九州大学学術情報リポジトリ
Kyushu University Institutional Repository

ON AN $ ¥varepsilon $-OPTIMAL POLICY OF DISCRETE TIME STOCHASTIC CONTROL PROCESSES

Tanaka, Kensuke
Department of Mathematics, Faculty of Science, Niigata University

Hoshino, Mitsuhiro
Department of Mathematical Science , Graduate School of Science and Technology, Niigata University

Kuroiwa, Daishi
Department of Mathematical Science , Graduate School of Science and Technology, Niigata University

https://doi.org/10.5109/13445

出版情報：Bulletin of informatics and cybernetics. 27 (1), pp.107-119, 1995-03. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

Bulletin of Informatics and Cybernetics, Vol. 27, No. 1, 1995

ON AN ε-OPTIMAL POLICY OF DISCRETE TIME STOCHASTIC CONTROL PROCESSES

By

Kensuke TANAKA,* Mitsuhiro HOSHINO† and Daishi KUROIWA†

Abstract

In this paper, stochastic control processes have been investigated as dynamic programming models with an infinite horizon. In many cases, it is our main purpose to seek for an optimal policy under the various conditions. However, optimal policy may not exist under weak conditions. In such situation, it will be necessary to seek for ε-optimal policy. Thus, after an ε-optimal policy has been sought, we show that there exists at least a better policy than the given ε-optimal one and that the better policy is near to the given ε-optimal one. Then, the principle of Ekeland's theorem will play an important role. Moreover, we introduce a modified control model with the perturbed one-stage cost functions and show that the better policy is an optimal one of the modified model.

Key words: dynamic programming, optimal policy, ε-optimal policy, and lower semicontinuous

1. Introduction

In recent years, the dynamic programming problems with an infinite horizon have actively been investigated by many authors. Much of the earlier works of this area were done by Blackwell [4] [5] and Strauch [14]. Further, Dynkin and Yushkevich [8] and Hinderer [11] gave extensive accounts of dynamic programming with discrete time parameter. In many cases, the concept of optimal policy is introduced and, then, the existence of an optimal one is shown. However, in order to show the existence of an optimal policy, we need to assume strong condition that action space for player is compact. See Badler [2] and Schäll [12] [13] in detail. Then, from mathematical and/or practical view points, the compactness is weakened. Further, an optimal policy in an error of a given $\varepsilon > 0$, which is said to be ε-optimal, is studied. In Bertsekas and Shreve [3], it is discussed in detail that the existence of such an ε-optimal policy depends on the measurability structure of the model.

In this paper, we will describe a basic minimization problem with constraints for Markovian control model on an infinite horizon. We exclude the compactness of the

* Department of Mathematics, Faculty of Science, Niigata University, 950-21, Niigata, Japan.

† Department of Mathematical Science, Graduate School of Science and Technology, Niigata University, 950-21, Niigata, Japan.

action space. Under such weak condition, it may be impossible to seek an explicit optimal policy for the control model. Thus, it is important to seek an ε-optimal policy by which an optimal value was approximated. But, in the paper, it is not our purpose to show that there exists an ε-optimal one. After an ε-optimal policy has been sought, we want to show that it is exactly possible to seek at least a better policy than the ε-optimal one. Further, we show that the better policy is near to the ε-optimal one in a sense of some distance between both the policies. The way and the principle of Ekeland's theorem [9] [10], which gives an approximate optimal solution, will play an important role. Moreover, we introduce a modified control model with one-stage cost functions perturbed by the linear forms and show that the better policy is an optimal one for the modified model. To do this, it is a useful condition that the one-stage cost function and the integral operator are lower semicontinuous on the action space for each state.

This paper is organized in the following way. In Section 2, we formulate a basic minimization problem for the control model and give the definitions of optimal policy and ε-optimal one. In Section 3, from the base on Ekeland's theorem, we show that there exists exactly a better policy than the given ε-optimal one. Further, we show that the better policy is near to the given ε-optimal one. In Section 4, we introduce the modified control model with the perturbed one-stage cost functions and show that the better policy is an optimal one of the modified model.

2. Formulation of a stochastic optimal control model

A stochastic optimal control model is specified by six-tuple

$$(S, C, U, q, r, \beta), \tag{2.1}$$

where

(i) S is a nonempty Borel subset of a Polish (i.e., complete, separable, metric) space, the *state space.*

(ii) C is a nonempty Borel subset of a Polish space, namely, the *control space.*

(iii) U is a multifunction which assigns to each state $x \in S$ a non-empty permissible set of controls $U(x) \subset C$. We assume that $GrU = \{(x,u) | x \in S, u \in U(x)\}$ is analytic in SC, where SC denotes the Cartesian product of sets S, C. See Bertsekas and Shreve [3, Section 7.6 page 156] for analytic sets in details.

(iv) q is a Borel measurable stochastic kernel on S given SC, that is, $q(B|x,u)$ is a probability of a Borel subset $B \subset S$ for each $(x,u) \in GrU$ and a Borel measurable function of $(x,u) \in GrU$ for each Borel subset B. The law of motion is given by q.

(v) r is an extended real valued function, $GrU \to R^*$, where $R^* = R \cup \{\infty\}$, which is lower semianalytic, i.e., for any $c \in R^*$, $\{(x,u) \in GrU | r(x,u) < c\}$ is analytic, the *one-stage cost function.*

(vi) β is a *discount factor.*

In the specification, we should note that the permissible set of controls $U(x)$ depends on state $x \in S$ and $q(\cdot|x,u)$ is independent of time.

Then, a policy π for the model is defined as an infinite sequence $\pi = (\mu_0, \mu_1, \cdots, \mu_k, \cdots)$ with the property that, for each k, $\mu_k(du_k|x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$ is a universally measurable stochastic kernel on C given $SCS \cdots CS$ satisfying

$$\mu_k(U(x_k)|x_0, u_0, x_1, \cdots, u_{k-1}, x_k) = 1$$

for every $(x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$, where x_k and u_k denote k-th state and control, respectively. See [3, Section 7.7] for universally measurable stochastic kernels in details. If, for each k, μ_k is parametrized only by x_k, π is said to be a *Markov policy.* If, for each k and $(x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$, the stochastic kernel $\mu_k(du_k|x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$ assigns mass one to some point in C, π is said to be *nonrandomized.* In this case, π can be considered to be an infinite sequence $\pi = (\mu_0, \mu_1, \cdots)$, each component μ_k of which is an universally measurable mapping from $SCS \cdots CS$ to C with the property that, for every $(x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$,

$$\mu_k(x_0, u_0, x_1, \cdots, u_{k-1}, x_k) \in U(x_k).$$

If $\mathcal{F}$ is a σ-algebra on a Polish space and each stochastic kernel component of a policy is $\mathcal{F}$-measurable, we say that policy is $\mathcal{F}$-measurable. We denote by Π the set of all policies and by Π_M the set of all Markov policies. If π is a Markov policy of the form $\pi = (\mu, \mu, \cdots)$, it is said to be *stationary.*

Thus, the control system is interpreted as following. If a policy $\pi = (\mu_0, \mu_1, \cdots, \mu_k, \cdots)$ is employed, at the successive k-th stage, we observe k-th state and classify it to a possible state $x_k \in S$. Then, we choose a control $u_k \in U(x_k)$ by k-th universally measurable stochastic kernel μ_k depending on a *history* $(x_0, u_0, x_1, \cdots, u_{k-1}, x_k)$ up to k-th stage. As a result of state x_k and control u_k at k-th stage, we will incur a cost $r(x_k, u_k)$. The control system moves to a new state $x_{k+1} \in S$ according to the stochastic kernel $q(\cdot|x_k, u_k)$. After that, the process is analogously developed from x_{k+1}. Since GrU is analytic, from the Jankov von-Neumann theorem, it follows that there exists at least one nonrandomized Markov policy.

In the subsequent discussion, we will often use S_k and C_k, $k = 0, 1, 2, \cdots$ as copies of S and C respectively. Given an initial distribution p on S and any policy $\pi = (\mu_0, \mu_1, \mu_2, \cdots)$ together with the stochastic kernel q, there is a sequence of unique probability measure $P_t^{\pi,p}$ on $S_0C_0S_1C_1 \cdots S_{t-1}C_{t-1}$, $t = 1, 2, \cdots$, of future up to t-th stage. Further, there exists a unique probability measure $P^{\pi,p}$ on $S_0C_0S_1C_1 \cdots$ such that for each t, marginal measure of $P^{\pi,p}$ on $S_0C_0S_1C_1 \cdots S_{t-1}C_{t-1}$, is $P_t^{\pi,p}$, (see Hinderer [11, page 80]).

Let $h_t = (x_0, u_0, x_1, u_1, \cdots, x_t, u_t)$ denote a history of the control system up to $(t+1)$-th stage. The expected cost at $(t+1)$-th stage is given by

$$E_\pi\left[r(x_t, u_t)\right] = \int_{S_0C_0 \cdots S_tC_t} r(x_t, u_t) P_{t+1}^{\pi,p}(dh_t).$$

and the *total expected discounted cost* is given by

$$I(\pi) = \sum_{t=0}^{\infty} \beta^t E_\pi \left[r(x_t, u_t) \right]. \tag{2.2}$$

If $\pi = (\mu, \mu, \cdots)$ is stationary, we often write $I(\mu)$ in place of $I(\pi)$. Further, if an initial distribution p assigns mass one to a point $x \in S$, p is written as p_x and if p_x is used, we write $I(\pi)(x)$ in place of $I(\pi)$.

Then, we consider a basic minimization problem (MP) for the dynamic control system :

$$\text{(MP)} \qquad \text{minimize} \quad I(\pi)(x) \quad \text{subject to} \quad \pi \in \Pi.$$

DEFINITION 2.1. The *optimal discounted cost* $I^*(x)$ at an initial state x is defined by

$$I^*(x) = \inf_{\pi \in \Pi} I(\pi)(x). \tag{2.3}$$

DEFINITION 2.2. If $I^*(x) = I(\pi^*)(x)$, the policy $\pi^* \in \Pi$ is said to be *optimal at an initial state* x. If π^* is optimal at every initial state x, it is said to be *optimal.*

In this paper we will study the problem (MP) in only the case such that the optimal cost is finite for all initial states, that is, $I^*(x) \in R$ for all $x \in S$. Thus, we define an ε-optimal policy as follows.

DEFINITION 2.3. For a given $\varepsilon > 0$, a policy $\pi_\varepsilon \in \Pi$ is said to be *ε-optimal at an initial state* x if

$$I(\pi_\varepsilon)(x) \leq I^*(x) + \varepsilon.$$

If π_ε is ε-optimal at every initial state x, it is said to be *ε-optimal.*

3. The existence of a better policy than a given ε-optimal one

When the stochastic control models are generally discussed, we will treat the following three cases for one-stage cost function :

(D) $\quad 0 < \beta < 1$ and for some $M \in R$,

$\quad |r(x,u)| \leq M$ for every $(x,u) \in GrU$,

(P) $\quad \beta = 1$ and $0 \leq r(x,u)$ for every $(x,u) \in GrU$,

(N) $\quad \beta = 1$ and $r(x,u) \leq 0$ for every $(x,u) \in GrU$.

Further, in these cases, we can get the following lemma.

LEMMA 3.1. *For any initial state $x \in S$ and any policy $\pi \in \Pi$, there exists a Markov policy $\pi_M \in \Pi_M$ such that*

$$I(\pi_M)(x) = I(\pi)(x).$$

The proof is given in Bertsekas and Shreve [3, Prop.8.1 and Prop.9.1].

Further, in cases (D) and (P), Bertsekas and Shreve [3, Prop. 9.19] give the following lemma for ε-optimal policy.

LEMMA 3.2. *For each $\varepsilon > 0$, there exists an ε-optimal nonrandomized Markov policy and, if $0 < \beta < 1$, it can be taken to be stationary.*

From the results of these lemmas, we will treat only the nonrandomized Markov policies, the set of which is denoted by $\Pi_N \subset \Pi_M$, in cases (D) and (P). Thus, for the original problem (MP), we consider the following minimization problem ($\overline{\text{MP}}$).

$$(\overline{\text{MP}}) \qquad \text{minimize} \quad I(\pi)(x) \quad \text{subject to} \quad \pi \in \Pi_N.$$

Let $M(S)$ be the set of all extended real valued functions on S, which are lower semianalytic, $\not\equiv \infty$, and bounded from below. Further, let $N(C|S)$ be the set of all universally measurable functions $f : S \to C$ such that $f(x) \in U(x)$ for each $x \in S$. Such a measurable function f will be called a *selector* of U. Then, a nonrandomized Markov policy π is described by a sequence of selectors $\pi = (f_0, f_1, f_2, \cdots)$ The set of nonrandomized Markov policies is considered as $\Pi_N = N(C|S)N(C|S)\cdots$. For each selector $f \in N(C|S)$, we define an operator T_f on $M(S)$ as follows : for each $v \in M(S)$ and $x \in S$,

$$T_f v(x) = r(x, f(x)) + \beta \int_S v(x')q(dx'|x, f(x)). \tag{3.1}$$

The operator T_f is a mapping from $M(S)$ into $M(S)$. Let $v_0 \in M(S)$ be identically zero. Then, a successive use of the operators $T_{f_0}, T_{f_1}, \cdots, T_{f_{t-1}}, \cdots$ for $\pi = (f_0, \cdots, f_{t-1}, \cdots)$ yields the *t-stage cost function* $I_t(\pi)(x)$ and *the total expected cost function* as follows:

$$I_t(\pi)(x) = (T_{f_0} T_{f_1} \cdots T_{f_{t-1}}) v_0(x) \tag{3.2}$$

and

$$I(\pi)(x) = \lim_{t \to \infty} I_t(\pi)(x). \tag{3.3}$$

From Lemma 3.2, there exists ε-optimal nonrandomized Markov policy for the fairly general control models. Thus, in this paper, after we sought an ε-optimal nonrandomized Markov policy in cases (D) and (P), we want to show that there exists at least a better nonrandomized Markov policy than the ε-optimal one sought. Further, we show that the better policy exists at any near place of the ε-optimal one sought. Then, in order to show main results, we will impose the following assumptions on the control model.

(A1) For each $x \in S$, the non-empty permissible set of controls $U(x)$ is closed in C.

(A2) r is bounded from below on SC, $\not\equiv \infty$, and, for each $x \in S$, $r(x,u)$ is l.s.c. on $U(x)$, that is, for any convergent sequence of controls $\{u_k\}_{k=1,2,\cdots}$ in $U(x)$ such that $\rho(u_k, u) \to 0$ as $k \to \infty$,

$$\liminf_{k\to\infty} r(x, u_k) \geq r(x, u),$$

where ρ is the metric on the control space C.

(A3) For any $w \in M(S)$ and $x \in S$, the integral operator

$$\int_S w(y) q(dy|x,u)$$

is a l.s.c. function with respect to $u \in U(x)$.

Let $\pi_\varepsilon = (f_0^\varepsilon, f_1^\varepsilon, \cdots)$ be an ε-optimal nonrandomized Markov policy sought in the control model. To keep the notation short, we shall write $\rho(f,g)(x) = \rho(f(x), g(x))$. From the base on Ekeland's theorem, we prove the following lemma.

LEMMA 3.3. *For each n-th selector f_n^ε of the given ε-optimal policy $\pi_\varepsilon \in \Pi_N$ and any $w \in M(S)$, there exists $f_n^* \in N(C|S)$ such that for all $x \in S$,*

$$T_{f_n^*} w(x) + \varepsilon\beta\rho(f_n^*, f_n^\varepsilon)(x) \leq T_{f_n^\varepsilon} w(x), \tag{3.4}$$

and for any $f \in N(C|S)$ such that $\rho(f, f_n^*)(x) > 0$,

$$T_f w(x) > T_{f_n^*} w(x) - \varepsilon\beta\rho(f_n^*, f)(x). \tag{3.5}$$

Proof. Let n and $w \in M(S)$ be fixed. For the given selector $f_n^\varepsilon \in N(C|S)$, we put $f_{n,0}^*(x) = f_n^\varepsilon(x)$ for any $x \in S$. We define inductively a sequence of selectors $f_{n,k}^*$, $k \geq 1$, by a sequence of controls $f_{n,k}^*(x)$, $k \geq 1$ at each $x \in S$. Suppose a selector $f_{n,k}^*$ is given. We use the following notation

$$N_n(k,x) = \left\{ f \in N(C|S) \;\middle|\; \begin{array}{l} f(x) \neq f_{n,k}^*(x), \\ T_f w(x) \leq T_{f_{n,k}^*} w(x) - \varepsilon\beta\rho(f, f_{n,k}^*)(x) \end{array} \right\}.$$

Then if $N_n(k,x) = \phi$, we define $f_{n,k+1}^*(x) = f_{n,k}^*(x)$, If $N_n(k,x) \neq \phi$, we can select $f_{n,k+1}^* \in N_n(k,x)$ satisfying

$$2T_{f_{n,k+1}^*} w(x) - T_{f_{n,k}^*} w(x) \leq \inf_{f \in N_n(k,x)} T_f w(x). \tag{3.6}$$

Indeed, if we choose a function $g \in N_n(k,x)$, we have

$$\begin{aligned} 0 &< \varepsilon\beta\rho(g, f_{n,k}^*)(x) \\ &\leq T_{f_{n,k}^*} w(x) - T_g w(x) \\ &\leq T_{f_{n,k}^*} w(x) - \inf_{f \in N_n(k,x)} T_f w(x). \end{aligned} \tag{3.7}$$

For $\frac{1}{2}\varepsilon\beta\rho(g, f^*_{n,k})(x) > 0$, from the definition of infimum, it follows that there exists $g' \in N_n(k,x)$ such that

$$T_{g'}w(x) \leq \inf_{f \in N_n(k,x)} T_f w(x) + \frac{1}{2}\varepsilon\beta\rho(g, f^*_{n,k})(x). \tag{3.8}$$

According to (3.7) and (3.8), we obtain

$$2T_{g'}w(x) - T_{f^*_{n,k}}w(x) \leq \inf_{f \in N_n(k,x)} T_f w(x). \tag{3.9}$$

Thus, we can select $g' \in N_n(k,x)$ in (3.9) as $f^*_{n,k+1}$ in (3.6).

From the above construction for the sequence $\{f^*_{n,k}\}_{k=0,1,2,\cdots} \subset N(C|S)$, we have

$$0 \leq \varepsilon\beta\rho(f^*_{n,k}, f^*_{n,k+1})(x) \leq T_{f^*_{n,k}}w(x) - T_{f^*_{n,k+1}}w(x) \tag{3.10}$$

for $k = 0, 1, 2, \cdots$. Adding them up for $m > k$, we get for all $m > k$

$$\begin{aligned} \varepsilon\beta\rho(f^*_{n,k}, f^*_{n,m})(x) &\leq \varepsilon\beta \sum_{j=k}^{m-1} \rho(f^*_{n,j}, f^*_{n,j+1})(x) \\ &\leq \sum_{j=k}^{m-1} \left\{T_{f^*_{n,j}}w(x) - T_{f^*_{n,j+1}}w(x)\right\} \\ &= T_{f^*_{n,k}}w(x) - T_{f^*_{n,m}}w(x). \end{aligned} \tag{3.11}$$

From (3.10), the sequence $\{T_{f^*_{n,k}}w(x)\}_{k=0,1,2,\cdots}$ is decreasing. Further, from $r, w \in M(S)$, it is bounded from below. Therefore, $\{T_{f^*_{n,k}}w(x)\}_{k=0,1,2,\cdots}$ converges. Therefore, the right-hand side of the inequality (3.11) goes to zero as $k, m \to \infty$. This shows that $\{f^*_{n,k}(x)\}_{k=0,1,2,\cdots}$ is a Cauchy sequence on $U(x)$ in C. Since the control space C is complete, the sequence of controls $\{f^*_{n,k}(x)\}_{k=0,1,2,\cdots}$ converges to some limit $f^*_n(x)$ in the set $U(x)$ for each $x \in S$. Since each $f^*_{n,k} \in N(C|S)$ is universally measurable on S, f^*_n is also universally measurable on S. See Dudley [6, Theorem 4.2.2] for more details of measurability. Further, since $U(x)$ is closed, $f^*_n(x)$ belongs to $U(x)$, that is, f^*_n is a selector of U.

Letting $k = 0$ in (3.11), we obtain, for $m, n = 1, 2, \cdots$

$$\varepsilon\beta\rho(f^\varepsilon_n, f^*_{n,m})(x) \leq T_{f^\varepsilon_n}w(x) - T_{f^*_{n,m}}w(x) \tag{3.12}$$

Since $T_f w(x)$ is l.s.c. on $U(x)$ by (A2) and (A3), from (3.12) it follows that

$$\begin{aligned} T_{f^*_n}w(x) &\leq \liminf_{m\to\infty} T_{f^*_{n,m}}w(x) \\ &= \lim_{m\to\infty} T_{f^*_{n,m}}w(x) \\ &\leq \lim_{m\to\infty} \{T_{f^\varepsilon_n}w(x) - \varepsilon\beta\rho(f^\varepsilon_n, f^*_{n,m})(x)\} \\ &= T_{f^\varepsilon_n}w(x) - \varepsilon\beta\rho(f^\varepsilon_n, f^*_n)(x). \end{aligned} \tag{3.13}$$

Thus, (3.13) shows that (3.4) holds.

Next, in order to show that the strictly inequality (3.5) holds, for any fixed $x \in S$, we consider two cases : (i) there exists an integer N such that $N_n(N,x) = \phi$, (ii) $N_n(k,x) \neq \phi$ for all k.

In case (i), for any $f \in N(C|S)$ with $\rho(f, f_n^*)(x) > 0$, $T_f w(x) > T_{f_{n,N}^*} w(x) - \varepsilon\beta\rho(f, f_{n,N}^*)(x)$. Since $f_{n,N}^*(x) = f_{n,k}^*(x)$ for all $k \geq N$, we have $f_{n,N}^*(x) = \lim_{k\to\infty} f_{n,k}^*(x) = f_n^*(x)$. Thus we obtain (3.5).

In case (ii), suppose that there exists $\tilde{f} \in N(C|S)$ such that $\tilde{f}(x) \neq f_n^*(x)$ and

$$T_{\tilde{f}} w(x) \leq T_{f_n^*} w(x) - \varepsilon\beta\rho(f_n^*, \tilde{f})(x). \tag{3.14}$$

Taking the limit as $m \to \infty$ in (3.11), we have

$$\varepsilon\beta\rho(f_{n,k}^*, f_n^*)(x) \leq T_{f_{n,k}^*} w(x) - \lim_{m\to\infty} T_{f_{n,m}^*} w(x) \tag{3.15}$$

for $k = 0, 1, 2, \cdots$. Taking account of (3.14) and (3.15), we have for $k = 0, 1, 2, \cdots$

$$\begin{aligned} T_{\tilde{f}} w(x) &\leq T_{f_n^*} w(x) - \varepsilon\beta\rho(f_n^*, \tilde{f})(x) \\ &\leq \lim_{m\to\infty} T_{f_{n,m}^*} w(x) - \varepsilon\beta\rho(f_n^*, \tilde{f})(x) \\ &\leq T_{f_{n,k}^*} w(x) - \varepsilon\beta\rho(f_{n,k}^*, f_n^*)(x) - \varepsilon\beta\rho(f_n^*, \tilde{f})(x) \\ &\leq T_{f_{n,k}^*} w(x) - \varepsilon\beta\rho(f_{n,k}^*, \tilde{f})(x). \end{aligned} \tag{3.16}$$

Here, we have $\tilde{f}(x) \neq f_{n,k}^*(x)$ for $k = 0, 1, 2, \cdots$. Because, if $\tilde{f}(x) = f_{n,k_0}^*(x)$ holds for some integer k_0, then (3.16) shows

$$\begin{aligned} T_{\tilde{f}} w(x) &\leq T_{f_{n,k_0}^*} w(x) - \varepsilon\beta \left\{ \rho(f_{n,k_0}^*, f_n^*)(x) + \rho(f_n^*, \tilde{f})(x) \right\} \\ &= T_{\tilde{f}} w(x) - 2\varepsilon\beta\rho(f_n^*, \tilde{f})(x), \end{aligned}$$

which contradicts $\tilde{f}(x) \neq f_n^*(x)$. Thus, we obtain $\tilde{f} \in N_n(k,x)$, $k = 0, 1, 2, \cdots$. From the facts that $N_n(k,x) \neq \phi$ and that the function $f_{n,k+1}^* \in N_n(k,x)$ satisfies the inequality (3.6) for $k = 0, 1, 2, \cdots$, we have

$$2T_{f_{n,k+1}^*} w(x) - T_{f_{n,k}^*} w(x) \leq T_{\tilde{f}} w(x),$$

for $k = 0, 1, 2, \cdots$. Taking the limit as $k \to \infty$, we have $\lim_{k\to\infty} T_{f_{n,k}^*} w(x) \leq T_{\tilde{f}} w(x)$. By (3.13) and (3.14), we have

$$\begin{aligned} T_{f_n^*} w(x) &\leq \lim_{k\to\infty} T_{f_{n,k}^*} w(x) \\ &\leq T_{\tilde{f}} w(x) \\ &\leq T_{f_n^*} w(x) - \varepsilon\beta\rho(f_n^*, \tilde{f})(x) \\ &< T_{f_n^*} w(x), \end{aligned}$$

which is a contradiction. Thus, (3.5) holds and the proof of the lemma is complete. □

For any given nonrandomized ε-optimal policy $\pi_\varepsilon = (f_0^\varepsilon, f_1^\varepsilon, \cdots) \in \Pi_N$, we can get the following theorem.

THEOREM 3.4. *For any given ε-optimal policy $\pi_\varepsilon \in \Pi_N$, there exists a nonrandomized Markov policy $\pi^* = (f_0^*, f_1^*, \cdots) \in \Pi_N$ such that for all initial state $x \in S$,*

$$I(\pi^*)(x) + \varepsilon \sum_{k=0}^{\infty} \beta^{k+1} E_{\pi_\varepsilon} [\rho(f_k^*, f_k^\varepsilon)(x_k)|x_0 = x] \leq I(\pi_\varepsilon)(x) \tag{3.17}$$

and

$$\sum_{k=0}^{\infty} \beta^{k+1} E_{\pi_\varepsilon} [\rho(f_k^*, f_k^\varepsilon)(x_k)|x_0 = x] \leq 1. \tag{3.18}$$

Proof Let the stationary policy $\pi_\varepsilon = (f_0^\varepsilon, f_1^\varepsilon, \cdots)$ be an ε-optimal nonrandomized Markov policy such that for every initial state $x \in S$,

$$I(\pi_\varepsilon)(x) < I^*(x) + \varepsilon. \tag{3.19}$$

For the control model in cases (D) and (P), it follows from (3.3) that

$$I(\pi_\varepsilon)(x) = \lim_{t \to \infty} I_t(\pi_\varepsilon)(x), \tag{3.20}$$

where, starting from the zero function $v_0 \in M(S)$ yields the corresponding t-stage cost function $I_t(\pi_\varepsilon)(x)$ as follows

$$I_t(\pi_\varepsilon)(x) = (T_{f_0^\varepsilon}, T_{f_1^\varepsilon}, \cdots, T_{f_{t-1}^\varepsilon})v_0(x).$$

From (3.19) and (3.20), there exists a sufficiently large integer $N > 0$ such that for all $n \geq N$ and $x \in S$,

$$I_n(\pi_\varepsilon)(x) < I^*(x) + \varepsilon.$$

Thus, at first, we need to show that the results are valid for a sufficiently large integer $n \geq N$. For such an integer n, from Lemma 3.3 we have a selector $f_n^* \in N(C|S)$ such that for all $x \in S$,

$$T_{f_n^*} v_0(x) + \varepsilon\beta\rho(f_n^*, f_n^\varepsilon)(x) \leq T_{f_n^\varepsilon} v_0(x). \tag{3.21}$$

Since the operator $T_{f_n^\varepsilon}$ is monotone, operating $T_{f_n^\varepsilon}$ on both sides of (3.21), we have

$$T_{f_{n-1}^\varepsilon} \left(T_{f_n^*} v_0 + \varepsilon\beta\rho(f_n^*, f_n^\varepsilon)\right)(x) \leq T_{f_{n-1}^\varepsilon} T_{f_n^\varepsilon} v_0(x). \tag{3.22}$$

The left-side of (3.22) is rewritten as follows

$$T_{f_{n-1}^\varepsilon} T_{f_n^*} v_0(x) + \varepsilon\beta^2 E_{f_{n-1}^\varepsilon} [\rho(f_n^*, f_n^\varepsilon)(x_n)|x_{n-1} = x] \leq T_{f_{n-1}^\varepsilon} T_{f_n^\varepsilon} v_0(x). \tag{3.23}$$

Then, applying Lemma 3.3 to the first term of the left-side of (3.23), it follows that there exists $f_{n-1}^* \in N(C|S)$ such that

$$T_{f_{n-1}^*} T_{f_n^*} v_0(x) + \varepsilon\beta\rho(f_{n-1}^*, f_{n-1}^\varepsilon)(x) \leq T_{f_{n-1}^\varepsilon} T_{f_n^*} v_0(x). \tag{3.24}$$

Combining (3.23) and (3.24), we obtain

$$T_{f^*_{n-1}} T_{f^*_n} v_0(x) + \varepsilon\beta\rho(f^*_{n-1}, f^\varepsilon_{n-1})(x) + \varepsilon\beta^2 E_{f^\varepsilon_{n-1}} [\rho(f^*_n, f^\varepsilon_n)(x_n)|x_{n-1} = x] \\ \leq T_{f^\varepsilon_{n-1}} T_{f^\varepsilon_n} v_0(x). \tag{3.25}$$

Therefore, a successive operation yields that for each $x \in S$,

$$T_{f^*_0} T_{f^*_1} \cdots T_{f^*_n} v_0(x) + \sum_{k=0}^{n} \varepsilon\beta^{k+1} E_{f^\varepsilon_0 f^\varepsilon_1 \cdots f^\varepsilon_{k-1}} [\rho(f^*_k, f^\varepsilon_k)(x_k)|x_0 = x] \leq I_{n+1}(\pi_\varepsilon)(x). \tag{3.26}$$

Then, letting $n \to \infty$ in (3.26), we get for each $x \in S$,

$$I(\pi^*)(x) + \sum_{k=0}^{\infty} \varepsilon\beta^{k+1} E_{\pi_\varepsilon} [\rho(f^*_k, f^\varepsilon_k)(x_k)|x_0 = x] \leq I(\pi_\varepsilon)(x). \tag{3.27}$$

Here, from Definition 2.1, we have the following inequalities for the given ε-optimal policy π_ε,

$$I^*(x) \leq I(\pi^*)(x)$$

and

$$I(\pi_\varepsilon)(x) \leq I^*(x) + \varepsilon.$$

Applying the preceding two inequalities to (3.27), we have for each $x \in S$,

$$I^*(x) + \sum_{k=0}^{\infty} \varepsilon\beta^{k+1} E_{\pi_\varepsilon} [\rho(f^*_k, f^\varepsilon_k)(x_k)|x_0 = x] \leq I^*(x) + \varepsilon. \tag{3.28}$$

Since $I^*(x)$ is finite, from (3.28) it follows that

$$\sum_{k=0}^{\infty} \beta^{k+1} E_{\pi_\varepsilon} [\rho(f^*_k, f^\varepsilon_k)(x_k)|x_0 = x] \leq 1.$$

This shows that (3.18) holds. Hence the proof is complete.

□

Note : This theorem shows that there exists a better policy $\pi^* \in \Pi_N$ than a given ε-optimal one $\pi_\varepsilon \in \Pi_N$ and, further, shows that π^* is near to π_ε in a sense of (3.18).

4. The minimization problem of a modified control system

The better policy $\pi^* = (f^*_0, f^*_1, \cdots) \in \Pi_N$ given in Theorem 3.4 depends only on the given ε-optimal policy $\pi_\varepsilon \in \Pi_N$ and $v_0 \in M(S)$. Using the policy $\pi^* \in \Pi_N$, we introduce the *modified one-stage cost functions* $m_n : GrU \to R^*$, $n = 0, 1, 2, \cdots$ as follows : for each $(x, u) \in GrU$,

$$m_n(x, u) = r(x, u) + \varepsilon\beta\rho(f^*_n(x), u).$$

In this section, we consider a modified stochastic control model given by a six-tuple

$$(S, C, U, q, m_{(\cdot)}, \beta),$$

where, S, C, U, q, β are the same ones as described in Section 2. The cost function $m_{(\cdot)}$ is perturbed by control and selector in the policy π^*. As in Section 3, we will treat cases (D) and (P). We impose the assumptions (A1), (A2) and (A3) on the control model.

If we use a policy $\pi \in \Pi$, the *total expected discounted cost for the modified control system* is defined by

$$M(\pi) = \sum_{t=0}^{\infty} \beta^t E_\pi \left[m_t(x_t, u_t) \right], \tag{4.1}$$

where

$$E_\pi \left[m_t(x_t, u_t) \right] = \int_{S_0 C_0 \cdots S_t C_t} m_t(x_t, u_t) P_{t+1}^{\pi,p}(dh_t).$$

Further, we define $M(\pi)(x)$ in the same way as $I(\pi)(x)$ in Section 2.

Then, we consider the following modified minimization problem (MM$\overline{\text{P}}$) :

$$(\text{MM}\overline{\text{P}}) \qquad \text{minimize} \quad M(\pi)(x) \quad \text{subject to} \quad \pi \in \Pi_N.$$

For the modified cost functions m_n, $n = 0, 1, 2, \cdots$ and each selector $f \in N(C|S)$, we define an operator $S_{n,f}$ on $M(S)$ as follows : for each $w \in M(S)$ and $x \in S$,

$$S_{n,f} w(x) = m_n(x, f(x)) + \beta \int_S w(x') q(dx'|x, f(x)).$$

That is,

$$S_{n,f} w(x) = T_f w(x) + \varepsilon \beta \rho(f_n^*, f)(x). \tag{4.2}$$

Since $m_n(\cdot, f(\cdot))$ is bounded from below, we have for each $\pi = (f_0, f_1, f_2, \cdots) \in \Pi_N$,

$$M(\pi)(x) = \lim_{n \to \infty} \left(S_{0,f_0} S_{1,f_1} \cdots S_{n,f_n} \right) v_0(x). \tag{4.3}$$

THEOREM 4.1. *The policy $\pi^* = (f_0^*, f_1^*, \cdots) \in \Pi_N$ given in Theorem 3.4 for an ε-optimal policy $\pi_\varepsilon \in \Pi_N$ is optimal for the modified minimization problem* (MM$\overline{\text{P}}$). *That is, for every initial state $x \in S$,*

$$M(\pi^*)(x) = \inf_{\pi \in \Pi_N} M(\pi)(x). \tag{4.4}$$

Proof Let $v_0 \in M(S)$ be identically zero. Using (3.5) in Lemma 3.3, for all $f_n \in N(C|S)$, we have

$$T_{f_n} v_0(x) \geq T_{f_n^*} v_0(x) - \varepsilon \beta \rho(f_n^*, f_n)(x).$$

From (4.2), we have for all $f_n \in N(C|S)$,

$$T_{f_n} v_0(x) = S_{n,f_n} v_0(x) - \varepsilon\beta\rho(f_n^*, f_n)(x).$$

In particular, $T_{f_n^*} v_0(x) = S_{n,f_n^*} v_0(x)$. Therefore, we obtain for all $f_n \in N(C|S)$,

$$S_{n,f_n} v_0(x) \geq S_{n,f_n^*} v_0(x). \tag{4.5}$$

Again, using (3.5), we get

$$T_{f_{n-1}} T_{f_n^*} v_0(x) \geq T_{f_{n-1}^*} T_{f_n^*} v_0(x) - \varepsilon\beta\rho(f_{n-1}^*, f_{n-1})(x) \tag{4.6}$$

for all $f_{n-1} \in N(C|S)$. Operating $T_{f_n^*} v_0 = S_{n,f_n^*} v_0$ on (4.6), we have

$$T_{f_{n-1}} S_{n,f_n^*} v_0(x) \geq T_{f_{n-1}^*} S_{n,f_n^*} v_0(x) - \varepsilon\beta\rho(f_{n-1}^*, f_{n-1})(x).$$

Further, from the definition of $S_{(\cdot,\cdot)}$,

$$T_{f_{n-1}} S_{n,f_n^*} v_0(x) = S_{n-1,f_{n-1}} S_{n,f_n^*} v_0(x) - \varepsilon\beta\rho(f_{n-1}^*, f_{n-1})(x)$$

for all $f_{n-1} \in N(C|S)$. In particular, $T_{f_{n-1}^*} S_{n,f_n^*} v_0(x) = S_{n-1,f_{n-1}^*} S_{n,f_n^*} v_0(x)$. Thus, for all $f_{n-1} \in N(C|S)$, we have

$$S_{n-1,f_{n-1}} S_{n,f_n^*} v_0(x) \geq S_{n-1,f_{n-1}^*} S_{n,f_n^*} v_0(x). \tag{4.7}$$

Therefore, a successive operation yields that for all $f_0, f_1, \cdots, f_n \in N(C|S)$,

$$S_{0,f_0} S_{1,f_1} \cdots S_{n,f_n} v_0(x) \geq S_{0,f_0^*} S_{1,f_1^*} \cdots S_{n,f_n^*} v_0(x). \tag{4.8}$$

Taking the limit as $n \to \infty$ in (4.8), we have $M(\pi) \geq M(\pi^*)$ for all $\pi = (f_0, f_1, \cdots, f_n, \cdots) \in \Pi_N$, which shows that (4.4) holds. Thus, the proof is complete. □

Acknowledgments

The authors would like to thank the referee for his valuable comments.

References

[1] J.P. Aubin, Mathematical Methods of Game and Economic Theory, Revised Edition, North-Holland, Amsterdam, 1982.

[2] E.J. Badler, On compactness of the space of policies in stochastic dynamic programming, Stochastic Proc. Appl. 32 (1989) 141-150.

[3] D.P. Bertsekas and S.E. Shreve, Stochastic Optimal Control: The Discrete Time Case, Academic Press, New York, 1978.

[4] D. Blackwell, Discrete dynamic programming, Ann. Math. Statist. 33 (1962) 719-726.

[5] D. Blackwell, Discounted dynamic programming, Ann. Math. Statist. 36 (1965) 226-235.

[6] R.M. Dudley, Real Analysis and Probability, Wadsworth & Brooks, 1989.

[7] E.B. Dynkin, Markov Processes-I, Springer-Verlag, Berlin, 1965.

[8] E.B. Dynkin and A.A. Yushkevich, Controlled Markov Processes, Springer-Verlag, Berlin, 1979.

[9] I. Ekeland, On the variational principle, J.Math.Anal.Appl., 47 (1974) 324-353.

[10] I. Ekeland, Nonconvex minimization problems, Bull. Amer. Math., 47 (1979) 443-474.

[11] K. Hinderer, Foundations of non-stationary dynamic programming with discrete time parameter, Lecture Notes on Operations Research and Mathematical Systems 33, Springer-Verlag, Berlin, 1970.

[12] M. Schäll, On dynamic programming: Compactness of the space of policies, Stochastic Processes Appl. 3 (1975) 345-364.

[13] M. Schäll, On dynamic programming and statistical decision theory, Ann. Statist. 7 (1979) 432-445.

[14] R. Strauch, Negative dynamic programming, Ann. Math. Statist., 37 (1966) 871-890.

[15] K. Tanaka, On discounted dynamic programming with constraints, J. Math. Anal. Appl. 155 (1991) 264-277.

Received June 15, 1994
Revised October 4, 1994