Journal of Mathematical Analysis and Applications **231,** 47–67 (1999)
Article ID jmaa.1998.6203, available online at http://www.idealibrary.com on IDEAL®

Minimizing Risk Models in Markov Decision Processes with Policies Depending on Target Values

Congbin Wu and Yuanlie Lin*

Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China

Submitted by E. Stanley Lee

Received November 3, 1997

This paper studies the minimizing risk problems in Markov decision processes with countable state space and reward set. The objective is to find a policy which minimizes the probability (risk) that the total discounted rewards do not exceed a specified value (target). In this sort of model, the decision made by the decision maker depends not only on system's states, but also on his target values. By introducing the decision-maker's state, we formulate a framework for minimizing risk models. The policies discussed depend on target values and the rewards may be arbitrary real numbers. For the finite horizon model, the main results obtained are: (i) The optimal value functions are distribution functions of the target, (ii) there exists an optimal deterministic Markov policy, and (iii) a policy is optimal if and only if at each realizable state it always takes optimal action. In addition, we obtain a sufficient condition and a necessary condition for the existence of finite horizon optimal policy independent of targets and we give an algorithm computing finite horizon optimal policies and optimal value functions. For an infinite horizon model, we establish the optimality equation and we obtain the structure property of optimal policy. We prove that the optimal value function is a distribution function of target and we present a new approximation formula which is the generalization of the nonnegative rewards cases. An example which illustrates the mistakes of previous literature shows that the existence of optimal policy has not been proved really. In this paper, we give an existence condition, which is a sufficient and necessary condition for the existence of an infinite horizon optimal policy independent of targets, and we point out that whether there exists an optimal policy remains an open problem in the general case.

* E-mail: ylin@math.tsinghua.edu.cn.

0022-247X/99 $30.00

1. INTRODUCTION

In this paper, we consider the following discrete-time and stationary Markov decision processes (MDP, for short),

$$\Gamma = (S, A, W, P, \beta). \tag{1}$$

The state space S and the action space $A = \bigcup_{i \in S} A(i)$ both are nonempty and countable, where $A(i)$ is the set of admissible actions when the system is in state i. For each $i \in S$, $A(i)$ is nonempty and finite. Reward set W is a countable subset of $\mathbb{R} = (-\infty, +\infty)$. Let X_n, Δ_n, and R_n denote the state of the system, the action taken by the decision maker, and the reward received at stage n, respectively. Then the stationary conditional transition probability P is given by

$$P(X_{n+1} = j, R_n = r \mid X_n = i, \Delta_n = a) = p_{ijr}^a, \quad i, j \in S, a \in A(i), r \in W, n \geq 1. \tag{2}$$

$$\sum_{j \in S,\, r \in W} p_{ijr}^a = 1, \qquad i \in S, \qquad a \in A(i). \tag{3}$$

Discounted factor $\beta \in (0, 1)$.

We assume that W is bounded.

A decision rule π_n, at stage n, specifies the action to take at stage n. A policy π is a sequence of decision rules: $\pi = (\pi_1, \pi_2, \ldots, \pi_n, \ldots)$ (The precise definitions of policies will be given in the next section.)

Let B_n^π and B^π denote the random total discounted rewards generated by policy π for finite and infinite horizon problems, respectively. Then,

$$B_n^\pi = \sum_{k=1}^{n} \beta^{k-1} R_k^\pi, \qquad n \geq 1, \qquad B^\pi = \sum_{k=1}^{\infty} \beta^{k-1} R_k^\pi. \tag{4}$$

Our optimization problem for minimizing risk models is the following: Find a policy which minimizes the probability (risk) that the total discounted rewards do not exceed a specified value (target), that is, for the finite horizon model, find a policy π^* such that

$$\text{Prob}(B_n^{\pi^*} \leq x \mid X_1 = i) = \inf_\pi \{\text{Prob}(B_n^\pi \leq x \mid X_1 = i)\}, \qquad i \in S, x \in \mathbb{R}, \tag{5}$$

and for the infinite horizon model, one seeks a policy π^* such that

$$\text{Prob}(B^{\pi^*} \le x \mid X_1 = i) = \inf_{\pi} \{\text{Prob}(B^{\pi} \le x \mid X_1 = i)\}, \qquad i \in S, x \in \mathbb{R}, \tag{6}$$

where the infimum is taken over all policies.

This sort of problem is studied by Sobel [4] and Chung and Sobel [5]. In [4], Sobel only gives the recurrence equation of the objective function for stationary policies which component decision rules are only functions of X_n. In [5], Chung and Sobel illustrate that the notation p_{ijr}^a is essential for this sort of problem, but they only establish several properties of operator T^δ (defined in the next section) where δ is the only function of X_n.

The general study of this problem is done by Bouakiz and Kebir [1] and White [3]. In [1], Bouakiz and Kebir study a finite MDP with a positive and a finite reward set. Their main results include the optimality equations of the model and the property of the optimal value function for finite and infinite horizon. In [3], White considers a finite MDP with a nonnegative and bounded reward set and gives a set of results for optimal policy, optimal value function, policy iteration, and error bounds.

However, both [1] and [3] have not given a structure result of optimal policies and an effective algorithm computing optimal policies and optimal value functions. In addition [1] and [3] hold incorrectly that all objective functions generated by policies are the distribution functions of the target, this brings about that the proof of part of the results in [1] needs to be modified and the key Lemma 3 in [3] does not hold in the general case (see Section 4 of this paper). Thus, in [3], some main results (e.g., the existence of the optimal policy for infinite horizon) have not been proved really and part of the treatment (e.g., the policy space procedure in [3]) is inappropriate. This shows that it is necessary to give a clear and precise description for policy and objective functions.

Another related paper [6] studies a finite MDP with percentile performance criteria where the decision maker is interested in finding a policy that achieves a specified value (target) of the long-run limiting average reward at a specified probability level. Reference [6] points out that the satisfactory treatment of the discounted case with percentile performance criteria is an important open problem. The results and the approaches in this paper give an answer to this problem in a certain sense.

In this paper, we study not only the optimality equations of the model and the property of optimal value functions, but also the existence and structure of optimal policy and the algorithm for computing optimal policies. The policies discussed depend on target values and rewards may be arbitrary real numbers. The technique taken in this paper is different from Bouakiz and Kebir [1] and White [3].

In Section 2, by introducing the decision-maker's state, we formulate a framework for minimizing risk models and we give a clear and precise description for policy and objective function.

Section 3 considers the finite horizon model. We prove that (i) the optimal value functions are distribution functions of target, (ii) there exists an optimal deterministic Markov policy, and (iii) a policy is optimal if and only if at each realizable state it always takes optimal action, and it gives a sufficient condition and a necessary condition for the existence of a finite horizon optimal policy independent of targets.

Section 4 deals with the infinite horizon model. We establish the optimality equation and we obtain the structure property of optimal policy. We prove that the optimal value function is a distribution function of target. For the case in which rewards may take negative values, where the approximation relation of optimal value functions in [1] and [3] does not hold, we obtain a new approximation formula which is the generalization of the nonnegative rewards cases. An example which illustrates the mistakes of previous literature shows that the existence of optimal policy has not been proved really. In this section, we give an existence condition, which is a sufficient and necessary condition for the existence of an infinite horizon optimal policy independent of targets, and we point out that whether there exists an optimal policy remains an open problem in the general case.

In Section 5, state space and reward set both are finite. We give an algorithm computing finite horizon optimal policies and optimal value functions and we point out that the optimal value functions are step functions of the target with finite jump points and there exists an optimal deterministic policy which structure is analogous to that of optimal value functions.

2. TREATMENT AND DEFINITION

Different from the standard optimization criterion in MDP, which maximizes the expected value of the total discounted rewards, minimizing risk criterion is risk-sensitive (see [2]). The decision maker considers not only the system's state but also his target when making decision and taking action at each stage. Therefore, the policies depend on the system's state and target.

We refer to (i, x) as the state of the decision maker to distinguish from the system's state i, where x is the target value. Before giving the definition of policy, we first expand MDP Γ by enlarging state space. Note that if the initial state of the decision maker is (i, x) and an action a is taken, by (2), the decision-maker's state translates to $(j, (x - r)/\beta)$ from

(i, x) in probability p^a_{ijr}. Thus, if we denote E as the space of decision-maker's state and Γ_1 as MDP generated by expanding Γ, then the MDP Γ_1 has the following structure,

$$\Gamma_1 = (E, A, W, P, \beta), \tag{7}$$

where the state space $E = S \times \mathbb{R}$, the action space $A = \bigcup_{(i,x)\in E} A(i, x) = \bigcup_{i \in S} A(i)$, and $A(i, x) = A(i)$, $(i, x) \in E$, the stationary conditional transition probability P,

$$P\left(Y_{n+1} = \left(j, \frac{x - r}{\beta}\right) \middle| Y_n = (i, x), \Delta_n = a\right) = p^a_{ijr},$$
$$i, j \in S, a \in A(i), r \in W, x \in \mathbb{R}, \tag{8}$$

and the reward set W and the discounted factor β are the same as MDP Γ.

Let H_n denote the set of all (admissible) histories up to n. A generic element h_n of H_n is a vector of the form $h_n = (i_1, x_1, a_1, \ldots, i_{n-1}, x_{n-1}, a_{n-1}, i_n, x_n)$ where $(i_k, x_k) \in E$, $a_k \in A(i_k)$, $k = 1, \ldots, n - 1$, and $(i_n, x_n) \in E$.

A decision rule π_n at time n is a conditional transition probability measure from H_n to A satisfying the constraint: for any $h_n \in H_n$ and $C \subset A$, $\pi_n(\cdot \mid h_n)$ is a probability measure on A such that $\pi_n(A(i_n) \mid h_n) = 1$ and $\pi_n(C \mid \cdot)$ is a measurable function from H_n to $[0, 1]$, where H_n is endowed with the natural Borel sigma-algebra.

A policy π is a sequence of decision rules $\pi = (\pi_1, \pi_2, \ldots, \pi_n, \ldots)$.

A Markov policy π is one in which each π_n only depends on the current state at time n, that is, $\pi_n(\cdot \mid h_n) = \pi_n(\cdot \mid i_n, x_n)$ for all $h_n \in H_n$.

A stationary policy π which can be denoted by $\pi = \pi_1^\infty$ is a Markov policy with an identical decision rule.

A deterministic policy π is one in which each π_n is nonrandomized, that is, π_n is a measurable mapping from H_n to A such that $\pi_n(h_n) \in A(i_n)$ for all $h_n \in H_n$.

Let Π, Π_m, Π^d_m, Π_s, and Π^d_s denote the sets of all policies, all Markov policies, all deterministic Markov policies, all stationary policies, and all deterministic stationary policies, respectively.

Let Π_0 denote the set of all policies which are independent of targets $x_n (n \geq 1)$.

For any $\pi = (\pi_k, k \geq 1) \in \Pi$ and some given history (i, x, a), the cut-head policy of π to (i, x, a) is defined by ${}^1\pi^{(i,x,a)} = (\pi_k^{(i,x,a)}, k \geq 1)$, where $\pi_k^{(i,x,a)}(\cdot \mid h_k) = \pi_{k+1}(\cdot \mid (i, x, a), h_k)$ for all $h_k \in H_k$, $k \geq 1$.

For any $\pi = (\pi_k,\ k \geq 1)$, $\sigma = (\sigma_k,\ k \geq 1) \in \Pi$. Let $\pi(n) = (\pi_1, \pi_2, \ldots, \pi_n)$ denote the truncation of π to n stages and $(\pi(n), \sigma) = (\pi_1, \pi_2, \ldots, \pi_n, \sigma_1, \sigma_2, \ldots)$ denote the policy in which at first n stages π is taken and from $(n+1)$th stage downward σ is taken, starting from σ_1. $(\pi(n), \sigma)$ is called an n stages delay policy of σ to π.

Let P_π denote the conditional probability measure determined by π and P.

To simplify the notations, we will use B and B_n instead of B^π and B_n^π, respectively.

For any $\pi \in \Pi$, the objective functions generated by π are

$$F_n^\pi(i,x) = P_\pi\big(B_n \leq x \mid Y_1 = (i,x)\big), \qquad n \geq 1, \tag{9}$$

$$F^\pi(i,x) = P_\pi\big(B \leq x \mid Y_1 = (i,x)\big). \tag{10}$$

The optimal value functions are

$$F_n^*(i,x) = \inf_{\pi\in\Pi} F_n^\pi(i,x), \qquad F^*(i,x) = \inf_{\pi\in\Pi} F^\pi(i,x), \qquad (i,x) \in E. \tag{11}$$

Obviously, F_n^π, F_n^*, F^π, and F^* satisfy that

$$F_n^\pi(i,x) = F_n^*(i,x) = \begin{cases} 0, & \text{if } x < \dfrac{b(1-\beta^n)}{1-\beta}, \\ 1, & \text{if } x \geq \dfrac{d(1-\beta^n)}{1-\beta}, \end{cases} \tag{12}$$

$$F^\pi(i,x) = F^*(i,x) = \begin{cases} 0, & \text{if } x < \dfrac{b}{1-\beta}, \\ 1, & \text{if } x \geq \dfrac{d}{1-\beta}, \end{cases} \tag{13}$$

where $b = \inf\{r \mid r \in W\}$ and $d = \sup\{r \mid r \in W\}$.

If $\pi^* \in \Pi$ such that $F_n^{\pi^*}(i,x) = F_n^*(i,x)$ for all $(i,x) \in E$, then π^* is called an n stages optimal policy.

If $\pi^* \in \Pi$ such that $F^{\pi^*}(i,x) = F^*(i,x)$ for all $(i,x) \in E$, then π^* is called an infinite horizon optimal policy, simply, optimal policy.

Remark 1. For any policy $\pi \in \Pi_0$, $F_n^\pi(i,x)$ and $F^\pi(i,x)$ are the distribution functions of x, but for general policy $\pi \in \Pi$, this result does not hold (see the example in Section 4).

To help the conciseness of analysis, we need to define the following notations: $D = \{u \mid u\colon E \to [0,1]$, a measurable function$\}$.

Let $\delta^\infty \in \Pi_s$. Define the operators G, T^δ, and T: for each $u \in D$,

$$Gu(i,x,a) = \sum_{j\in S,\, r\in W} p_{ijr}^a u\left(j, \frac{x-r}{\beta}\right), \qquad (i,x) \in E,\ a \in A(i). \tag{14}$$

$$T^\delta u(i,x) = \sum_{a\in A(i)} \delta(a \mid i,x) Gu(i,x,a), \qquad (i,x) \in E. \tag{15}$$

$$Tu(i,x) = \min_{a\in A(i)} Gu(i,x,a), \qquad (i,x) \in E. \tag{16}$$

$$\begin{gathered} (T^\delta)^0 u = u, \qquad (T^\delta)^n u = T^\delta\left((T^\delta)^{n-1}u\right), \\ T^0 u = u, \qquad T^n u = T(T^{n-1}u). \end{gathered} \tag{17}$$

Obviously, when $\delta^\infty \in \Pi_s^d$, $T^\delta u(i,x) = Gu(i,x,\delta(i,x))$.

In addition, we supplement $B_0 = 0$ and $F_0^\pi(i,x) = P_\pi(B_0 \le x \mid Y_1 = (i,x))$ for any $\pi = (\pi_k,\ k \ge 1) \in \Pi$. Then, we have

$$F_0^*(i,x) = F_0^\pi(i,x) = I_{[0,+\infty)}(x), \tag{18}$$

where $I_{[0,+\infty)}(x)$ is the indictor function of set $[0,+\infty)$.

LEMMA 1. (i) *If $u, v \in D$, $u \le v$, then $Gu \le Gv$, $T^\delta u \le T^\delta v$, $Tu \le Tv$.*

(ii) *Let $u \in D$. If $u(i,x)$ is a nondecreasing and a right continuous function of x for any $i \in S$, then, $Tu(i,x)$ is also a nondecreasing and a right continuous function of x for each $i \in S$.*

Proof. The proof is obvious. ∎

3. FINITE HORIZON MODEL

LEMMA 2. *Let $\pi = (\pi_k,\ k \ge 1) \in \Pi$. Then,*

$$F_n^\pi(i,x) = \sum_{a\in A(i)} \pi_1(a \mid i,x) \sum_{j\in S,\, r\in W} p_{ijr}^a F_{n-1}^{1\pi^{(i,x,a)}}\left(j, \frac{x-r}{\beta}\right),$$
$$(i,x) \in E,\ n \ge 1, \tag{19a}$$

and F_n^π is determined by $\pi(n)$.

Proof. By the properties of P_π (see Hernández-Lerma [7]), we have

$$
\begin{aligned}
&F_n^\pi(i,x)\\
&\quad = P_\pi\big(B_n \le x \mid Y_1 = (i,x)\big)\\
&\quad = \sum_{a\in A(i)} \pi_1(a \mid i,x) \sum_{j\in S,\, r\in W} P\left(Y_2 = \left(j, \frac{x-r}{\beta}\right) \mid Y_1 = (i,x),\, \Delta_1 = a\right)\\
&\qquad \cdot P_\pi\left(B_n \le x \mid Y_1 = (i,x),\, \Delta_1 = a,\, Y_2 = \left(j, \frac{x-r}{\beta}\right)\right)\\
&\quad = \sum_{a\in A(i)} \pi_1(a \mid i,x)\\
&\qquad \times \sum_{j\in S,\, r\in W} p_{ijr}^a P_{1\pi^{(i,x,a)}}\left(B_{n-1} \le \frac{x-r}{\beta} \mid Y_1 = \left(j, \frac{x-r}{\beta}\right)\right)\\
&\quad = \sum_{a\in A(i)} \pi_1(a \mid i,x) \sum_{j\in S,\, r\in W} p_{ijr}^a F_{n-1}^{1\pi^{(i,x,a)}}\left(j, \frac{x-r}{\beta}\right).
\end{aligned}
$$

Equation (19a) is proved.

Using (19a) repeatedly, we obtain another part of Lemma 2 immediately. ∎

By slightly abusing the notation T^{π_1}, we denote the right side of (19a) by $T^{\pi_1}F_{n-1}^{1\pi}(i,x)$. Thus, the equality (19a) can be simplified into

$$F_n^\pi = T^{\pi_1}F_{n-1}^{1\pi}, \qquad n \ge 1. \tag{19b}$$

Theorem 1 is one of the main results in this paper.

THEOREM 1. (i) $\{F_n^*,\ n \ge 0\}$ *satisfies optimality equations,*

$$F_0^* = I_{[0,+\infty)}, \qquad F_n^* = TF_{n-1}^*, \qquad n \ge 1. \tag{20}$$

(ii) *For any* $n \ge 0$ *and* $i \in S$, $F_n^*(i,x)$ *is a distribution function of* x.

(iii) *For any* $n \ge 0$, *there exists a policy* $\pi \in \Pi_m^d$ *such that* $F_n^\pi = F_n^*$.

Proof. We prove Theorem 1 by induction. When $n = 0$, by (18), Theorem 1 is true. Assume that Theorem 1 holds when $n = k$.

By induction assumption, for any $i \in S$, $F_k^*(i,x)$ is a distribution function of x. Note that $A(i,x) = A(i)$ is finite for any $(i,x) \in E$. By the measurable selection theorem [7, Proposition D3, p. 130], there exists a measurable mapping δ from E to A such that $\delta(i,x) \in A(i)$ and

$GF_k^*(i, x, \delta(i, x)) = TF_k^*(i, x)$ for all $(i, x) \in E$, that is, $\delta^\infty \in \Pi_s^d$ and $T^\delta F_k^* = TF_k^*$.

By induction assumption, there exists a policy $\sigma \in \Pi_m^d$ such that $F_k^\sigma = F_k^*$. Let $\pi = (\delta, \sigma)$. Then, $\pi \in \Pi_m^d$. By Lemma 2, we have

$$F_{k+1}^*(i, x) \le F_{k+1}^\pi(i, x) = T^\delta F_k^\sigma(i, x) = T^\delta F_k^*(i, x) = TF_k^*(i, x). \quad (21)$$

On the other hand, for any $\eta \in \Pi$, by Lemma 2 we have

$$F_{k+1}^\eta(i, x) = T^{\eta_1} F_k^{{}^1\eta}(i, x) \ge T^{\eta_1} F_k^*(i, x) \ge TF_k^*(i, x).$$

Hence, $F_{k+1}^*(i, x) \ge TF_k^*(i, x)$.

Associating it with (21), we obtain $TF_k^* = F_{k+1}^* = F_{k+1}^\pi$. Thus, by Lemma 1 and (12), $F_{k+1}^*(i, x)$ is a distribution function of x.

Earlier results imply that Theorem 1 is also true when $n = k + 1$. By induction, for any $n \ge 0$, Theorem 1 holds. This completes the proof of Theorem 1. ∎

COROLLARY 1. $F_n^*(i, x) = \inf_{\pi \in \Pi} F_n^\pi(i, x) = \inf_{\pi \in \Pi_m^d} F_n^\pi(i, x), (i, x) \in E$, $n \ge 1$.

We define

$$A_n(i, x) = \{a \mid a \in A(i) \text{ and } GF_{n-1}^*(i, x, a) = F_n^*(i, x)\}, \quad n \ge 1, (i, x) \in E,$$
$$A_n(i) = \bigcap_{x \in \mathbb{R}} A_n(i, x), \qquad n \ge 1, i \in S. \quad (22)$$

Then, by Theorem 1, $A_n(i, x) \ne \varnothing$ for any $n \ge 1$, $(i, x) \in E$. But, it is possible that $A_n(i) = \varnothing$. In Theorem 4, we see that $\{A_n(i): i \in S, n \ge 1\}$ plays a crucial role for the existence of the optimal policy independent of targets x_n $(n \ge 1)$.

THEOREM 2. *Let δ_k be a measurable mapping from E to A and satisfy $\delta_k(i, x) \in A_k(i, x)$ for all $(i, x) \in E$, $k \ge 1$. Then any policy π which satisfies $\pi(n) = (\delta_n, \delta_{n-1}, \ldots, \delta_1)$ is n stages optimal.*

Proof. Note that $T^{\delta_k} F_{k-1}^* = F_k^*$ for all $k \ge 1$. When $n = 1$, by Lemma 2 and (18), we have that $F_1^\pi = T^{\pi_1} F_0^{{}^1\pi} = T^{\delta_1} F_0^* = F_1^*$. Assume that Theorem 2 holds when $n = k$.

Let $n = k + 1$. Then, $F_k^{{}^1\pi^{(i, x, a)}} = F_k^*$ because ${}^1\pi^{(i, x, a)}(k) = (\delta_k, \delta_{k-1}, \ldots, \delta_1)$. Hence, by Lemma 2, we obtain that $F_{k+1}^\pi = T^{\pi_1} F_k^{{}^1\pi} = T^{\delta_{k+1}} F_k^* = F_{k+1}^*$. By induction, Theorem 2 is proved. ∎

With respect to the structure of n stages optimal policy, we have the following result.

THEOREM 3. *Let* $\pi = (\pi_k,\ k \geq 1) \in \Pi$. *For any given* $(i, x) \in E$, $F_n^{\pi}(i, x) = F_n^*(i, x)$ *if and only if* $\pi_1(A_n(i, x) \mid i, x) = 1$ *and* $F_{n-1}^{1_\pi(i,x,a)}(j, (x - r)/\beta) = F_{n-1}^*(j, (x - r)/\beta)$ *when* $\pi_1(a \mid i, x) p_{ijr}^a > 0$.

Proof. Assume that $F_n^{\pi}(i, x) = F_n^*(i, x)$. By Theorem 1, there exists a policy σ such that $F_{n-1}^{\sigma} = F_{n-1}^*$. Hence, by Lemma 2, we have

$$\begin{aligned} F_n^*(i, x) = F_n^{\pi}(i, x) &= T^{\pi_1} F_{n-1}^{1\pi}(i, x) \geq T^{\pi_1} F_{n-1}^*(i, x) \\ &= T^{\pi_1} F_{n-1}^{\sigma}(i, x) = F_n^{(\pi_1, \sigma)}(i, x) \geq F_n^*(i, x), \end{aligned}$$

and so

$$F_n^*(i, x) = T^{\pi_1} F_{n-1}^*(i, x), \qquad T^{\pi_1} F_{n-1}^{1\pi}(i, x) = T^{\pi_1} F_{n-1}^*(i, x).$$

Hence, by (14), (15), and (19), we conclude that

$$\sum_{a \in A(i)} \pi_1(a \mid i, x) \{ G F_{n-1}^*(i, x, a) - F_n^*(i, x) \} = 0, \tag{23}$$

$$\sum_{a \in A(i)} \sum_{j \in S,\, r \in W} \pi_1(a \mid i, x) p_{ijr}^a \left\{ F_{n-1}^{1_\pi(i,x,a)}\left(j, \frac{x - r}{\beta} \right) - F_{n-1}^*\left(j, \frac{x - r}{\beta} \right) \right\} = 0. \tag{24}$$

Thus, by Theorem 1 and (23), we have $\pi_1(A_n(i, x) \mid i, x) = 1$ for all $(i, x) \in E$. In addition, by (24), when $\pi_1(a \mid i, x) p_{ijr}^a > 0$, we obtain that $F_{n-1}^{1_\pi(i,x,a)}(j, (x - r)/\beta) = F_{n-1}^*(j, (x - r)/\beta)$.

The necessity of Theorem 3 is proved. Note that the preceding proof is reversible. So the sufficiency of Theorem 3 is also true. Theorem 3 is proved. ∎

Because of Theorem 3, we call $A_n(i, x)$ the optimal action set and it's element optimal action for n stages.

Remark 2. (i) Theorem 3 shows that a policy π is optimal for a finite horizon model if and only if the action taken by π at each realizable state is an optimal action and the corresponding cut-head policy is also optimal at each stage.

(ii) From Lemma 2 and Theorem 1, we can further see that π is n stages optimal if and only if the actions taken by π in the preceding n stages are optimal.

Theorem 4 gives a sufficient condition and a necessary condition for the existence of a finite horizon optimal policy independent of targets.

THEOREM 4. (i) *If there exists a policy* $\pi \in \Pi_0$ *such that* $F_n^{\pi} = F_n^*$, *then* $A_n(i) \neq \varnothing$ *and* $\pi_1(A_n(i) \mid i) = 1$ *for each* $i \in S$;

(ii) *If* $A_k(i) \neq \varnothing$ *for each* $i \in S$ *and* $1 \leq k \leq n$, *then there exists a policy* $\pi \in \Pi_0$ *such that* $F_n^\pi = F_n^*$.

Proof. (i) Let $\pi \in \Pi_0$ and $F_n^\pi = F_n^*$. Then, by Theorem 3, $\pi_1(A_n(i, x) \mid i) = 1$ for all $x \in \mathbb{R}$ and $i \in S$, it follows that $\pi_1(A_n(i) \mid i) = 1$ for each $i \in S$. Hence, $A_n(i) \neq \varnothing$ for each $i \in S$.

(ii) Select δ_k: $S \to A$ such that $\delta_k(i) \in A_k(i)$ for each $i \in S$ and $1 \leq k \leq n$. Then, by Theorem 3, for any policy $\pi \in \Pi_0$ which satisfies $\pi(n) = (\delta_n, \delta_{n-1}, \ldots, \delta_1)$, $F_n^\pi = F_n^*$ holds. ∎

4. INFINITE HORIZON MODEL

LEMMA 3. *Let* $\pi \in \Pi$. *Then,*

$$F^\pi(i, x) = \sum_{a \in A(i)} \pi_1(a \mid i, x) \sum_{j \in S,\, r \in W} p_{ijr}^a F^{1_\pi(i, x, a)}\left(j, \frac{x - r}{\beta}\right), \quad (i, x) \in E. \tag{25a}$$

Proof. The proof is similar to Lemma 2's. ∎

Similarly, we denote the right side of (25a) by $T^{\pi_1}F^{1\pi}(i, x)$ and we simplify the equality (25a) into

$$F^\pi = T^{\pi_1}F^{1\pi}. \tag{25b}$$

Especially, $F^{\delta^\infty} = T^\delta F^{\delta^\infty}$.

LEMMA 4. *If* W *is a nonnegative set, then*

(i) *For any* $\pi \in \Pi$, $F_n^\pi \geq F_{n+1}^\pi \geq F^\pi$, *and* $\lim_{n \to \infty} F_n^\pi = F^\pi$;

(ii) $F_n^* \geq F_{n+1}^* \geq F^*$ *and* $\lim_{n \to \infty} F_n^* = F^*$.

(iii) F^* *satisfies optimality equation*: $F^* = TF^*$.

Proof. (i) Let $(i, x) \in E$ be the initial state. Because the rewards are nonnegative,

$$\{B_{n+1} \leq x\} \subset \{B_n \leq x\}, \qquad \bigcap_{n=1}^{\infty} \{B_n \leq x\} = \{B \leq x\}.$$

Thus, $F_n^\pi \geq F_{n+1}^\pi \geq F^\pi$ and by the continuity of probability measure,

$$\begin{aligned} F^\pi(i, x) &= P_\pi(B \leq x \mid Y_1 = (i, x)) \\ &= \lim_{n \to \infty} P_\pi(B_n \leq x \mid Y_1 = (i, x)) = \lim_{n \to \infty} F_n^\pi(i, x), \end{aligned}$$

that is, $\lim_{n \to \infty} F_n^\pi = F^\pi$.

(ii) By (i), $F_n^* \geq F_{n+1}^* \geq F^*$. Hence, $\lim_{n\to\infty} F_n^*$ exists and $F^* \leq \lim_{n\to\infty} F_n^*$.

For arbitrary $\pi \in \Pi$, by $F_n^\pi \geq F_n^*$ and (i) we have $F^\pi = \lim_{n\to\infty} F_n^\pi \geq \lim_{n\to\infty} F_n^*$. Hence, from the arbitrariness of π, $F^* \geq \lim_{n\to\infty} F_n^*$. Therefore, $\lim_{n\to\infty} F_n^* = F^*$.

(iii) By Lemma 3, for any $\pi \in \Pi$, $F^\pi = T^{\pi_1}F^{1\pi} \geq TF^*$. It follows that

$$F^* \geq TF^*. \tag{26}$$

On the other hand, for any $(i, x) \in E$, $a \in A(i)$, by Theorem 1 and (16),

$$F_n^*(i, x) = TF_{n-1}^*(i, x) \leq GF_{n-1}^*(i, x, a) = \sum_{j\in S,\, r\in W} p_{ijr}^a F_{n-1}^*\left(j, \frac{x-r}{\beta}\right).$$

By dominated convergence theorem and (ii), we obtain

$$\begin{aligned} F^*(i, x) &= \lim_{n\to\infty} F_n^*(i, x) \leq \sum_{j\in S,\, r\in W} p_{ijr}^a \lim_{n\to\infty} F_{n-1}^*\left(j, \frac{x-r}{\beta}\right) \\ &= \sum_{j\in S,\, r\in W} p_{ijr}^a F^*\left(j, \frac{x-r}{\beta}\right) = GF^*(i, x, a). \end{aligned}$$

So, $F^*(i, x) \leq TF^*(i, x) = \min_{a\in A(i)} GF^*(i, x, a)$, that is, $F^* \leq TF^*$. Associating it with (26), we have $F^* = TF^*$. ∎

If W may not be a nonnegative set, then we have the following important results.

THEOREM 5. *Assume that constants* $b \leq \inf\{r \mid r \in W\}$ *and* $d \geq \sup\{r \mid r \in W\}$ *and let* $b_n = b\beta^n/(1-\beta)$, $d_n = d\beta^n/(1-\beta)$, $n \geq 1$. *Then,*

(i) *For any* $(i, x) \in E$, $F_n^*(i, x - b_n)$ *is a nonincreasing sequence and* $\lim_{n\to\infty} F_n^*(i, x - b_n) = F^*(i, x)$;

(ii) F^* *satisfies optimality equation*: $F^* = TF^*$;

(iii) $0 \leq F_n^*(i, x - b_n) - F^*(i, x) \leq F_n^*(i, x - b_n) - F_n^*(i, x - d_n)$, $(i, x) \in E$ $n \geq 1$;

(iv) *For each* $i \in S$, $F^*(i, x)$ *is a distribution function of* x;

(v) *There exists a policy* $\delta^\infty \in \Pi_s^d$ *such that* $T^\delta F^* = F^*$.

Proof. (i) First, we introduce a new MDP $\Gamma_2 = (E, A, \tilde{P}, \tilde{W}, \beta)$, where E and A are the same to Γ_1s, $\tilde{W} = W - b = \{r - b \mid r \in W\}$, and $\tilde{P}$ satisfies

$$\tilde{p}_{ij(r-b)}^a = p_{ijr}^a, \qquad i, j \in S,\ a \in A(i),\ r \in W.$$

Let $\tilde{F}_n^{\pi}, \tilde{F}^{\pi}$, and $\tilde{F}_n^{*}, \tilde{F}^{*}$ denote the corresponding objective functions generated by π and the optimal value functions, respectively.

Second, we prove that

$$F_n^*(i,x) = \tilde{F}_n^*(i, x - c_n), \qquad F^*(i,x) = \tilde{F}^*(i, x - c),$$
$$(i,x) \in E,\ n \geq 1. \tag{27}$$

where $c_n = \sum_{k=1}^{n} \beta^{k-1} b$ and $c = \sum_{k=1}^{\infty} \beta^{k-1} b$.

In fact, for any $\pi = (\pi_k,\ k \geq 1) \in \Pi$, define policies $\sigma = (\sigma_k,\ k \geq 1)$ and $\theta = (\theta_k,\ k \geq 1)$ as the following,

$$\sigma_k(\cdot \mid i_1, x_1, a_1, \ldots, i_{k-1}, x_{k-1}, a_{k-1}, i_k, x_k)$$
$$= \pi_k(\cdot \mid i_1, y_1, a_1, \ldots, i_{k-1}, y_{k-1}, a_{k-1}, i_k, y_k),$$
$$\theta_k(\cdot \mid i_1, x_1, a_1, \ldots, i_{k-1}, x_{k-1}, a_{k-1}, i_k, x_k)$$
$$= \pi_k(\cdot \mid i_1, z_1, a_1, \ldots, i_{k-1}, z_{k-1}, a_{k-1}, i_k, z_k),$$

where $y_k = x_k + c_{n-k+1}$, $k \leq n$; $y_k = x_k$, $k > n$ and $z_k = x_k + c$, $k \geq 1$.

Then, we have

$$\tilde{F}_n^{\sigma}(i, x - c_n) = F_n^{\pi}(i,x), \qquad \tilde{F}^{\theta}(i, x - c) = F^{\pi}(i,x), \qquad (i,x) \in E. \tag{28}$$

Similarly, for any $\pi \in \Pi$, if we define policies $\sigma = (\sigma_k,\ k \geq 1)$ and $\theta = (\theta_k,\ k \geq 1)$ by

$$\sigma_k(\cdot \mid i_1, x_1, a_1, \ldots, i_{k-1}, x_{k-1}, a_{k-1}, i_k, x_k)$$
$$= \pi_k(\cdot \mid i_1, y_1, a_1, \ldots, i_{k-1}, y_{k-1}, a_{k-1}, i_k, y_k),$$
$$\theta_k(\cdot \mid i_1, x_1, a_1, \ldots, i_{k-1}, x_{k-1}, a_{k-1}, i_k, x_k)$$
$$= \pi_k(\cdot \mid i_1, z_1, a_1, \ldots, i_{k-1}k, z_{k-1}, a_{k-1}, i_k, z_k),$$

where $y_k = x_k - c_{n-k+1}$, $k \leq n$; $y_k = x_k$, $k > n$, and $z_k = x_k - c$, $k \geq 1$, then, we have

$$F_n^{\sigma}(i,x) = \tilde{F}_n^{\pi}(i, x - c_n), \qquad F^{\theta}(i,x) = \tilde{F}^{\pi}(i, x - c), \qquad (i,x) \in E. \tag{29}$$

Thus, by (28) and (29), we obtain (27). By (27), we have

$$F_n^*(i, x - b_n) = \tilde{F}_n^*(i, x - c), \qquad (i,x) \in E,\ n \geq 1.$$

Note that $\tilde{W}$ is a nonnegative set, by Lemma 4 and (27), (i) is proved.

(ii) For MDP Γ_2, let $\tilde{T}$ denote the corresponding operator T. Then, by Lemma 4, $\tilde{F}^* = \tilde{T}\tilde{F}^*$. Hence, for any $(i, x) \in E$, by (27), we have

$$\begin{aligned} F^*(i,x) &= \tilde{F}^*(i, x-c) = \tilde{T}\tilde{F}^*(i, x-c) \\ &= \min_{a \in A(i)} \sum_{j \in S,\, r \in W} \tilde{p}^a_{ij(r-b)} \tilde{F}^* \left(j, \frac{x - c - (r-b)}{\beta} \right) \\ &= \min_{a \in A(i)} \sum_{j \in S,\, r \in W} p^a_{ijr} \tilde{F}^* \left(j, \frac{x-r}{\beta} - c \right) \\ &= \min_{a \in A(i)} \sum_{j \in S,\, r \in W} p^a_{ijr} F^* \left(j, \frac{x-r}{\beta} \right) \\ &= TF^*(i,x). \end{aligned}$$

Therefore $F^* = TF^*$.

(iii) The first inequality is obtained by (i). To obtain the second inequality, it suffices to prove that $F_n^*(i, x - d_n) \le F^*(i, x)$ for all $(i, x) \in E$. Thus, we need only to prove that for any $\pi \in \Pi$ there exists $\sigma \in \Pi$ such that $F_n^\sigma(i, x - d_n) \le F^\pi(i, x)$ for all $(i, x) \in E$.

In fact, fixing $n \ge 1$, for any $\pi = (\pi_k, k \ge 1) \in \Pi$, we define

$$\begin{aligned} &\sigma_k(\cdot \mid i_1, x_1, a_1, \ldots, i_{k-1}, x_{k-1}, a_{k-1}, i_k, x_k) \\ &\quad = \pi_k(\cdot \mid i_1, y_1, a_1, \ldots, i_{k-1}, y_{k-1}, a_{k-1}, i_k, y_k), \end{aligned}$$

where $y_k = x_k + d_n \beta^{-k+1}$, $k \ge 1$. Then $P_\sigma(B \le x \mid i, x - d_n) = P_\pi(B \le x \mid i, x)$ for all $(i, x) \in E$. Because $B \le B_n + d_n$, we obtain

$$\begin{aligned} F_n^\sigma(i, x - d_n) &= P_\sigma(B_n \le x - d_n \mid i, x - d_n) \\ &\le P_\sigma(B \le x \mid i, x - d_n) = P_\pi(B \le x \mid i, x) = F^\pi(i, x). \end{aligned}$$

(iv) For each $i \in S$ and $x < y$, by Theorem 1, $F_n^*(i, x - b_n) \le F_n^*(i, y - b_n)$. It follows that, by letting $n \to \infty$, $F^*(i, x) \le F^*(i, y)$, that is, $F^*(i,x)$ is a nondecreasing function.

From $F^*(i, x) \le F^*(i, y) \le F_n^*(i, y - b_n)$ and the right continuity of $F_n^*(i, x)$, letting $y \to x (y > x)$, we obtain

$$F^*(i, x) \le F^*(i, x + 0) \le F_n^*(i, x - b_n).$$

In addition, letting $n \to \infty$ again, by (i), we have

$$F^*(i, x) \le F^*(i, x + 0) \le F^*(i, x).$$

that is, $F^*(i, x + 0) = F^*(i, x)$. So, $F^*(i, x)$ is right continuous.

Thus, by (13), $F^*(i, x)$ is a distribution function.

(v) As $A(i, x) = A(i)$ are finite for all $(i, x) \in E$, by the measurable selection theorem [7, Proposition D3, p. 130], there exists a measurable mapping δ from E to A such that $\delta(i, x) \in A(i)$ and $GF^*(i, x, \delta(i, x)) = F^*(i, x)$ for all $(i, x) \in E$. That is to say that $\delta^\infty \in \Pi_s^d$ and $T^\delta F^* = F^*$.

The proof of Theorem 5 is complete. ▌

Remark 3. If W is a nonnegative set and let $b = 0$, Theorem 5(i) is just Lemma 4(ii). In addition, Theorem 5(iii) is the generalization of Theorem 4.10 in [1]. But for the case in which the reward may take a negative value, the approximation relation $\lim_{n\to\infty} F_n^* = F^*$ does not hold. See the following example.

EXAMPLE 1. Let $S = \{1, 2\}$, $A(1) = A(2) = \{a, b\}$, $W = \{-2, -1\}$, $p^a_{11(-1)} = p^a_{12(-1)} = p^a_{21(-2)} = p^a_{22(-2)} = p^b_{11(-2)} = p^b_{12(-2)} = p^b_{21(-1)} = p^b_{22(-1)} = 0.5$, and $\beta = 0.5$. Then

$$F_n^*(i, x) = I_{[-2+0.5^{n-1}, +\infty)}(x), \qquad F^*(i, x) = I_{[-2, +\infty)}(x).$$

Thus, $\lim_{n\to\infty} F_n^*(i, x) = I_{(-2,\infty)}(x) \ne F^*(i, x)$.

Let $A^*(i, x) = \{a \mid a \in A(i) \text{ and } GF^*(i, x, a) = F^*(i, x)\}$ for each $(i, x) \in E$. Then by the finiteness of $A(i)$ and Theorem 5, $A^*(i, x) \ne \varnothing$ for all $(i, x) \in E$.

THEOREM 6. *Let* $\pi = (\pi_k, k \ge 1) \in \Pi$ *and* $\delta^\infty \in \Pi_s$. *Then,*

(i) $F^\pi = F^*$ *if and only if for all* $(i, x) \in E$, $\pi_1(A^*(i, x) \mid i, x) = 1$ *and* $F^{1_\pi(i,x,a)}(j, (x - r)/\beta) = F^*(j, (x - r)/\beta)$ *when* $\pi_1(a \mid i, x)p^a_{ijr} > 0$.

(ii) *If* $F^\pi = F^*$ *and* $T^\delta F^* = F^*$, *then for any* $n \ge 1$, $F^{(\delta^n, \pi)} = F^*$.

Proof. (i) By Lemmas 1, 3, and Theorem 5, using the approach similar to the proof of Theorem 3, we obtain (i) immediately.

(ii) Let $F^\pi = F^*$ and $T^\delta F^* = F^*$. By Lemma 3, we have $F^{(\delta,\pi)} = T^\delta F^\pi = T^\delta F^* = F^*$. It follows that $F^{(\delta^n,\pi)} = F^*$ for any $n \ge 1$. ▌

Similarly, as a result of Theorem 6, $A^*(i, x)$ and it's elements are called an optimal action set and optimal actions, respectively.

Remark 4. For an infinite horizon model, even if δ^∞ is a policy which takes optimal actions and (δ^n, π) is optimal for any $n \ge 1$, δ^∞ may not be optimal (see the following example).

Now, we give an example which illustrates that $F_n^\pi(i, x)$ and $F^\pi(i, x)$ may not be the distribution functions of x. This example is also a counterexample for Lemma 3 in [3].

EXAMPLE 2. Let $S = \{1, 2\}$, $A(1) = A(2) = \{a, b\}$, $W = \{1, 2\}$, $p_{111}^a = p_{121}^a = p_{212}^a = p_{222}^a = p_{112}^b = p_{122}^b = p_{211}^b = p_{221}^b = 0.5$, and $\beta = 0.5$. Then

$$F^*(i, x) = I_{[4, +\infty)}(x),$$

and

$$A^*(1, x) = \begin{cases} \{a, b\}, & x < 3 \quad \text{or} \quad x \geq 4 \\ \{b\}, & 3 \leq x < 4 \end{cases},$$

$$A^*(2, x) = \begin{cases} \{a, b\}, & x < 3 \quad \text{or} \quad x \geq 4 \\ \{a\}, & 3 \leq x < 4 \end{cases}.$$

Let

$$\theta(1, x) = b, \qquad \theta(2, x) = a, \qquad x \in (-\infty, +\infty),$$

$$\rho(1, x) = \begin{cases} a, & x < 3 \\ b, & x \geq 3 \end{cases}, \qquad \rho(2, x) = \begin{cases} b, & x < 3 \\ a, & x \geq 3 \end{cases},$$

and

$$\mu(1, x) = \begin{cases} a, & x < 2.5 \\ b, & x \geq 2.5 \end{cases}, \qquad \mu(2, x) = \begin{cases} b, & x < 2.5 \\ a, & x \geq 2.5 \end{cases}.$$

Then, $T^\theta F^* = T^\rho F^* = T^\mu F^* = F^*$. But, $F^{\theta^\infty} = F^*$, $F^{\rho^\infty}(i, x) = I_{[2, +\infty)}(x)$, $F^{\mu^\infty}(i, 2) = 1$, $F^{\mu^\infty}(i, 2.6) = 0$, $F_4^{\mu^\infty}(i, 3) = 1$, $F_4^{\mu^\infty}(i, 3.25) = 0$, $i = 1, 2$, that is, θ^∞ is an optimal policy, ρ^∞ and μ^∞ both are not optimal, and $F_4^{\mu^\infty}(i, x)$ and $F^{\mu^\infty}(i, x)$ are not distribution functions of x.

Example 2 shows that

(i) $F_n^\pi(i, x)$ and $F^\pi(i, x)$ may not be distribution functions of x. For an infinite horizon model, policy δ^∞ which takes optimal action at each realizable state, that is, $T^\delta F^* = F^*$, may not be optimal.

(ii) The proof of Theorem 4.10 in [1] is not correct because that $F_n^\pi(i, x - d_n) \leq F^\pi(i, x)$ may not hold. In fact, let $\pi = \mu^\infty$, $x = 3.25$, and $n = 4$, then $F_n^\pi(i, x - d_n) = 1$ and $F^\pi(i, x) = 0$.

(iii) Lemma 3 in [3] does not hold. In fact, Lemma 3 in [3] implies that equation $T^\delta u = u$ has an unique solution in $\mathscr{F}$, where $\mathscr{F} = \{u: u \in D$ and for any $i \in S$, $u(i, x)$ is a distribution function such that $u(i, x) = 0$ if $x < 0$ and $u(i, x) = 1$ if $x \geq H/(1 - \beta)\}$ and $\delta^\infty \in \Pi_s^d$. In [3], W is nonnegative and H is an upbound of W. However, in the foregoing

example, by Lemma 3, $T^{\rho}u = u$ has at least two solutions $F^{\rho^{\infty}}$ and $F^* = F^{\theta^{\infty}}$ in $\mathscr{F}$.

Thus, some main results in [3] (e.g., Theorem 1(ii) and (iii) and Theorem 4 in [3]) have not been proved really. Although Lemma 3 in [3] does not hold for the general case, we point out that it is correct under a special condition in Lemma 5.

Let $\mathscr{F} = \{u \mid u \in D$ and for each $i \in S$, $u(i, x)$ is a distribution function of x satisfying $u(i, x) = 0$ if $x < b/(1 - \beta)$ and $u(i, x) = 1$ if $x \geq d/(1 - \beta)\}$, where $b = \inf\{r \mid r \in W\}$ and $d = \sup\{r \mid r \in W\}$.

LEMMA 5. *Let $u, v \in \mathscr{F}$, $\delta^{\infty} \in \Pi_0$, and $u - v \geq T^{\delta}(u - v)$. Then $u \geq v$.*

Proof. Let $b_n = \beta^n b/(1 - \beta)$ and $d_n = \beta^n d/(1 - \beta)$, where $b = \inf\{r \mid r \in W\}$ and $d = \sup\{r \mid r \in W\}$. Then,

$$(u - v)(i, x) \geq (T^{\delta})^n(u - v)(i, x) \geq F_n^{\delta^{\infty}}(i, x - d_n) - F_n^{\delta^{\infty}}(i, x - b_n). \tag{30}$$

In fact, the first inequality follows from Lemma 1. Now, we prove the second inequality by induction. When $n = 1$, it is easy to prove because $u, v \in \mathscr{F}$. We assume that the second inequality holds when $n = k$. Then, for $n = k + 1$, by Theorem 1, we have

$$\begin{aligned}
&(T^{\delta})^{k+1}(u - v)(i, x) \\
&\quad = T^{\delta}\left[(T^{\delta})^k(u - v)\right](i, x) \\
&\quad = \sum_{a \in A(i)} \delta(a \mid i) \sum_{i \in S,\, r \in W} p_{ijr}^a (T^{\delta})^k(u - v)\left(j, \frac{x - r}{\beta}\right) \\
&\quad \geq \sum_{a \in A(i)} \delta(a \mid i) \sum_{i \in S,\, r \in W} p_{ijr}^a \left(F_k^{\delta^{\infty}}\left(j, \frac{x - r}{\beta} - d_k\right)\right. \\
&\qquad\qquad \left. - F_k^{\delta^{\infty}}\left(j, \frac{x - r}{\beta} - b_k\right)\right) \\
&\quad = \sum_{a \in A(i)} \delta(a \mid i) \sum_{i \in S,\, r \in W} p_{ijr}^a \left(F_k^{\delta^{\infty}}\left(j, \frac{x - d_{k+1} - r}{\beta}\right)\right. \\
&\qquad\qquad \left. - F_k^{\delta^{\infty}}\left(j, \frac{x - b_{k+1} - r}{\beta}\right)\right) \\
&\quad = F_{k+1}^{\delta^{\infty}}(i, x - d_{k+1}) - F_{k+1}^{\delta^{\infty}}(i, x - b_{k+1}).
\end{aligned}$$

The second inequality holds for $n = k + 1$. By induction, (30) is proved.

Owing to $B_n + b_n \le B \le B_n + d_n$ and $\delta^\infty \in \Pi_0$, we have

$$F^{\delta^\infty}(i, x - d_n + b_n) \le F_n^{\delta^\infty}(i, x - d_n) \le F_n^{\delta^\infty}(i, x - b_n) \le F^{\delta^\infty}(i, x + d_n - b_n). \quad (31)$$

Thus, if x is a point of continuity of $F^{\delta^\infty}(i, \cdot)$, letting $n \to \infty$ in (30), by (31), we have $(u - v)(i, x) \ge 0$. But u and v are right continuous and the set of all discontinuity points of u and v combined is countable, by Lemma 1 in [3], $(u - v)(i, x) \ge 0$ for all $x \in \mathbb{R}$. Lemma 5 is proved. ∎

COROLLARY 2. *Let $\delta^\infty \in \Pi_0$. Then, F^{δ^∞} is the unique solution of the equation $T^\delta u = u$ in $\mathscr{F}$.*

Proof. By Lemma 3, F^{δ^∞} is a solution of $T^\delta u = u$ in $\mathscr{F}$. Assume that $u \in \mathscr{F}$ and $T^\delta u = u$. Then, $F^{\delta^\infty} - u = T^\delta(F^{\delta^\infty} - u)$. By Lemma 5, we obtain $u = F^{\delta^\infty}$. ∎

Remark 5. What differs from Lemma 3 in [3] is that Lemma 5 requires $\delta^\infty \in \Pi_0$. The technique of the proof of Lemma 5 comes from Lemma 3's in [3].

Theorem 7 states another main result in this paper.

THEOREM 7. *There exists a policy $\pi \in \Pi_0$ such that $F^\pi = F^*$ if and only if $A^*(i) = \bigcap_{x \in \mathbb{R}} A^*(i, x) \ne \varnothing$ for all $i \in S$.*

Proof. $\Rightarrow$. Let $\pi \in \Pi_0$ and $F^\pi = F^*$. Then, by Theorem 6, $\pi_1(A^*(i, x) \mid i) = 1$ for all $x \in \mathbb{R}$ and $i \in S$. Note that $A(i)$ is finite, it follows that $\pi_1(A^*(i) \mid i) = 1$. Hence, $A^*(i) \ne \varnothing$ for all $i \in S$.

$\Leftarrow$. Let $A^*(i) \ne \varnothing$ for all $i \in S$. Select δ: $S \to A$ such that $\delta(i) \in A^*(i)$ for all $i \in S$. Then, $\pi = \delta^\infty \in \Pi_0$ and $T^\delta F^* = F^*$. By Corollary 2, we obtain $F^\pi = F^*$. ∎

COROLLARY 3. *If $A^*(i) \ne \varnothing$ for all $i \in S$, then $F^*(i, x) = \inf_{\pi \in \Pi_0} F^\pi(i, x)$, $(i, x) \in E$.*

Remark 6. Theorem 7 gives a sufficient condition under which there exists an optimal policy. Because Lemma 5 does not hold for general policy $\delta^\infty \in \Pi$, for an infinite horizon model, whether there exists an optimal policy remains an open problem.

5. ALGORITHM

In this section we give an algorithm computing optimal value functions, optimal action sets, and optimal policies for a finite horizon model.

In this section, we assume that S and W are finite and we let $W = \{r_1, r_2, \ldots, r_m\}$ and $r_1 < r_2 < \cdots < r_m$. Then, by Theorems 1, 2, and the finiteness of S, A, and W, we have the following conclusions:

(i) For each $i \in S$ and $n \geq 1$, $F_n^*(i, x)$ is a step distribution function of x with finite jump points;

(ii) For each $i \in S$ and $n \geq 1$, $A_n(i, x)$ is a set-valued function from $\mathbb{R}$ to $A(i)$ with finite discontinuity points;

(iii) For each $n \geq 1$, there exists an n stages optimal deterministic Markov policy which kth decision rule has the structure analogous to that of $F_k^*(i, x)$ and $A_k(i, x)$, $1 \leq k \leq n$.

The following algorithm is just the proof of the earlier conclusions.

By Theorem 1,

$$\begin{aligned} F_0^*(i, x) &= I_{[0, +\infty)}(x), \\ F_n^*(i, x) &= \min_{a \in A(i)} \sum_{j \in S,\, r \in W} p_{ijr}^a F_{n-1}^*\left(j, \frac{x - r}{\beta}\right), \qquad i \in S, n \geq 1. \end{aligned} \tag{32}$$

Let

$$\begin{aligned} b_n(i, x, a) &= \sum_{j \in S,\, r \in W} p_{ijr}^a F_{n-1}^*\left(j, \frac{x - r}{\beta}\right), \qquad i \in S, a \in A(i), \\ I_n(i, x) &= \min_{a \in A(i)} b_n(i, x, a), \qquad i \in S. \end{aligned}$$

ALGORITHM.

Step 1. Calculate

$$\begin{aligned} b_1(i, r_k, a) &= \sum_{j \in S} \sum_{r \leq r_k} p_{ijr}^a, \qquad i \in S, a \in A(i), \\ I_1(i, r_k) &= \min_{a \in A(i)} b_1(i, r_k, a), \qquad i \in S, \\ A_1(i, r_k) &= \{a \mid a \in A(i), b_1(i, r_k, a) = I_1(i, r_k)\}, \qquad i \in S, \end{aligned}$$

and select $\delta_1(i, r_k) \in A_1(i, r_k)$, $k = 1, \ldots, m-1$, $\delta_1(i, r_m) \in A(i)$. Then, by (32) and (22),

$$F_1^*(i, x) = \begin{cases} 0, & x < r_1, \\ I_1(i, r_k), & r_k \le x < r_{k+1}, k = 1, \ldots, m-1, \\ 1, & x \ge r_m, \end{cases}$$

$$A_1(i, x) = \begin{cases} A(i), & x < r_1 \quad \text{or} \quad x \ge r_m, \\ A_1(i, r_k), & r_k \le x < r_{k+1}, k = 1, \ldots, m-1. \end{cases}$$

Let

$$\delta_1(i, x) = \begin{cases} \delta_1(i, r_m), & x < r_1 \quad \text{or} \quad x \ge r_m, \\ \delta_1(i, r_k), & r_k \le x < r_{k+1}, k = 1, \ldots, m-1. \end{cases}$$

Step 2. Assume that F_{l-1}^*, A_{l-1}, and δ_{l-1} have been obtained and the all jump points of F_{l-1}^* are $t_1 < t_2 < \cdots < t_M$. Arranging $\beta t_k + r_l$ $(k = 1, 2, \ldots, M,\ l = 1, 2, \ldots, m)$ in ascending order and denoting them by $u_1 < u_2 < \cdots < u_N$ $(N \le mM)$, then, for any $j \in S$ and $r \in W$, we have

$$F_{l-1}^*\left(j, \frac{x-r}{\beta}\right) = \begin{cases} 0, & x < u_1, \\ F_{l-1}^*\left(j, \frac{u_k - r}{\beta}\right) & u_k \le x < u_{k+1}, k = 1, \ldots, N-1, \\ 1, & x \ge u_N. \end{cases} \tag{33}$$

Calculate

$$b_l(i, u_k, a) = \sum_{j \in S,\, r \in W} p_{ijr}^a F_{l-1}^*\left(j, \frac{u_k - r}{\beta}\right), \qquad i \in S,\ a \in A(i),$$

$$I_l(i, u_k) = \min_{a \in A(i)} b_l(i, u_k, a), \qquad i \in S,$$

$$A_l(i, u_k) = \{a \mid a \in A(i),\ b_l(i, u_k, a) = I_l(i, u_k)\}, \qquad i \in S,$$

and select $\delta_l(i, u_k) \in A_l(i, u_k)$, $k = 1, \ldots, N-1$, $\delta_l(i, u_N) \in A(i)$. Then, by (33), (32), and (22),

$$F_l^*(i, x) = \begin{cases} 0, & x < u_1, \\ I_l(i, u_k), & u_k \le x < u_{k+1}, k = 1, \ldots, N-1, \\ 1, & x \ge u_N, \end{cases}$$

$$A_l(i, x) = \begin{cases} A(i), & x < u_1 \quad \text{or} \quad x \ge u_N, \\ A_l(i, u_k), & u_k \le x < u_{k+1}, k = 1, \ldots, N-1, \end{cases}$$

Let

$$\delta_l(i, x) = \begin{cases} \delta_l(i, u_N), & x < u_1 \quad \text{or} \quad x \ge u_N, \\ \delta_l(i, u_k), & u_k \le x < u_{k+1}, k = 1, \ldots, N-1. \end{cases}$$

Repeating Step 2 up to $l = n$, we obtain the optimal function F_n^* and an optimal policy $\pi = (\delta_n, \delta_{n-1}, \ldots, \delta_1)^\infty$. In addition, $A_1(i, x), A_2(i, x), \ldots, A_n(i, x)$ are the corresponding optimal action sets, from them all n stages optimal policies can be obtained.

REFERENCES

1. M. Bouakiz and Y. Kebir, Target-level criterion in Markov decision processes, *J. Optim. Theory Appl.* **86** (1995), 1–15.
2. D. J. White, Mean, variance, and probabilistic criterion in finite Markov decision processes: A review, *J. Optim. Theory Appl.* **56** (1988), 1–29.
3. D. J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173** (1993), 634–646.
4. M. J. Sobel, The variance of discounted Markov decision processes, *J. Appl. Probab.* **19** (1982), 794–802.
5. K. J. Chung and M. J. Sobel, Discounted MDP's: Distribution functions and exponential utility maximization, *SIAM J. Contr. Optim.* **25** (1987), 49–62.
6. J. A. Filar, D. Krass, and K. W. Ross, Percentile performance criteria for limiting average Markov decision processes, *IEEE Trans. Automat. Contr.* **40** (1995), 2–10.
7. O. Hernández-Lerma, "Adaptive Markov Control Processes," Springer-Verlag, New York, 1989.