ACADEMIC PRESS

J. Math. Anal. Appl. 271 (2002) 66–81

Journal of MATHEMATICAL ANALYSIS AND APPLICATIONS

www.academicpress.com

Optimal policy for minimizing risk models in Markov decision processes

Y. Ohtsubo [a,*] and K. Toyonaga [b]

[a] *Department of Mathematics, Faculty of Science, Kochi University, Kochi 780-8520, Japan*
[b] *Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan*

Received 15 March 1999

Submitted by J. Filar

Abstract

We consider the minimizing risk problems in discounted Markov decisions processes with countable state space and bounded general rewards. We characterize optimal values for finite and infinite horizon cases and give two sufficient conditions for the existence of an optimal policy in an infinite horizon case. These conditions are closely connected with Lemma 3 in White (1993), which is not correct as Wu and Lin (1999) point out. We obtain a condition for the lemma to be true, under which we show that there is an optimal policy. Under another condition we show that an optimal value is a unique solution to some optimality equation and there is an optimal policy on a transient set. © 2002 Elsevier Science (USA). All rights reserved.

Keywords: Markov decision process; Minimizing risk model; Maximal fixed point; Existence of optimal policy

1. Introduction

In the area of Markov decision processes, standard criteria have been the expected discounted total reward over a finite or an infinite time horizon, or the av-

* Corresponding author.
E-mail addresses: ohtsubo@math.kochi-u.ac.jp (Y. Ohtsubo), toyonaga@math.kyushu-u.ac.jp (K. Toyonaga).

PII: S0022-247X(02)00097-5

erage expected reward per unit over an infinite time horizon (e.g., see Derman [1], Hordijk [2], Puterman [3] and White [4]). To make up for insufficiency of these criteria, other criteria (e.g., utility, probabilistic constraints and mean-variance criteria) have been proposed and investigated by many authors (see White [5] for survey, or [6], Filar et al. [7,8] and Kadota et al. [9]).

As a special case of utility criteria, several authors [10–14] consider a problem in which we minimize a threshold probability $P_s^\pi(Z \leqslant r)$ with respect to policy π in Markov decision processes, where Z is a discounted total reward, r is a threshold (target) value and s is an initial state. In [13], White investigates finite Markov decision processes with a bounded reward set and shows that optimal value function is a unique solution to optimal equation and there exists an optimal policy. However, by giving a counterexample, Wu and Lin [14] point out that threshold probability, which is generated by a policy, an initial state and a threshold value, is not necessarily a distribution of threshold value, and that Lemma 3 in [13] does not hold in general case and hence the existence of an optimal policy has not been proved really. They prove that the optimal value functions for finite and infinite horizon cases are distributions of threshold value. They also show by measurable selection theorem that there exists an optimal deterministic Markov policy in a finite horizon model, but point out that the existence of an optimal policy in an infinite horizon case is open. Instead of existence theorem, they give a sufficient and necessary condition for a policy independent of a threshold value to be optimal in finite and infinite cases.

In this paper, we concern ourselves with such a problem in discounted Markov decision processes with countable state space and nonnegative bounded general rewards. In Section 2, we give notations and formulate our problem. In Section 3, we show that the threshold probability is measurable with respect to a threshold value, and give slight extensions of [10] and [14] for properties of an optimal operator and optimal values in finite and infinite horizon cases. Especially, there exists a right continuous optimal policy in a finite horizon case and an optimal value function in an infinite horizon case is a maximal fixed point of the operator. In Section 4, we give two sufficient conditions for the existence of an optimal policy in an infinite horizon case. These are closely related with Lemma 3 in White [13], which is not correct as Wu and Lin [14] point out. We first obtain a sufficient condition for the lemma to be true and we show that there is an optimal policy under the condition. We show under another condition that an optimal value on some transient set is a unique solution to some optimality equation and there exists an optimal policy.

2. Notations and formulation

Discounted Markov decision processes $\Gamma = ((X_t), (A_t), (Y_t), p, \beta)$ with a discrete time space $N = \{1, 2, \ldots\}$ are defined by the following: the state space S

is a countable set and denote the state at time $t \in N$ by X_t; the action space $A = \bigcup_{s\in S} A(s)$ is countable, where $A(s)$ is a nonempty set of admissible finite actions when the system is in state $s \in S$, and denote the action at time $t \in N$ by A_t; the reward space $R_H = [H', H]$ is a bounded interval where H' is a nonpositive constant and H is positive, and $Y_t \in R_H$ is a random immediate reward function at time $t \in N$, which is defined by a conditional probability for (X_{t+1}, Y_t) given (X_t, A_t):

$$p^a(s', y \mid s) = P(X_{t+1} = s', Y_t \leqslant y \mid X_t = s, A_t = a);$$

the discount factor β is a real number so that $0 < \beta < 1$. We use $S_R = S \times R$ as a new state space where $R = (-\infty, \infty)$. For convenience sake, we assume that $H' = 0$, because we can extend our results to a model with $H' < 0$, by the method similar to that in Wu and Lin [14].

Let $H_1 = S_R$ and $H_{t+1} = H_t \times A \times S \times R_H$ for each $t \in N$. Then H_t is the set of all possible histories of the system when the tth action must be chosen, and denote by θ_t the history at time t. A decision rule δ_t for time $t \in N$ is a conditional probability given θ_t: $\delta_t(a_t \mid h_t) = P(A_t = a_t \mid \theta_t = h_t)$, where $h_t = (s_1, r, a_1, s_2, y_1, \ldots, a_{t-1}, s_t, y_{t-1}) \in H_t$ which is a realising value of $\theta_t = (X_1, r, A_1, X_2, Y_1, \ldots, A_{t-1}, X_t, Y_{t-1})$ and r is a given real number. It is assumed that $\delta_t(A_t \in A(s_t) \mid h_t) = 1$ for every history $h_t = (s_1, r, a_1, \ldots, s_t, y_{t-1}) \in H_t$ and $\delta_t(a_t \mid \cdot)$ is a Lebesgue–Stieltjes measurable function on H_t. We denote by Δ the set of all decision rules. A policy π is an infinite sequence of decision rules $(\delta_1, \delta_2, \ldots, \delta_t, \ldots)$. We denote by C the set of all such policies.

We define the random total discounted rewards for a finite horizon case $(\mathcal{P}_n)$ and an infinite horizon case $(\mathcal{P})$ by

$$Z_0 = 0, \qquad Z_n = \sum_{t=1}^{n} \beta^{t-1} Y_t, \quad n \geqslant 1,$$

and

$$Z = \sum_{t=1}^{\infty} \beta^{t-1} Y_t,$$

respectively. We define a new random sequence by

$$W_1 = r, \qquad W_n = (W_1 - Z_{n-1})/\beta^{n-1} = (W_{n-1} - Y_{n-1})/\beta, \quad n \geqslant 2,$$

which depend upon $(Y_1, Y_2, \ldots, Y_{n-1})$, where r is a given real number. Then we are able to replace histories $h_t = (s_1, r, a_1, s_2, y_1, \ldots, a_{t-1}, s_t, y_{t-1})$ by $h'_t = (s_1, w_1, a_1, s_2, w_2, \ldots, a_{t-1}, s_t, w_t)$ which is a realising value of $(X_1, W_1, A_1, X_2, W_2, \ldots, A_{t-1}, X_t, W_t)$. Hence a decision rule and a policy may be equivalently defined for the new histories. A policy $\pi = (\delta_1, \delta_2, \ldots, \delta_t, \ldots)$ is said to be Markov when the decision rule δ_t for the new histories is a function of $(X_t, W_t) = (s_t, w_t)$ for every $t \in N$, and we denote by Δ_M the set of such decision rules. Also,

a policy π is called a deterministic Markov policy if π is Markov and δ_t determines some action $a_t \in A(s_t)$ deterministically, and we denote by Δ_D the set of such decision rules. We may consider $\delta \in \Delta_D$ as a mapping from S_R into A, and hence when $\delta(a \mid s, r) = 1$ we denote it by $\delta(s, r) = a$. We denote the set of all Markov policies by C_M, and the set of all deterministic Markov policies by C_D. When $\pi = (\delta, \delta, \ldots, \delta, \ldots) \in C_D$, we write $\pi = \delta^\infty$, which is called a stationary policy, and we denote the set of stationary policies by C_D^s.

A decision rule $\delta \in \Delta_D$ from S_R into A is said to be right continuous (on R) if for every $(s, r) \in S_R$, there is a positive real number μ such that $\delta(s, r) = \delta(s, r + u)$ for all u: $0 \leqslant u < \mu$. A policy $\pi = (\delta_t) \in C_D$ is said to be right continuous if the decision rule δ_t is right continuous for every $t \in N$.

We denote by $P_s^\pi(Z \leqslant r)$ the conditional probability of event $\{Z \leqslant r\}$ given an initial state $X_1 = s$ and a policy π. Here a random variable Z depends upon not only s and π but also r. For the sake of convenience, we consider optimization problems only on a set C_M of Markov policies.

We define criterion functions for $(\mathcal{P}_n)$ and $(\mathcal{P})$ by

$$F_n^\pi(s, r) = P_s^\pi(Z_n \leqslant r), \qquad F^\pi(s, r) = P_s^\pi(Z \leqslant r),$$

respectively, for each $(s, r) \in S_R$ and $\pi \in C_M$. Wu and Lin [14] give an example in which $F_n^\pi(s, r)$ and $F^\pi(s, r)$ are not distribution functions of r. We also denote optimal value functions F_n^* and F^* for $(\mathcal{P}_n)$ and $(\mathcal{P})$ by, respectively,

$$F_n^*(s, r) = \inf_{\pi \in C_M} F_n^\pi(s, r), \qquad F^*(s, r) = \inf_{\pi \in C_M} F^\pi(s, r).$$

A policy $\pi \in C_M$ is said to be optimal in $(\mathcal{P}_n)$ if $F_n^*(s, r) = F_n^\pi(s, r)$ for every $(s, r) \in S_R$. Similarly, we define an optimal policy in an infinite horizon case $(\mathcal{P})$.

We define the following sets of functions: let $\mathcal{F}$ be the set of functions F from S_R into interval $[0, 1]$ such that for each $s \in S$, $F(s, \cdot)$ is measurable on R, and let $\mathcal{F}_m$ be the set of functions $F \in \mathcal{F}$ such that for each $s \in S$, $F(s, r) = 0$ if $r < 0$ and $F(s, r) = 1$ if $r > H/(1 - \beta)$. Also, let $\mathcal{F}_r$ be the set of functions $F \in \mathcal{F}_m$ such that for each $s \in S$, $F(s, \cdot)$ is nondecreasing and right continuous on R.

We define operators T^a, T^δ and T from $\mathcal{F}$ into $\mathcal{F}$ as follows: for $F \in \mathcal{F}$, $(s, r) \in S_R$, $a \in A(s)$ and $\delta \in \Delta$,

$$T^a F(s, r) = \int_{S_R} F\big(s', (r - y)/\beta\big)\, dp^a(s', y \mid s),$$

$$T^\delta F(s, r) = \sum_{a \in A(s)} T^a F(s, r) \delta(a \mid s, r),$$

$$T F(s, r) = \inf_{\delta \in \Delta} T^\delta F(s, r) = \min_{a \in A(s)} T^a F(s, r).$$

Here we see from Fubini's theorem that

$$\int_{S_R} F(s',\cdot)\,dp^a(s',y\mid s)=\sum_{s'\in S} q^a(s'\mid s)\int_R F(s',\cdot)\,p(dy\mid s,a,s'),$$

where $q^a(s'\mid s)=\int_{y\in R} p^a(s',dy\mid s)$ and $p(y\mid s,a,s')=p^a(s',y\mid s)/q^a(s'\mid s)$. We also define operators T^n and $T^1=T$ and $T^{n+1}=T(T^n)$ for each $n\geqslant 1$. Similarly, $(T^\delta)^n$ is defined for $\delta\in\Delta_M$.

In all argument, we use notations $=$, $\geqslant$ and $\leqslant$ for simultaneous equalities or inequalities; e.g., for $F,G\in\mathcal{F}$, $F\geqslant G$ means that $F(s,r)\geqslant G(s,r)$ for all $(s,r)\in S_R$. We give fundamental lemmas below.

Lemma 2.1.

(i) *For $F,G\in\mathcal{F}$ and $\delta\in\Delta$, $T^\delta F-T^\delta G=T^\delta(F-G)$.*
(ii) *If $F,G\in\mathcal{F}$ and $F\geqslant G$, then $T^aF\geqslant T^aG$, $T^\delta F\geqslant T^\delta G$ and $TF\geqslant TG$ for each $a\in A(\cdot)$ and $\delta\in\Delta$.*
(iii) *If $G\in\mathcal{F}_r$, then $T^aG(s,\cdot)$ is nondecreasing and right continuous on R for each $s\in S$ and any $a\in A(s)$. Also, T is an operator from $\mathcal{F}$, $\mathcal{F}_m$ or $\mathcal{F}_r$ into itself.*
(iv) *If $G_n\in\mathcal{F}_r$ and $G_n\geqslant G_{n+1}$ for each $n\geqslant 0$, then $\lim_{n\to\infty}G_n\in\mathcal{F}_r$.*

Proof. The statements (i), (ii) and the former part of (iii) are immediate results of definitions. It is also easy to see that if $F\in\mathcal{F}_m$, then $TF(s,\cdot)$ is measurable on R, and $TF(s,r)=0$ when $r<0$ and $TF(s,r)=1$ when $r>H/(1-\beta)$, which imply that $TF\in\mathcal{F}_m$. We can similarly prove for $F\in\mathcal{F}_r$. Hence the latter part of (iii) is proved; (iv) is a result of Lemma 1 in White [13]. □

The following lemma is a result related to the existence of a right continuous decision rule.

Lemma 2.2. *For each $F\in\mathcal{F}_r$, there exists a right continuous decision rule $\delta\in\Delta_D$ satisfying $TF=T^\delta F$.*

Proof. Let $F\in\mathcal{F}_r$ and $(s,r)\in S_R$ be arbitrarily fixed. From Lemma 2.1, $T^aF(s,\cdot)$ is right continuous at r for each $a\in A(s)$. Since $A(s)$ is finite, we see that there exist $\mu>0$ and $a\in A(s)$ such that $TF(s,u)=T^aF(s,u)$ for all u satisfying $r\leqslant u<r+\mu$. For such an action a, if we define $\delta\in\Delta_D$ by $\delta(s,u)=a$ for every u so that $r\leqslant u<r+\mu$, then δ is right continuous and $TF(s,r)=T^\delta F(s,r)$. □

3. Fundamental results of optimal values

In this section, we characterize optimal values for a finite horizon case $(\mathcal{P}_n)$ and an infinite horizon case $(\mathcal{P})$ and investigate properties of the optimal op-

erator T. These results are slight extensions of Bouakiz and Kebir [10] and Wu and Lin [14]. We also show that there is a right continuous optimal policy for a finite horizon model $(\mathcal{P}_n)$.

To guarantee that $T^a F^\pi$ and TF^π are well defined, it is required that F^π is in $\mathcal{F}_m$; that is, $F^\pi(s,\cdot)$ is measurable on R for each $s \in S$.

Lemma 3.1. *Let $\pi \in C_M$ be arbitrary.*

(i) *For $n \geqslant 0$, $F_n^\pi \geqslant F_{n+1}^\pi \geqslant \lim_{n\to\infty} F_n^\pi = F^\pi$.*
(ii) *For each $n \geqslant 0$, $F_n^\pi \in \mathcal{F}_m$ and $F^\pi \in \mathcal{F}_m$.*
(iii) *For each $n \geqslant 1$, $F_n^\pi = T^\delta F_{n-1}^\tau$ and $F^\pi = T^\delta F^\tau$, where $\pi = (\delta, \tau)$. Especially, $F^\pi = T^\delta F^\pi$ when $\pi = \delta^\infty$ is a stationary policy.*

Proof. (i) This is proved by the same method as Lemma 4(i) in Wu and Lin [14].

(ii) To show that $F_n^\pi \in \mathcal{F}_m$, it suffices to prove that $F_n^\pi(s,\cdot)$ is measurable on R. Since $F_0^\pi(s,r) = I_{[0,\infty)}(r)$, where I_A is the indicator function on a set A, we see that $F_0^\pi(s,\cdot)$ is measurable for every $\pi \in C_M$ and each $s \in S$. We assume that $F_n^\tau(s,\cdot)$ is measurable for every $\tau \in C_M$ and each $s \in S$. It then follows from Lemma 2.1(iii) that for any $\pi = (\delta, \tau) \in C_M$,

$$T^\delta F_n^\tau(s,r) = \sum_{a\in A(s)} \delta(a \mid s,r) \int_{(s',y)\in S_R} F_n^\tau\left(s', \frac{r-y}{\beta}\right) dp^a(s', y \mid s)$$

is well defined and measurable at r. However, by Markov property, we have

$$\begin{aligned} T^\delta F_n^\tau(s,r) &= \sum_{a\in A(s)} \delta(a \mid s,r) \int_{(s',y)} P^\tau_{(s',(r-y)/\beta)}(y + \beta Z_n \leqslant r)\, dp^a(s', y \mid s) \\ &= P_s^\pi(Z_{n+1} \leqslant r) = F_{n+1}^\pi(s,r), \end{aligned}$$

since $\tau \in C_M$, where $P^\tau_{(s,r)}(B)$ is a conditional probability of an event B given an initial state (s,r) and a policy τ. Hence, $F_{n+1}^\pi(s,\cdot)$ is measurable. Thus, by induction, $F_n^\pi(s,\cdot)$ is measurable for every $n \geqslant 0$. Furthermore, it follows from (i) that $F^\pi(s,\cdot) = \lim_{n\to\infty} F_n^\pi(s,\cdot)$ is also measurable.

(iii) From proof of (ii), we have $F_{n+1}^\pi(s,r) = T^\delta F_n^\tau(s,r)$, when $\pi = (\delta,\tau)$. Similarly, it is easy to see that $F^\pi = T^\delta F^\tau$. □

Theorem 3.1.

(i) *For each $n \geqslant 0$, $F_n^* \in \mathcal{F}_r$ and $\{F_n^*,\ n \geqslant 0\}$ satisfies optimality equations*:

$$F_0^* = I_{[0,\infty)}, \qquad F_n^* = TF_{n-1}^*, \quad n \geqslant 1.$$

(ii) *For each $n \geqslant 0$, there exists a right continuous optimal policy $\pi \in C_D$ in $(\mathcal{P}_n)$.*

Proof. We prove this theorem by induction. When $n = 0$, we see that $F_0^*(s, r) = I_{[0,\infty)}(r) = F_0^\pi(s, r)$ for any right continuous policy $\pi \in C_D$, so this theorem holds. Assume that this theorem is true for $n = k$. Thus, $F_k^* \in \mathcal{F}_r$ and there exists a right continuous optimal policy $\sigma \in C_D$ such that $F_k^* = F_k^\sigma$. It follows from Lemma 2.2 that there exists a right continuous decision rule $\delta \in \Delta_D$ such that $TF_k^* = T^\delta F_k^*$, which implies that $\pi = (\delta, \sigma)$ is a right continuous policy in C_D. By the same argument as Theorem 1 in [14], we have $TF_k^* = F_{k+1}^* = F_{k+1}^\pi$. Hence, π is optimal in $(\mathcal{P}_{k+1})$, and from Lemma 2.1(iii), we have $F_{k+1}^* \in \mathcal{F}_r$. By induction, the proof of this theorem is complete. □

From Theorem 3.1, we have $F_n^* = T^n F_0^*$, $n \geqslant 1$.

Theorem 3.2.

(i) *For each* $n \geqslant 0$, $F_n^* \geqslant F_{n+1}^* \geqslant \lim_{n\to\infty} F_n^* = F^*$ *and* $F^* \in \mathcal{F}_r$.
(ii) F^* *satisfies optimality equation* $F^* = TF^*$.
(iii) *There exists a right continuous decision rule* $\delta \in \Delta_D$ *such that* $F^* = T^\delta F^*$.

Proof. (i) By the same way as Lemma 4(ii) in [14], we obtain that $F_n^* \geqslant F_{n+1}^*$ for each $n \geqslant 0$ and $F^* = \lim_{n\to\infty} F_n^*$. Also, since $F_n^* \in \mathcal{F}_r$, $n \geqslant 1$, by Theorem 3.1 and $F_n^* \geqslant F_{n+1}^*$ it follows from Lemma 2.1(iv) that $F^* \in \mathcal{F}_r$.

(ii) It follows from Lemma 3.1 that $F^\pi = T^\delta F^\tau \geqslant T^\delta F^* \geqslant TF^*$ for any $\pi = (\delta, \tau) \in C_M$. Hence we have $F^* \geqslant TF^*$. Conversely, for $(s, r) \in S_R$, it follows from Theorem 3.1 that $F_n^*(s, r) = TF_{n-1}^*(s, r) \leqslant T^a F_{n-1}^*(s, r)$ for any $a \in A(s)$. By (i) and dominated convergence theorem, we have $F^*(s, r) \leqslant T^a F^*(s, r)$ for any $a \in A(s)$, so $F^*(s, r) \leqslant \min_a T^a F^*(s, r) = TF^*(s, r)$. Therefore we obtain the desired equation $F^* = TF^*$.

(iii) From (i) and (ii), $F^* \in \mathcal{F}_r$ and F^* satisfies an equation $F^* = TF^*$. Thus Lemma 2.2 leads that there exists a right continuous decision rule $\delta \in \Delta_D$ such that $F^* = T^\delta F^*$. □

In the following theorem, we give characterization of optimal value function F^* for optimal operator T in $\mathcal{F}_m$ (but not $\mathcal{F}_r$).

Theorem 3.3.

(i) F^* *is the maximal fixed point of* T *in* $\mathcal{F}_m$.
(ii) *Let* $G \in \mathcal{F}_m$ *be a function such that* $G \geqslant F^*$. *Then* $\{T^n G\}$ *converges and* $\lim_{n\to\infty} T^n G = F^*$.

Proof. (i) Let $G \in \mathcal{F}_m$ be a fixed point of T. Then we have $G \leqslant F_0^* = I_{[0,\infty)}$ and $G = T^n G$. Hence it follows from Theorems 3.1 and 3.2 that

$$G = \lim_n T^n G \leqslant \lim_n T^n F_0^* = \lim_n F_n^* = F^*.$$

(ii) If G is a function in $\mathcal{F}_m$, then $G \leqslant F_0^*$ and hence $T^n G \leqslant T^n F_0^* = F_n^*$. Thus, we have $\limsup_n T^n G \leqslant \lim_n F_n^* = F^*$. Conversely, since $G \geqslant F^*$, we have $T^n G \geqslant T^n F^* = F^*$, so $\liminf_n T^n G \geqslant F^*$. Hence, combining with the previous inequality, we have $\lim_n T^n G = F^*$. □

Corollary 3.1.

(i) *For any policy* $\pi \in C_M$, $\lim_{n\to\infty} T^n F^\pi = F^*$.
(ii) F^* *is the unique fixed point of* T *in the class of functions dominating* F^*; *that is, if* $G = TG$, $G \in \mathcal{F}_m$ *and* $G \geqslant F^*$, *then* $G = F^*$.
(iii) *If there exist a policy* $\pi \in C_M$ *such that* $F^\pi = TF^\pi$, *then* π *is optimal in* $(\mathcal{P})$.

Proof. The statement (i) is an immediate result of Theorem 3.3(ii), since $F^\pi \geqslant F^*$. We easily see that Theorem 3.3(i) leads (ii). It also follows from (i) or (ii) that $F^\pi = F^*$ and hence π is optimal in $(\mathcal{P})$. □

4. Sufficient conditions for existence of optimal policy

Wu and Lin [14] give a counterexample for Lemma 5 in White [13] and point out that Lemma 3 in [13] does not hold. In this section we first give a sufficient condition for Lemma 3 and Lemma 5 in [13] to be true.

We define another conditional probability of an event $\{Z \leqslant z\}$, given an initial state $(X_1, W_1) = (s, r)$ and a policy π, by

$$F^\pi(s, r; z) = P^\pi_{(s,r)}(Z \leqslant z).$$

Here Z is a random variable which depends upon (s, r) and π, but not z. Thus $F^\pi(s, r; z)$ is a distribution function of z and $F^\pi(s, r; r) = F^\pi(s, r)$. We denote by C_D^* the set of all policies $\pi = \delta^\infty \in C_D^s$ for which there is a countable subset $E \subset R$ such that $F^\pi(s, r; z)$ is continuous at $z = r$ for every $(s, r) \in S \times E^c$, where E^c is the complement of E. We notice in Example 4.1 that there is a policy $\pi \in C_D^s - C_D^*$.

Lemma 4.1. *Let* $F, G \in \mathcal{F}_r$ *and let* $\pi = \delta^\infty \in C_D^*$. *If* $F - G \leqslant T^\delta(F - G)$, *then* $F \leqslant G$.

Proof. Let $(s, r) \in S_R$ be arbitrarily fixed and $F, G \in \mathcal{F}_r$. From Lemma 2.1, we have $F(s, r) - G(s, r) \leqslant (T^\delta)^n(F - G)(s, r)$ for any $n \geqslant 1$. Now, we prove by induction that

$$(T^\delta)^n(F - G)(s, r) \leqslant F_n^\pi(s, r; r) - F_n^\pi(s, r; r - b_n), \tag{1}$$

where $b_n = \beta^n H/(1 - \beta)$, $n \geqslant 1$. Since $(F - G)(s, r) = 0$ if $r < 0$ or $r \geqslant H/(1 - \beta)$, and $(F - G)(s, r) \leqslant 1$ otherwise, when $n = 1$, we have

$$
\begin{aligned}
T^{\delta}(F-G)(s,r) &= \int_{S_R} (F-G)\big(s',(r-y)/\beta\big)\,dp^{\delta(s,r)}(s',y\mid s)\\
&= \int_{S\times(r-b_1,r]} (F-G)\big(s',(r-y)/\beta\big)\,dp^{\delta(s,r)}(s',y\mid s)\\
&\leqslant \int_{S\times(r-b_1,r]} dp^{\delta(s,r)}(s',y\mid s)\\
&= P^{\pi}_{(s,r)}(r-b_1 < Z_1 \leqslant r)\\
&= F^{\pi}_1(s,r;r) - F^{\pi}_1(s,r;r-b_1).
\end{aligned}
$$

Hence, the inequality (1) is true for $n=1$. Assume that the inequality (1) holds for $n=k$. Then, by Markov property, we have

$$
\begin{aligned}
&(T^{\delta})^{k+1}(F-G)(s,r)\\
&\quad = T^{\delta}(T^{\delta})^{k}(F-G)(s,r)\\
&\quad = \int_{S_R} (T^{\delta})^{k}(F-G)\big(s',(r-y)/\beta\big)\,dp^{\delta(s,r)}(s',y\mid s)\\
&\quad \leqslant \int_{S_R} P^{\pi}_{(s',(r-y)/\beta)}\left(\frac{r-y}{\beta}-b_k < Z_k \leqslant \frac{r-y}{\beta}\right) dp^{\delta(s,r)}(s',y\mid s)\\
&\quad = P^{\pi}_{(s,r)}\left(\frac{r-Y_1}{\beta}-b_k < Z_k \leqslant \frac{r-Y_1}{\beta}\right)\\
&\quad = P^{\pi}_{(s,r)}(r-b_{k+1} < Z_{k+1} \leqslant r)\\
&\quad = F^{\pi}_{k+1}(s,r;r) - F^{\pi}_{k+1}(s,r;r-b_{k+1}).
\end{aligned}
$$

Hence, by induction, the inequality (1) holds for every n.

Now, all rewards are nonnegative, $F^{\pi}(s,r;z)$ is a distribution function of z for any $\pi\in C_M$ and $(s,r)\in S_R$ and $\{Z\leqslant z\}\subset\{Z_n\leqslant z\}\subset\{Z\leqslant z+b_n\}$, where Z and Z_n depend upon an initial state (s,r). Hence, we have

$$
F^{\pi}(s,r;r-b_n)\leqslant F^{\pi}_n(s,r;r-b_n)\leqslant F^{\pi}_n(s,r;r)\leqslant F^{\pi}(s,r;r+b_n).
$$

Thus we obtain

$$
\begin{aligned}
F(s,r)-G(s,r) &\leqslant (T^{\delta})^{n}(F-G)(s,r)\\
&\leqslant F^{\pi}(s,r;r+b_n)-F^{\pi}(s,r;r-b_n).
\end{aligned}
$$

We see from the definition of C^{*}_{D} that if $r\in E^{c}$, then the right-hand side of the above inequality tends to zero as $n\to\infty$, since $\lim_{n\to\infty} b_n = 0$ and $F^{\pi}(s,r;z)$ is continuous at $z=r$. Thus we have $F(s,r)\leqslant G(s,r)$ for every $(s,r)\in S\times E^{c}$. Since E is countable and F, G are right continuous, it follows from Lemma 1 in White [13] that $F(s,r)\leqslant G(s,r)$ for every $(s,r)\in S_R$. □

In [14], Wu and Lin point out that a policy $\pi = \delta^\infty \in C_D$ such that $F^* = T^\delta F^*$ is not necessarily optimal. In the following theorem, a sufficient condition for such a policy to be optimal is given.

Theorem 4.1.

(i) *If $\pi = \delta^\infty \in C_D^*$ and $F^\pi \in \mathcal{F}_r$, then F^π is the unique solution to $F = T^\delta F$ in $\mathcal{F}_r$.*
(ii) *If there is a policy $\pi = \delta^\infty \in C_D^*$ such that $F^\pi \in \mathcal{F}_r$ and $F^* = T^\delta F^*$, then π is an optimal policy in $(\mathcal{P})$.*

Proof. (i) From Lemma 3.1, we have $F^\pi = T^\delta F^\pi$. Let $F \in \mathcal{F}_r$ be a solution to $F = T^\delta F$. Then we have $F - F^\pi = T^\delta(F - F^\pi)$ and hence Lemma 4.1 implies that $F = F^\pi$. Thus F^π is the unique solution to $F = T^\delta F$ in $\mathcal{F}_r$.

(ii) From the above (i) it follows that F^π is the unique solution to $F = T^\delta F$ in $\mathcal{F}_r$. Thus, since $F^* = T^\delta F^*$, we have $F^* = F^\pi$, which implies that π is optimal. □

We give another sufficient condition for Lemma 3 in [13] to be true. A subset S_0 of S is said to be closed for a policy π if $P^\pi_{(s,r)}(X_2 \in S_0) = 1$ for every initial state $(s, r) \in S_0 \times R$, and S_0 is said to be reachable (with probability one) for π if

$$P^\pi_{(s,r)}\left(\bigcup_{n=2}^{\infty}\{X_n \in S_0\}\right) = P^\pi_{(s,r)}(X_n \in S_0 \text{ for some } n \geqslant 2) = 1$$

for every initial state $(s, r) \in S_R$.

Lemma 4.2. *Let $\pi = \delta^\infty \in C_D$. Suppose there is a subset S_0 of S such that S_0 is closed and reachable for π.*

(i) *Let $F, G \in \mathcal{F}_m$. If $F - G \leqslant T^\delta(F - G)$ on $S_0^c \times R$ and $F = G$ on $S_0 \times R$, then $F \leqslant G$.*
(ii) *F^π is the unique solution in $\mathcal{F}_m$ to equation $F = T^\delta F$ with $F = F^\pi$ on $S_0 \times R$.*

Proof. (i) Since $F = G$ on $S_0 \times R$ and $F - G \leqslant 1$, it follows from condition on S_0 that if $s \in S_0$, then

$$T^\delta(F - G)(s, r) = \int_{(s',y)\in S_0\times R} (F - G)\big(s', (r - y)/\beta\big)\, dp^{\delta(s,r)}(s', y \mid s) = 0,$$

and if $s \notin S_0$, then

$$
\begin{aligned}
T^{\delta}(F-G)(s,r) &= \int_{(s',y)\in S_0^c\times R} (F-G)\bigl(s',(r-y)/\beta\bigr)\,dp^{\delta(s,r)}(s',y\mid s)\\
&\leqslant \int_{(s',y)\in S_0^c\times R} dp^{\delta(s,r)}(s',y\mid s)\\
&= P^{\pi}_{(s,r)}\bigl(X_2\in S_0^c\bigr).
\end{aligned}
$$

For $n\geqslant 1$, assume that $(T^{\delta})^n(F-G)(s,r)=0$ if $s\in S_0$, and $(T^{\delta})^n(F-G)(s,r)\leqslant P^{\pi}_{(s,r)}(\bigcap_{k=2}^{n+1}\{X_k\in S_0^c\})$ otherwise. Then, it follows from Markov property that if $s\in S_0$, then

$$
\begin{aligned}
&(T^{\delta})^{n+1}(F-G)(s,r)\\
&\quad= T^{\delta}(T^{\delta})^n(F-G)(s,r)\\
&\quad= \int_{(s',y)\in S_0\times R} (T^{\delta})^n(F-G)\bigl(s',(r-y)/\beta\bigr)\,dp^{\delta(s,r)}(s',y\mid s)=0,
\end{aligned}
$$

and if $s\notin S_0$, then

$$
\begin{aligned}
&(T^{\delta})^{n+1}(F-G)(s,r)\\
&\quad= \int_{(s',y)\in S_0^c\times R} (T^{\delta})^n(F-G)\bigl(s',(r-y)/\beta\bigr)\,dp^{\delta(s,r)}(s',y\mid s)\\
&\quad\leqslant \int_{(s',y)\in S_0^c\times R} P^{\pi}_{(s',(r-y)/\beta)}\left(\bigcap_{k=2}^{n+1}\bigl\{X_k\in S_0^c\bigr\}\right)dp^{\delta(s,r)}(s',y\mid s)\\
&\quad= P^{\pi}_{(s,r)}\left(\bigcap_{k=2}^{n+2}\bigl\{X_k\in S_0^c\bigr\}\right).
\end{aligned}
$$

By induction, it follows that

$$
(F-G)(s,r)\leqslant (T^{\delta})^n(F-G)(s,r)\leqslant P^{\pi}_{(s,r)}\left(\bigcap_{k=2}^{n+1}\bigl\{X_k\in S_0^c\bigr\}\right),
$$

for each $(s,r)\in S_0^c\times R$ and all $n\geqslant 1$. Since $P^{\pi}_{(s,r)}(\bigcup_{n=2}^{\infty}\{X_n\in S_0\})=1$, we have

$$
\lim_{n\to\infty} P^{\pi}_{(s,r)}\left(\bigcap_{k=2}^{n+1}\bigl\{X_k\in S_0^c\bigr\}\right)=1-P^{\pi}_{(s,r)}\left(\bigcup_{k=2}^{\infty}\{X_k\in S_0\}\right)=0.
$$

Letting $n\to\infty$ on the above inequality, we have $(F-G)(s,r)\leqslant 0$ for every $(s,r)\in S_0^c\times R$, which completes the proof of the statement (i).

(ii) Let $F \in \mathcal{F}_m$ be a solution to $F = T^\delta F$ with $F = F^\pi$ on $S_0 \times R$. Since F^π is a solution to $F = T^\delta F$ in $\mathcal{F}_m$, we have $F - F^\pi = T^\delta(F - F^\pi)$ on $S_0^c \times R$. Thus the statement (i) implies that $F = F^\pi$. □

Theorem 4.2. *Suppose there is a subset S_0 of S such that S_0 is closed and reachable for every $\tau \in C_D$.*

(i) *F^* is the unique solution to $F = TF$ with $F = F^*$ on $S_0 \times R$.*
(ii) *If a policy $\pi = \delta^\infty \in C_D$ satisfies equations $F^* = T^\delta F^*$ on $S_0^c \times R$ and $F^* = F^\pi$ on $S_0 \times R$, then π is optimal in $(\mathcal{P})$.*

Proof. (i) From Theorem 3.2, we see that $F^* = TF^* \in \mathcal{F}_r$ and there is a decision rule $\delta \in \Delta_D$ such that $F^* = T^\delta F^*$. Let $F \in \mathcal{F}_r$ be another solution to equation $F = TF$ with $F = F^*$ on $S_0 \times R$. It follows from Lemma 2.2 that there is a decision rule $\delta' \in \Delta_D$ such that $F = T^{\delta'} F$. Thus we have $F^* = T^\delta F^* \leqslant T^{\delta'} F^*$ and $F = T^{\delta'} F \leqslant T^\delta F$ on $S_0^c \times R$. Hence we have $F^* - F \geqslant T^\delta(F^* - F)$ and $F - F^* \geqslant T^{\delta'}(F - F^*)$ on $S_0^c \times R$. From Lemma 4.2(i), we obtain $F = F^*$, which completes the uniqueness of F^*.

(ii) From Lemma 4.2(ii), F^π is the unique solution to $F = T^\delta F$ with $F = F^\pi$ on $S_0 \times R$. However, F^* satisfies the same equations by assumption. Thus, we have $F^* = F^\pi$ on S_R, and hence π is optimal in $(\mathcal{P})$. □

We give a necessary condition for a policy to be optimal.

Theorem 4.3. *If a stationary policy $\pi = \delta^\infty \in C_M$ is optimal in $(\mathcal{P})$, then $F^* = T^\delta F^*$ holds.*

Proof. From Lemma 3.1, we obtain $F^\pi = T^\delta F^\pi \geqslant T^\delta F^* \geqslant TF^* = F^*$. However, we have $F^\pi = F^*$, by optimality of π. Hence the desired result is obtained. □

Finally, we give simple examples for an infinite horizon case. Theorems 4.1 and 4.2 are applied to the first and the second examples, respectively. The first example is the same as Example 2 in Wu and Lin [14], but notations are different from it.

Example 4.1. Let state space be $S = \{s_1, s_2\}$, action space $A = A(s_i) = \{a_1, a_2\}$ $(i = 1, 2)$ and a discount factor $\beta = 1/2$. Letting q be

$$q_{ij}^k(y) = P(X_{t+1} = s_j, Y_t = y \mid X_t = s_i, A_t = a_k),$$

we assume that stochastic behavior of (X_{t+1}, Y_t) is determined by

$$\begin{aligned} q_{11}^1(2) &= q_{12}^1(2) = q_{21}^1(1) = q_{22}^1(1) \\ &= q_{11}^2(1) = q_{12}^2(1) = q_{21}^2(2) = q_{22}^2(2) = 1/2. \end{aligned}$$

Then we first find fixed points of T in $\mathcal{F}_m$ below. Since $F = TF$ and $F(s_i, 2r - 4) \leqslant F(s_i, 2r - 2)$, $i = 1, 2$, we have

$$F(s_i, r) = \big(F(s_1, 2r - 4) + F(s_2, 2r - 4)\big)\big/2, \quad i = 1, 2.$$

By using the fact that $F(s, r) = 0$ if $r < 0$ and $F(s, r) = 1$ if $r > 4 = H/(1 - \beta)$, we obtain fixed points

$$G_\alpha(s, r) = \alpha I_{\{4\}}(r) + I_{(4,\infty)}(r)$$

for every $(s, r) \in S_R$, where α is an arbitrary constant so that $0 \leqslant \alpha \leqslant 1$. Since F^* is the maximal fixed point of T (Theorem 3.3), it follows that $F^* = G_1$. Let a decision rule δ be $\delta(s_i, r) = a_i$ for $i = 1, 2$ and $r \in R$. Then the decision rule satisfies the equation $G_\alpha = T^\delta G_\alpha$ for any α: $0 \leqslant \alpha \leqslant 1$. Also, when (s, r) is any initial state and a policy $\pi = \delta^\infty$ is used, we have $Y_n = 2$ $(n \geqslant 1)$, so $Z = 4$. Thus $F^\pi(s, r; z) = I_{[4,\infty)}(z)$ for every $(s, r, z) \in S_R \times R$ and hence $z = r = 4$ is only one discontinuous point of $F^\pi(s, r; \cdot)$ and $E = \{4\}$. Therefore $\pi \in C_D^*$. Since $F^\pi = G_1 \in \mathcal{F}_r$, it follows from Theorem 4.1(ii) that $F^* = F^\pi$ and π is optimal in $(\mathcal{P})$. Furthermore, Theorem 4.1(i) implies that F^* is the unique solution to $F = TF$ in $\mathcal{F}_r$.

On the other hand, let another decision rule be

$$\rho(s_1, r) = \begin{cases} a_2 & (r < 3), \\ a_1 & (r \geqslant 3), \end{cases} \qquad \rho(s_2, r) = \begin{cases} a_1 & (r < 3), \\ a_2 & (r \geqslant 3). \end{cases}$$

Then ρ satisfies $F^* = T^\rho F^*$, but $\tau = \rho^\infty \notin C_D^*$. Indeed, for an initial state (s, r), we have

$$Z = 2I_{(-\infty,2)}(r) + rI_{[2,4)}(r) + 4I_{[4,\infty)}(r).$$

We will show this equality below. It is obvious for $r < 2$ and $r \geqslant 4$. Let $2 \leqslant r < 4$. In order to $Z = r$, it suffices to show that $Z_n = h(n - 1, k - 1)$, $n \geqslant 1$, if $h(n - 1, k) \leqslant r < h(n - 1, k + 1)$, for $k = 0, 1, \ldots, 2^n - 1$, where $h(n, k) = 2 + k/2^n$. We prove the relation by induction. When $n = 1$, we easily see that $Z_1 = Y_1 = 1$ if $r < 3$ and $Z_1 = 2$ otherwise, which implies that the relation holds. Assume that it is true for n. Since $\rho(s_i, w_{n+1}) = a_i$ $(i = 1, 2)$ if $w_{n+1} \geqslant 3$, we obtain that $Y_{n+1} = 2$ if and only if $w_{n+1} = 2(w_n - Y_n) \geqslant 3$; that is, $w_n \geqslant 2 + 1/2$ when $w_n < 3$ $(Y_n = 1)$ and $w_n \geqslant 2 + 3/2$ when $w_n \geqslant 3$ $(Y_n = 2)$. By iterating this argument, $Y_{n+1} = 2$ if and only if we have the form of inequality $r = w_1 \geqslant h(n, 2k + 1)$, $k = 0, 1, \ldots, 2^n - 1$. Hence when $h(n, 2k + 1) \leqslant r$ $(< h(n - 1, k + 1))$, it follows that $Z_{n+1} = Z_n + 2 \cdot (1/2)^n = h(n, 2k)$. By a similar argument, when $(h(n - 1, k) \leqslant)$ $r < h(n, 2k + 1)$, we obtain that $Z_{n+1} = Z_n + (1/2)^n = h(n, 2k - 1)$. Hence the relation holds for $n + 1$. Thus the desired relation is proved. Therefore

$$F^\tau(s, r; z) = \begin{cases} I_{[2,\infty)}(z) & (r < 2), \\ I_{[r,\infty)}(z) & (2 \leqslant r < 4), \\ I_{[4,\infty)}(z) & (r \geqslant 4). \end{cases}$$

Hence, when r is in interval $E = [2, 4)$ which is not a countable set, $F^\tau(s, r; z)$ is discontinuous at $z = r$. Thus we see that $\tau \notin C_D^*$.

Example 4.2. Let state space be $S = \{s_1, s_2, s_3\}$ and let action space be $A = A(s_i) = \{a_1, a_2\}$, $i = 1, 2, 3$. Letting notation q be the same as Example 4.1, we assume that (X_{t+1}, Y_t) is determined by

$$q_{11}^1(b_1) = q_{12}^2(0) = 3/4, \qquad q_{13}^1(0) = q_{11}^2(b_2) = 1/4,$$
$$q_{22}^1(H) = q_{23}^1(H) = q_{32}^1(b_3) = q_{33}^1(b_3) = 1/2,$$
$$q_{22}^2(b_3) = q_{23}^2(b_3) = q_{32}^2(H) = q_{33}^2(H) = 1/2,$$

where b_1, b_2, b_3 and H are constants such that $0 < b_2 < b_1 < H$, $0 < b_3 < H$ and $b_1/H < \beta < 1$. Let $d = H/(1-\beta)$ and $\beta_i = b_i + \beta d$ $(i = 1, 2)$.

Then it is obvious that a subset $S_0 = \{s_2, s_3\}$ of S is closed and reachable for every policy $\pi \in C_D$ and the system on S_0 has the same behavior as that in Example 4.1. Thus, by the way similar to Example 4.1, it is easily checked that optimal value for $(\mathcal{P})$ on S_0 are $F^*(s, r) = I_{[d,\infty)}(r)$ for each $(s, r) \in S_0 \times R$, and optimal policy $\pi^* = \delta^\infty \in C_D^*$ on S_0 is determined by $\delta(s_i, r) = a_{i-1}$, $i = 2, 3$, $r \in R$.

From Theorem 4.2, F^* is the unique solution to $F = TF$ with $F = F^*$ on $S_0 \times R$. We can obtain $F^*(s_1, r)$ below. From the definition of T^a, we have

$$T^{a_1}F^*(s_1, r) = (3/4)F^*\big(s_1, (r - b_1)/\beta\big) + (1/4)F^*(s_3, r/\beta).$$

However, it follows that $F^*(s_3, r/\beta) = I_{[\beta d,\infty)}(r)$ and

$$F^*\big(s_1, (r - b_1)/\beta\big) = \begin{cases} 0 & \text{if } r < b_1, \\ 1 & \text{if } r \geqslant \beta_1, \end{cases}$$

since $F^*(s, r) = 0$ if $r < 0$ and $F^*(s, r) = 1$ if $r \geqslant d$. Hence we have

$$T^{a_1}F^*(s_1, r) = \begin{cases} 0 & \text{if } r < b_1, \\ (3/4)F^*(s_1, (r - b_1)/\beta) & \text{if } b_1 \leqslant r < \beta d, \\ (3/4)F^*(s_1, (r - b_1)/\beta) + 1/4 & \text{if } \beta d \leqslant r < \beta_1, \\ 1 & \text{if } r \geqslant \beta_1. \end{cases}$$

Similarly, we have

$$T^{a_2}F^*(s_1, r) = \begin{cases} 0 & \text{if } r < b_2, \\ (1/4)F^*(s_1, (r - b_2)/\beta) & \text{if } b_2 \leqslant r < \beta d, \\ (1/4)F^*(s_1, (r - b_2)/\beta) + 3/4 & \text{if } \beta d \leqslant r < \beta_2, \\ 1 & \text{if } r \geqslant \beta_2. \end{cases}$$

Since $F^* = TF^* = \min_i T^{a_i}F^*$, we obtain

$$F^*(s_1, r) = \begin{cases} 0 & \text{if } r < b_1/(1-\beta), \\ \min\{(3/4)F^*(s_1, (r - b_1)/\beta), (1/4)F^*(s_1, (r - b_2)/\beta)\} & \\ \qquad \text{if } b_1/(1-\beta) \leqslant r < \beta d, & \\ 1 & \text{if } r \geqslant \beta d. \end{cases}$$

Indeed, it is clear that if $r < b_1$ then $F^*(s_1, r) = 0$, and if $r \geqslant \beta_1$ then $F^*(s_1, r) = 1$. We also notice from condition that $b_1/(1-\beta) < \beta d < \beta_2$. When $b_1 \leqslant r < b_1/(1-\beta)$, we have

$$F^*(s_1, r) = \min\big((3/4)F^*\big(s_1, (r-b_1)/\beta\big), (1/4)F^*\big(s_1, (r-b_2)/\beta\big)\big).$$

If $b_1 \leqslant r < b_1(1+\beta)$, then $(r-b_1)/\beta < b_1$ and $F^*(s_1, (r-b_1)/\beta) = 0$, and hence $F^*(s_1, r) = 0$. Assume that if $b_1 \leqslant r < b_1 \sum_{k=0}^{n-1} \beta^k = b_1(1-\beta^n)/(1-\beta)$, then $F^*(s_1, r) = 0$. If $r < b_1 \sum_{k=0}^{n} \beta^k$, then $(r-b_1)/\beta < b_1(1-\beta^n)/(1-\beta)$ and $F^*(s_1, (r-b_1)/\beta) = 0$, and hence $F^*(s_1, r) = 0$. By induction, for every $n \geqslant 1$ if $r < b_1(1-\beta^n)/(1-\beta)$, then $F^*(s_1, r) = 0$. Hence, letting $n \to \infty$, we obtain $F^*(s_1, r) = 0$ when $b_1 \leqslant r < b_1/(1-\beta)$. When $\beta_2 \leqslant r < \beta_1$, we have

$$F^*(s_1, r) = (3/4)F^*\big(s_1, (r-b_1)/\beta\big) + 1/4.$$

If $\beta_1 > r \geqslant b_1(1+\beta) + \beta^2 d \geqslant \beta_2$, then $(r-b_1)\beta \geqslant \beta_1$ and $F^*(s_1, (r-b_1)/\beta) = 1$, and hence $F^*(s_1, r) = 1$. By induction, if $\beta_1 > r \geqslant b_1(1-\beta^n)/(1-\beta) + \beta^n d$ and $r \geqslant /\beta_2$, we obtain $F^*(s_1, r) = 1$. Letting $n \to \infty$ and noticing $\beta_2 \geqslant b_1/(1-\beta)$, we obtain $F^*(s_1, r) = 1$ when $\beta_2 \leqslant r < \beta_1$. Similarly, if $\beta d \leqslant r < \beta_2$, we have $F^*(s_1, r) = 1$.

A right continuous optimal policy $\pi^* = \delta^\infty$ in $(\mathcal{P})$ is given by $\delta(s_i, r) = a_{i-1}$ for every $r \in R$ and $i = 2, 3$, and

$$\delta(s_1, r) = \begin{cases} a_1 \text{ or } a_2 & \text{if } r < b_2 + \beta b_1/(1-\beta), \\ a_1 & \text{if } b_2 + \beta b_1/(1-\beta) \leqslant r < b_1/(1-\beta), \\ a_1 & \text{if } b_1/(1-\beta) \leqslant r < \beta d \text{ and} \\ & \quad F^*(s_1, r) = (3/4)F^*(s_1, (r-b_1)/\beta), \\ a_2 & \text{if } b_1/(1-\beta) \leqslant r < \beta d \text{ and} \\ & \quad F^*(s_1, r) = (1/4)F^*(s_1, (r-b_2)/\beta), \\ a_1 \text{ or } a_2 & \text{if } r \geqslant \beta d, \end{cases}$$

since, by Theorem 4.2, policy $\pi^* = \delta^\infty$ satisfying $F^* = T^\delta F^*$ on $S_0^c \times R$ is optimal.

Acknowledgments

The authors would like to thank Professors J.R. Birge and C.C. White for their valuable comments which were helpful to improve this paper. The authors are also grateful to anonymous reviewers for valuable remarks and helpful comments.

References

[1] C. Derman, Finite State Markovian Decision Processes, Academic Press, New York, 1970.
[2] A. Hordijk, Dynamic Programming and Markov Potential Theory, Mathematical Centre Tracts, Vol. 51, Mathematisch Centrum, Amsterdam, 1974.

[3] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley, New York, 1994.
[4] D.J. White, Markov Decision Processes, Wiley, New York, 1993.
[5] D.J. White, Mean, variance and probabilistic criteria in finite Markov decision processes: A review, J. Optim. Theory Appl. 56 (1988) 1–29.
[6] D.J. White, Utility, probabilistic constraint, mean and variance of discounted rewards in Markov decision processes, OR Spektrum 9 (1987) 13–22.
[7] J.A. Filar, L.C.M. Kallenberg, H.-M. Lee, Variance-penalized Markov decision processes, Math. Oper. Res. 14 (1989) 147–161.
[8] J.A. Filar, D. Krass, K.W. Ross, Percentile performance criteria for limiting average Markov decision processes, IEEE Trans. Automat. Control 40 (1995) 2–10.
[9] Y. Kadota, M. Kurano, M. Yasuda, A utility deviation in discounted Markov decision processes with general utility, Bull. Inform. Cybernet. 28 (1996) 71–78.
[10] M. Bouakiz, Y. Kebir, Target-level criterion in Markov decision processes, J. Optim. Theory Appl. 86 (1995) 1–15.
[11] K.J. Chung, M.J. Sobel, Discounted MDP's: Distribution functions and exponential utility maximization, SIAM J. Control Optim. 25 (1987) 49–62.
[12] M.J. Sobel, The variance of discounted Markov decision processes, J. Appl. Probab. 19 (1982) 794–802.
[13] D.J. White, Minimising a threshold probability in discounted Markov decision processes, J. Math. Anal. Appl. 173 (1993) 634–646.
[14] C. Wu, Y. Lin, Minimizing risk models in Markov decision processes with policies depending on target values, J. Math. Anal. Appl. 231 (1999) 47–67.