Amsterdam, 1965.

Sofia University "St. KL. Ohridski"
Faculty of Mathematics and Informatics
5, James Bouchier str
1126 Sofia
BULGARIA

Received 17.12.91
Revised 19.07.1993