International Journal of Pure and Applied Mathematics
Volume 59 No. 2 2010, 225-242

OPTIMAL THRESHOLD PROBABILITY AND POLICY ITERATION IN SEMI-MARKOV DECISION PROCESSES

Masahiko Sakaguchi[1], Yoshio Ohtsubo[2] §

[1]Graduate School of Integrated Arts and Sciences
Kochi University
2-5-1, Akebono-cho, Kochi, 780-8520, JAPAN

[2]Department of Mathematics
Faculty of Science
Kochi University
2-5-1, Akebono-cho, Kochi, 780-8520, JAPAN
e-mail: ohtsubo@kochi-u.ac.jp

Abstract: We consider undiscounted semi-Markov decision process with a target set and our main concern is a problem minimizing threshold probability. We formulate the problem as an infinite horizon case with a recurrent class. We show that an optimal value function is a unique solution to an optimality equation and there exists a stationary optimal policy. Also several value iteration methods and a policy improvement method are given in our model.

AMS Subject Classification: 90C40, 90C31
Key Words: semi-Markov decision process, optimal threshold probability, existence of optimal policy, value iteration, policy improvement method

1. Introduction

Semi-Markov decision processes are investigated in many references. Howard in [12] extends the theory of Markov decision processes to semi-Markov type with an average reward per unit time in the steady state. Such an average case is expanded by Ross [17] and Federgruen et al [7, 8]. For the expectation of total discounted reward in semi-Markovian type, Yasuda in [22] proves that there exists an optimal stationary policy and give the optimality equation.

Received: January 29, 2010

§Correspondence author

Bhattacharya and Majumdar in [1] and Feinberg in [9] obtain analogous results for general semi-Markov models with unbounded rewards and for multiple discounted criterion, respectively.

Undiscounted Markov decision processes with a target set are very important optimization problem. Eaton and Zadeh [6] formulate such a Markov decision process as a pursuit problem and show that the expected total cost corresponding to an optimal policy is a unique solution to an optimality equation on some conditions. Derman in [4, 5] considers finite Markov decision processes called a first passage problem and proves that the problem has an optimal stationary policy. Blackwell in [2] introduces a concept of a nearly optimal policy and he in [3] investigates positive Markov decision processes. Also Veinott [18] generalizes some results of [6] and [4] to transient Markov decision processes. Hernández-Lerma and Lasserre [10] further extend it to the transient case with Borel state spaces and compact action spaces.

On the other hand, several authors consider risk models in which we minimize a threshold probability $P_i^\pi(V_\beta \leq w)$ with respect to policy π, where $V_\beta = \sum_{k=1}^{\infty} \beta^k Y_k$ is a total discounted reward, w is a threshold value and i is an initial state. In [19, 20], White considers such a problem in finite Markov decision processes with a bounded reward set. Wu and Lin [21] prove that the optimal value functions for finite and infinite horizon cases are distributions of the threshold value, and also show that there exists an optimal deterministic Markov policy in a finite horizon model. Ohtsubo and Toyonaga [13] give two sufficient conditions for the existence of an optimal right continuous stationary policy in an infinite horizon case. All of these problems are concerned with a discounted case and correspond to one of the first equivalence class which is given in Ohtsubo and Toyonaga [14]. In [15], the author investigates minimizing risk models in a stochastic shortest path problem. The threshold probability is $P_i^\pi(V > w)$ where $V = \sum_{k=1}^{\infty} Y_k$ is a total undiscounted cost. This criterion corresponds to a problem of the second equivalence class in [14]. He in [16] also formulates the problem, which minimizes a threshold probability $P_i^\pi(V \leq w)$, as undiscounted finite Markov decision processes, proves that the optimal value function is a unique solution to an optimality equation and shows that there exists an optimal right continuous stationary policy.

Our aims are to generalize the results of Ohtsubo [16] to semi-Markovian case. Markov decision processes are special cases of semi-Markov decision processes, in which the distribution of every sojourn time is degenerated at unit time.

In this paper we formulate semi-Markov decision processes with countable

state space S and action space A and we denote a state and a action of the system at time t by $X(t)$ and $A(t)$, respectively. Our main concern is minimizing risk models with the threshold probability criterion $P_i^\pi(V \leq w)$, where $V = \int_0^\zeta r(X(t), A(t), t)dt$, r is a given function on $S \times A \times R$ and ζ is a first passage time to a given target set. Under a finiteness of ζ we show that the optimal value function is a unique solution to an optimality equation and that there exists an optimal right continuous stationary policy. Also we give several value iteration methods and a policy improvement method.

As real world situations described by such semi-Markov decision processes we can find many applications such as queueing systems, inventory-production systems and equipment maintenance problems.

2. Notations and Formulation

In this section we formulate our minimizing risk models as undiscounted semi-Markov decision processes $\Gamma = (Z(t), A(t), Y_n)$ with (S, A, r, Q):

(i) The state space S is countable;

(ii) The action space $A = \bigcup_{i \in S} A(i)$ is countable, where $A(i)$ is a nonempty set of admissible finite actions when the system is in state $i \in S$;

(iii) The reward r is a real-valued function on $S \times A \times \mathbb{R}$, which is nonnegative, bounded and integrable on any finite interval, and has a norm $||r|| := \sup\{|r(i, a, s)| \mid i \in S,\ a \in A,\ s \in \mathbb{R}\}$.

The random variables X_n, A_n and τ_n denote the state, the action and the time at the beginning of transition unit $n (n \geq 0)$. We define $\tau_0 = 0$.

(iv) If we take an action a in state i, the system obeys the semi-Markov kernel

$$Q_{ij}^a(t) = p_{ij}^a G_{ij}^a(t) = P(\tau_{n+1} - \tau_n \leq t, X_{n+1} = j | X_n = i, A_n = a),$$

where p_{ij}^a is the transition probability from state i to state j and $G_{ij}^a(t)$ is the distribution of the sojourn time in state i knowing that the next visiting state is j:

$$\begin{aligned} p_{ij}^a &= P(X_{n+1} = j \mid X_n = i, A_n = a), \\ G_{ij}^a(t) &= P(\tau_{n+1} - \tau_n \leq t \mid X_{n+1} = j, X_n = i, A_n = a). \end{aligned}$$

The following assumption is needed through this paper.

Assumption 2.1. For any $a \in A$ and $i, j \in S$, $G_{ij}^a(t) = 0$ if $t \leq 0$.

Lemma 2.1. *Under Assumption* 2.1, $P(\tau_{n+1} > \tau_n) = 1$ *for all* n.

Proof. It is obvious. □

The following is well defined under Assumption 2.1. Let $\xi(\omega) = \lim_{n\to\infty} \tau_n(\omega)$ and then for $t \geq 0$ we define

$$Z(t)(\omega) = \begin{cases} X_n(\omega) & \text{if } \tau_n(\omega) \leq t < \tau_{n+1}(\omega), \\ \Delta_s & \text{if } \xi(\omega) \leq t, \end{cases}$$

$$A(t)(\omega) = \begin{cases} A_n(\omega) & \text{if } \tau_n(\omega) \leq t < \tau_{n+1}(\omega), \\ \Delta_a & \text{if } \xi(\omega) \leq t, \end{cases}$$

where Δ_s and Δ_a are artificial points added to S and A respectively in the usual convention. However we do not really need them for Assumptions 2.2 and 2.3.

We shall consider the reward associated with semi-Markov decision processes. We suppose that the system obtains the reward $r(i, a, t)$ when it stays state i and we take action a at time t. Then the system obtains the unit reward

$$Y_n = \int_{\tau_n}^{\tau_{n+1}} r(Z(t), A(t), t)dt, \qquad n \geq 0.$$

Let a target set B be a nonempty subset of S and let a stopping time ζ be the smallest nonnegative real number t such that $Z(t) \in B$, where $\zeta = \infty$ if there does not exist such a real number t. We define the random total undiscounted reward by

$$V = \int_0^{\zeta} r(Z(t), A(t), t)dt.$$

Then our problem is to minimize a threshold probability $P_i^{\pi}(V \leq w)$ with respect to all policies π for a given threshold value w. To simplify the minimizing problem, we can redefine the semi-Markov decision processes as follows:

Assumption 2.2. B is a recurrent class, and it is reward-free, that is,

$$\sum_{j\in B} p_{ij}^a = 1 \quad \text{and} \quad r(i, a, t) = 0$$

for every $i \in B$, $a \in A(i)$ and all $t \geq 0$

Under Assumption 2.2 we have

$$V = \int_0^{\infty} r(Z(t), A(t), t)dt.$$

In order to analyse our problem we define the random total undiscounted reward

for a finite horizon case by

$$V_0 = 0, \quad V_n = \int_0^{\tau_n} r(Z(t), A(t), t)dt, \quad n \geq 1.$$

Further we define another random sequence by

$$W_0 = w, \quad W_n = W_0 - V_{n-1} = W_{n-1} - Y_{n-1}, \ n \geq 1,$$

where w is a given initial threshold value.

We use $S_R = S \times \mathbb{R}$ as a new state space where $\mathbb{R} = (-\infty, \infty)$. Let $H_0 = S_R$ and $H_n = H_{n-1} \times S_R$ for each $n \in \mathbb{N}$ where $\mathbb{N} = \{1, 2, 3, \cdots\}$. Then H_n represents the set of all possible histories $h_n = (i_0, w_0, i_1, w_1, \ldots, i_n, w_n)$ of the system when the n-th action must be chosen, and we denote by θ_n the history at time $n \geq 0$: $\theta_n = (X_0, W_0, X_1, W_1, \ldots, X_n, W_n)$. A decision rule f_n for time $n \geq 0$ is a mapping of H_n into A. We denote by Δ the set of all decision rules. A policy π is an infinite sequence of decision rules $(f_n, n \geq 0) = (f_0, f_1, \ldots, f_n, \ldots)$. We denote by C the set of all such policies.

A policy $\pi = (f_n, n \geq 0)$ is said to be Markov when the decision rule f_n is a function of $(X_n, W_n) = (i_n, w_n)$ for every $n \geq 0$. We denote the set of such decision rules by Δ_M and the set of all Markov policies by C_M. A policy $(g, f_0, f_1, \ldots, f_n, \ldots)$ is denoted by (g, π), where $g \in \Delta_M$ and $\pi \in C_M$. When $f_n = f$ for all $n \geq 0$, we write $\pi = \{f\}^\infty$, which is called a stationary policy, and we denote the set of all stationary policies by C_S.

We denote by $P_i^\pi(V \leq w)$ the conditional probability of event $\{V \leq w\}$ give an initial state $X_0 = i$ and a policy π. Since the random variable V depends upon not only i and π but also w, we may use a conditional probability $P_{(i,w)}^\pi(\cdot)$. Through this paper we assume the following.

Assumption 2.3. For every $\pi \in C$ and each $(i, w) \in S_R$,

$$P_{(i,w)}^\pi(\zeta < \infty) = 1, \ P_{(i,w)}^\pi(Z(t) \in B \text{ for some } t \in \mathbb{R}^+) = 1,$$

which means that the complement set B^c is a transient class when we use any policy $\pi \in C$.

Thus we easily see that $P_{(i,w)}^\pi(V < \infty) = 1$ for all $\pi \in C$ and each $(i, w) \in S_R$.

A decision rule $f \in \Delta_M$ is said to be right continuous (on $\mathbb{R}$) if for each $(i, w) \in S_R$ there is a positive real number μ such that $f(i, w) = f(i, w + u)$ for all u: $0 \leq u < \mu$. A policy $\pi = (f_n, n \geq 0) \in C_M$ is said to be right continuous if decision rule f_n is right continuous for every $n \geq 0$.

Our problem is to minimize a threshold probability $P_i^\pi(V \leq w)$ with respect to all policies π. We denote criterion functions for finite and infinite horizon

cases by

$$F_n^\pi(i,w) = P_i^\pi(V_n \le w), \quad F^\pi(i,w) = P_i^\pi(V \le w),$$

respectively, for each$(i,w) \in S_R$ and $\pi \in C$. We also define optimal value functions F_n^* and F^* for finite and infinite horizon cases by, respectively,

$$F_n^*(i,w) = \inf_{\pi \in C} F_n^\pi(i,w), \quad F^*(i,w) = \inf_{\pi \in C} F^\pi(i,w).$$

A policy π is said to be optimal if $F^*(i,w) = F^\pi(i,w)$ for every $(i,w) \in S_R$.

We define the following sets of functions: let $\mathcal{G}$ be the set of functions F from S_R into a finite interval such that $F(i,\cdot)$ is measurable on $\mathbb{R}$. Let $\mathcal{F}$ be the set of functions $F \in \mathcal{G}$ from S_R into $[0,1]$ such that $F(i,w) = 0$ for each $i \in S$ and every $w < 0$, and let $\mathcal{F}_r$ be the set of functions $F \in \mathcal{F}$ such that $F(i,\cdot)$ is monotone nondecreasing and right continuous on $\mathbb{R}$ for each $i \in S$. In Theorem 3.1 it is shown that $F^* \in \mathcal{F}_r$. However, it is not necessarily true that $F^\pi \in \mathcal{F}_r$ for each $\pi \in C$. We give a policy π such that $F^\pi \notin \mathcal{F}_r$ in Example 5.1.

We define operators L^a, L^f, U from $\mathcal{G}$ into itself as follows. For $F \in \mathcal{G}$, $(i,w) \in S_R$, $a \in A(i)$ and $f \in \Delta_M$,

$$L^a F(i,w) = \sum_{j \in S} \int_{t \in \mathbb{R}^+} F\left(j, w - \int_0^t r(i,a,s)ds\right) Q_{ij}^a(dt),$$

$$L^f F(i,w) = \sum_{j \in S} \int_{t \in \mathbb{R}^+} F\left(j, w - \int_0^t r(f)(i,w,s)ds\right) \times p^f(j \mid i,w) G^f(dt \mid i,w,j),$$

$$UF(i,w) = \inf_{f \in \Delta_M} [L^f F(i,w)] = \min_{a \in A(i)} L^a F(i,w).$$

where

$$r(f)(i,w,s) = r(i,f(i,w),s), \quad p^f(j|i,w) = p_{ij}^{f(i,w)}$$

and

$$G^f(t|i,w,j) = G_{ij}^{f(i,w)}(t).$$

Then we easily see that L^a, L^f and U operate from $\mathcal{G}$ into itself. We also define operators U^n by $U^1 = U$ and $U^{n+1} = U(U^n)$, $n \ge 1$. Similarly $(L^f)^n$ is defined for $f \in \Delta_M$. In all argument, $F, \hat{F} \in \mathcal{G}$, $F \ge \hat{F}$ means that $F(i,w) \ge \hat{F}(i,w)$ for all $(i,w) \in S_R$.

3. Optimal Value and Optimal Policy

In this section we prove that the optimal value function is a unique solution to an optimality equation and we show that there exists an optimal right continuous stationary policy. These results are composed of an argument similar to that in Ohtsubo [16].

We first give some fundamental lemmas below.

Lemma 3.1. *(i) For F, $\hat{F} \in \mathcal{F}$ and $f \in \Delta_M$, $L^f F - L^f \hat{F} = L^f(F - \hat{F})$.*

(ii) If F, $\hat{F} \in \mathcal{F}$ and $F \geq \hat{F}$, then $L^a F \geq L^a \hat{F}$ for each $a \in A(\cdot)$, $L^f F \geq L^f \hat{F}$ for each $f \in \Delta_M$ and $UF \geq U\hat{F}$.

(iii) If $F \in \mathcal{F}_r$, then $L^a F \in \mathcal{F}_r$ for any $a \in A(\cdot)$. Also, U is an operator from $\mathcal{F}$ (or $\mathcal{F}_r$) into itself.

(iv) If $\hat{F}_n \in \mathcal{F}_r$ and $\hat{F}_n \geq \hat{F}_{n+1}$ for each $n \geq 0$, then $\lim_{n\to\infty} \hat{F}_n \in \mathcal{F}_r$.

Proof. The statements (i), (ii) and the former part of (iii) are immediate results of definitions. Let $F \in \mathcal{F}$ and let $i \in S$ and $w < 0$ be arbitrary. Then $F(i, w) = 0$. Since r is nonnegative function, we have $L^a F(i, w) = 0$ for each $a \in A(i)$ and hence $UF(i, w) = 0$. Next, if $F(i, \cdot)$ is measurable on $\mathbb{R}$, $L^a F(i, \cdot)$ is also measurable for each $a \in A(i)$. and so is $UF(i, \cdot)$, thus $UF \in \mathcal{F}$. Similarly, it is easily proved that if $F \in \mathcal{F}_r$ then $UF \in \mathcal{F}_r$. Hence the latter part of (iii) is proved. Finally we show the statement (iv). It is clear that $\lim_n \hat{F}_n(i, w)$, say $\hat{F}(i, w)$, is nondecreasing in w for each $i \in S$ and that $\hat{F}(i, w) = 0$ for each $i \in S$ when $w < 0$. The right continuity of $\hat{F}(i, \cdot)$ holds by the same way as Lemma 1 in [20]. Thus $\hat{F} \in \mathcal{F}_r$. □

Lemma 3.2. *For each $F \in \mathcal{F}_r$, there exists a right continuous decision rule $F \in \Delta_M$ satisfying $UF = L^f F$.*

Proof. Let $F \in \mathcal{F}_r$ and $(i, w) \in S_R$ be arbitrarily fixed. From Lemma 3.1, $L^a F(i, \cdot)$ is right continuous on $\mathbb{R}$ for each $a \in A(i)$. Since $A(i)$ is finite, we see that there exists $\mu > 0$ and $a \in A(i)$ such that $UF(i, u) = L^a F(i, u)$ for all u satisfying $w \leq u < w + \mu$. For such an action a, if we define $f \in \Delta_M$ by $f(i, u) = a$ for every u so that $w \leq u < w + \mu$, then f is right continuous and $UF(i, w) = L^f F(i, w)$. □

For any $\pi = (f_n, n \geq 0) \in C$ and a given history $(i, w) \in S_R$, the cut-head policy of π to (i, w) is defined by ${}^1\pi^{(i,w)} = (f_n^{(i,w)}, n \geq 0)$ where $f_n^{(i,w)}(h_n) = f_{n+1}((i, w), h_n)$ for every $h_n \in H_n$ and each $n \geq 0$. Then we see that ${}^1\pi^{(i,w)} \in C$

for a fixed (i,w). For the sake of simplicity we use a notation:

$$L^{f_0}F^{1\pi}(i,w) = \sum_{j\in S}\int_{t\in\mathbb{R}^+} F^{1\pi^{(i,w)}}\left(j, w-\int_0^t r(f_0)(i,w,s)ds\right) \times p^{f_0}(j\mid i,w)G^{f_0}(dt\mid i,w,j)$$

for $\pi=(f_n, n\geq 0)\in C$ and $(i,w)\in S_R$.

The following assumption is needed for stationary property of the reward. If reward function r do not depend upon the time, we need not assume the following.

Assumption 3.1. For every $(i,a,t)\in S\times A\times\mathbb{R}_+$ and $n\in\mathbb{N}$,

$$r(i,a,t) = r(i,a,\tau_n+t) \quad \text{a.s.}$$

Lemma 3.3. *Let $\pi=(f_n, n\geq 0)\in C$ be arbitrary.*

(i) For $n\geq 0$, $F_n^\pi \geq F_{n+1}^\pi \geq \lim_{n\to\infty} F_n^\pi = F^\pi$.

(ii) For each $n\geq 0$, $F_n^\pi\in\mathcal{F}$ and $F^\pi\in\mathcal{F}$.

(iii) For each $n\geq 0$, $F_{n+1}^\pi = L^{f_0}F_n^{1\pi}$ and $F^\pi = L^{f_0}F^{1\pi}$. Especially, $F^\pi = L^f F^\pi$ when $\pi=\{f\}^\infty\in C_S$.

Proof. (i) For each initial state $(i,w)\in S_R$ and any $\pi\in C$, it is easy to prove that

$$F_n^\pi(i,w)\geq F_{n+1}^\pi(i,w)\geq \lim_{n\to\infty} F_n^\pi(i,w) = F^\pi(i,w),$$

since $\{V_n\leq w\}\supset\{V_{n+1}\leq w\}$ and $\bigcap_{n=1}^\infty\{V_n\leq w\}=\{V\leq w\}$.

(ii) To show that $F_n^\pi\in\mathcal{F}$, it suffices to prove that $F_n^\pi(i,\cdot)$ is measurable on $\mathbb{R}$ for each $i\in S$. Since $F_0^\pi(i,w)=I_{[0,\infty)}(w)$ where I_A is the indicator function on a set A, we see that $F_0^\pi(i,\cdot)$ is measurable for every $\pi\in C$ and each $i\in S$. We assume that $F_n^{1\pi}(i,\cdot)$ is measurable for each $i\in S$. It then follows from Lemma 3.1 (iii) that for any $\pi=(f_n,n\geq 0)\in C$,

$$L^{f_0}F_n^{1\pi}(i,w) = \sum_{j\in S}\int_{t\in\mathbb{R}^+} F_n^{1\pi^{(i,w)}}\left(j, w-\int_0^t r(f_0)(i,w,s)ds\right) \times p^{f_0}(j\mid i,w)G^{f_0}(dt\mid i,w,j)$$

is well defined and measurable at w. However we have

$$L(f_0)F_0^{1\pi}(i,w) = \sum_{j\in S}\int_{t\in\mathbb{R}^+} I_{[0,\infty]}\left(w-\int_0^t r(f_0)(i,w,s)ds\right) \times p^{f_0}(j\mid i,w)G^{f_0}(dt\mid i,w,j)$$

$$
\begin{aligned}
&= \sum_{j\in S} \int I_{\{u:\int_0^u r(f_0)(i,w,s)ds\le w\}}(t) G^{f_0}(dt \mid i,w,j) p^{f_0}(j \mid i,w) \\
&= \sum_{j\in S} E^{\pi}_{(i,w)}\Big[I_{\{\int_0^{\tau_1} r(f_0)(i,w,s)ds\le w\}} \Big| X_1 = j\Big] p^{f_0}(j \mid i,w) \\
&= \sum_{j\in S} P^{\pi}_{(i,w)}\Big(\int_0^{\tau_1} r(f_0)(i,w,s)ds \le w \Big| X_1 = j\Big) p^{f_0}(j \mid i,w) \\
&= P^{\pi}_{(i,w)}(V_1 \le w) = F^{\pi}_1(i,w),
\end{aligned}
$$

and, by Markov and stationary properties,

$$
\begin{aligned}
L^{f_0} F_n^{{}^1\pi}(i,w) =& \sum_{j\in S} \int_{t\in\mathbb{R}^+} P^{{}^1\pi^{(i,w)}}_{(j,w-y(t))}\left(y(t)+V_n \le w\right) p^{f_0}(j|i,w) G^{f_0}(dt|i,w,j) \\
=& P^{\pi}_i(V_{n+1}\le w) = F^{\pi}_{n+1}(i,w),
\end{aligned}
$$

since ${}^1\pi^{(i,w)} \in C$, where $y(t) = \int_0^t r(f_0)(i,w,s)ds$. Hence, $F^{\pi}_{n+1}(i,\cdot)$ is measurable. Thus, by induction, $F^{\pi}_n(i,\cdot)$ is measurable for every $n \ge 0$. Furthermore, it follows from (i) that $F^{\pi}(i,\cdot) = \lim_{n\to\infty} F^{\pi}_n(i,\cdot)$ is also measurable.

(iii) From the proof of (ii), we have $F^{\pi}_{n+1}(i,w) = L^{f_0} F_n^{{}^1\pi}(i,w)$. Similarly, it is easy to see that $F^{\pi} = L^{f_0} F^{{}^1\pi}$. □

We next give fundamental properties for optimal value functions of finite and infinite horizon cases.

Theorem 3.1. *(i) For each $n \ge 0$, $F^*_n \in \mathcal{F}_r$ and $\{F^*_n, n \ge 0\}$ satisfies equations:*

$$F^*_0 = I_{[0,\infty)}, \quad F^*_n = UF^*_{n-1}, \quad n \ge 1.$$

(ii) For each $n \ge 0$, there exists a right continuous policy $\pi \in C_M$ such that $F^*_n = F^{\pi}_n$.

(iii) For each $n \ge 0$, $F^*_n \ge F^*_{n+1} \ge \lim_{n\to\infty} F^*_n = F^*$ and $F^* \in \mathcal{F}_r$.

Proof. We first prove the statements (i) and (ii) by induction. When $n = 0$, we see that $F^*_0(i,w) = I_{[0,\infty)}(i,w) = F^{\pi}_0(i,w)$ for any right continuous policy $\pi \in C_M$ and every S_R and $F^*_0 \in \mathcal{F}_r$, which implies that (i) and (ii) hold for $n = 0$. We assume that these statements are true for $n = k$. Then, $F^*_k \in \mathcal{F}_r$ and there exists a right continuous policy $\sigma \in C_M$ such that $F^*_k = F^{\sigma}_k$. It follows from Lemma 3.2 that there exists a right continuous decision rule $\hat{f} \in \Delta_M$ such that $UF^*_k = L^{\hat{f}} F^*_k$, which implies that $\rho = (\hat{f}, \sigma)$ is a right continuous policy in C_M. It follows from Lemma 3.3(iii) that for each (i,w),

$$F^*_{k+1}(i,w) \le F^{\rho}_{k+1}(i,w) = L^{\hat{f}} F^{\sigma}_k(i,w) = L^{\hat{f}} F^*_k(i,w) = UF^*_k(i,w).$$

Conversely, we see from Lemma 3.3(iii) again that for any $\pi = (f_n, n \geq 1) \in C$,

$$F^{\pi}_{k+1}(i,w) = L^{f_1} F^{{}^1\pi}_k(i,w) \geq L^{f_1} F^*_k(i,w) \geq U F^*_k(i,w).$$

Taking infimum over C, we obtain $F^*_{k+1}(i,w) \geq U F^*_k(i,w)$. Thus, combining with the previous inequality, we have $U F^*_k(i,w) = F^*_{k+1}(i,w) = F^{\rho}_{k+1}(i,w)$. Hence, ρ satisfies $F^*_{k+1} = F^{\rho}_{k+1}$ and from Lemma 3.1(iii), we have $F^*_{k+1} = U F^*_k \in \mathcal{F}_r$. By induction the proof of statement (i) and (ii) is complete.

(iii) It follows from Lemma 3.3(i) that for any $\pi \in C$, $F^{\pi}_n \geq F^*_n \geq F^*_{n+1} \geq F^*$, which implies that $\lim_{n\to\infty} F^*_n \geq F^*$ and $F^{\pi} = \lim_{n\to\infty} F^{\pi}_n \geq \lim_{n\to\infty} F^*_n$, so $F^* \geq \lim_{n\to\infty} F^*_n$ by taking infimum over C. Thus we have $F^* = \lim_{n\to\infty} F^*_n$. Also, since $F^*_n \in \mathcal{F}_r$, $n \geq 0$, and $F^*_n \geq F^*_{n+1}$, it follows from Lemma 3.1(iv) that $F^* \in \mathcal{F}_r$. □

From Theorem 3.1, we have $F^* = \lim_{n\to\infty} U^n F^*_0$. In order to characterize the optimal value F^*, we need the following important lemma, in which the statement (iii) is a version of the policy improvement given in Howard [11].

Lemma 3.4. *Let $\pi = \{f\}^{\infty} \in C_M$ be arbitrary.*

(i) Let $F, \hat{F} \in \mathcal{F}$. If $F - \hat{F} \leq L^f(F - \hat{F})$ on $B^c \times \mathbb{R}$ and $F = \hat{F}$ on $B \times \mathbb{R}$, then $F \leq \hat{F}$.

(ii) F^{π} is a unique solution in $\mathcal{F}$ to equation $F = L^f F$ with $F = I_{[0,\infty)}$ on $B \times \mathbb{R}$.

(iii) *Let $\sigma \in C_M$. If $F^{(f,\sigma)} \leq F^{\sigma}$, then $F^{\pi} \leq F^{\sigma}$.*

Proof. (i) Since $F = \hat{F}$ on $B \times \mathbb{R}$ and the set B is a recurrent class and is reward-free, it follows that if $i \in B$ then

$$L^f(F - \hat{F})(i,w) = \sum_{j\in B} \int_{t\in\mathbb{R}^+} (F - \hat{F})(j,w) p^f(j \mid i,w) G^f(dt|i,w,j) = 0$$

for every $w \in \mathbb{R}$. By the fact that $F - \hat{F} \leq 1$, we also see that if $(i,w) \in B^c \times \mathbb{R}$, then

$$L^f(F - \hat{F})(i,w) = \sum_{j\in B^c} \int_{t\in\mathbb{R}^+} (F - \hat{F}) \left(j, w - \int_0^t r(f)(i,w,s) ds \right)$$
$$\times p^f(j \mid i,w) G^f(dt \mid i,w,j) \leq \sum_{j\in B^c} \int_{t\in\mathbb{R}^+} p^f(j \mid i,w) G^f(dt \mid i,w,j)$$
$$= P^{\pi}_{(i,w)}(X_1 \in B^c).$$

For $n \geq 1$, assume that $(L^f)^n(F - \hat{F})(i,w) = 0$ for any $(i,w) \in B \times \mathbb{R}$ and $(L^f)^n(F - \hat{F})(i,w) \leq P^{\pi}_{(i,w)}(\bigcap_{k=1}^n \{X_k \in B^c\})$ for any $(i,w) \in B^c \times \mathbb{R}$. Then, it

follows that when $(i,w) \in B \times \mathbb{R}$

$$(L^f)^{n+1}(F-\hat{F})(i,w) = L^f(L^f)^n(F-\hat{F})(i,w)$$
$$= \sum_{j\in B}\int_{t\in\mathbb{R}^+}(L^f)^n(F-\hat{F})(j,w)p^f(j \mid i,w)G^f(dt \mid i,w,j) = 0,$$

and from Markov and stationary property that when $(i,w) \in B^c \times \mathbb{R}$

$$(L^f)^{n+1}(F-\hat{F})(i,w) = L^f(L^f)^n(F-\hat{F})(i,w)$$
$$= \sum_{j\in B^c}\int_{t\in\mathbb{R}^+}(L^f)^n(F-\hat{F})\left(j, w-\int_0^t r(f)(i,w,s)ds\right)p^f(j \mid i,w)G^f(dt \mid i,w,j)$$
$$\le \sum_{j\in B^c}\int_{t\in\mathbb{R}^+}P^\pi_{\left(j,w-\int_0^t r(f)(i,w,s)ds\right)}\left(\bigcap_{k=1}^n\{X_k\in B^c\}\right)$$
$$\times p^f(j \mid i,w)G^f(dt \mid i,w,j) = P^\pi_{(i,w)}\left(\bigcap_{k=1}^{n+1}\{X_k\in B^c\}\right).$$

By induction, we have

$$(F-\hat{F})(i,w) \le (L^f)^n(F-\hat{F})(i,w) \le P^\pi_{(i,w)}\left(\bigcap_{k=1}^n\{X_k\in B^c\}\right),$$

for every $(i,w)\in B^c\times\mathbb{R}$ and $n\ge 1$. Since $P^\pi_{(i,w)}(X_n \in B \text{ for some } n\ge 1) = 1$ from Assumption 2.3, we obtain

$$\lim_{n\to\infty} P^\pi_{(i,w)}\left(\bigcap_{k=1}^n\{X_k\in B^c\}\right) = 1 - P^\pi_{(i,w)}\left(\bigcup_{k=1}^\infty\{X_k\in B^c\}\right) = 0.$$

Letting $n\to\infty$ on the above inequality, we have $(F-\hat{F})(i,w)\le 0$ for every $(i,w)\in B^c\times\mathbb{R}$, which completes the proof of the statement (i).

(ii) Let $F\in\mathcal{F}$ be a solution to $F = L^fF$ with $F = I_{[0,\infty)}$ on $B\times\mathbb{R}$. By Lemma 3.3(iii), F^π satisfies $F^\pi = L^fF^\pi$ and $F^\pi = I_{[0,\infty)}$ on $B\times\mathbb{R}$. Hence $F-F^\pi = L^f(F-F^\pi)$ on $B^c\times\mathbb{R}$ and $F = F^\pi$ on $B\times\mathbb{R}$. Thus the statement (i) implies $F = F^\pi$.

(iii) Let $F^{(f,\sigma)} \le F^\sigma$ for any $\sigma\in C_M$. It follows from Lemma 3.3(ii) and the statement (ii) that $F^\pi - F^\sigma \le F^\pi - F^{(f,\sigma)} = L^f(F^\pi - F^\sigma)$ and $F^\pi = F^\sigma$ on $B\times\mathbb{R}$. Thus the statement (i) implies $F^\pi \le F^\sigma$. □

Now we are in a position to give a main theorem.

Theorem 3.2. *(i) F^* is a unique solution in $\mathcal{F}$ to equation $F = UF$ with $F = I_{[0,\infty)}$ on $B\times\mathbb{R}$.*

(ii) There exists a right continuous policy $\pi = \{f\}^\infty \in C_M$ *satisfying* $F^* = L^f F^*$ *on* $B^c \times \mathbb{R}$ *and* π *is optimal.*

Proof. It follows from Lemma 3.3(iii) that $F^\pi = L^{f_1} F^{1\pi} \geq L^{f_1} F^* \geq UF^*$ for any $\pi = (f_n, n \geq 0) \in C$. Hence we have $F^* \geq UF^*$. Conversely, for $(i, w) \in S_R$, it follows from Theorem 3.1(i) that $F_n^*(i, w) = UF_{n-1}^*(i, w) \leq L^a F_{n-1}^*(i, w)$ for any $a \in A(i)$. By Theorem 3.1(iii) and the dominated convergence theorem, we have $F^*(i, w) \leq L^a F^*(i, w)$ for every $a \in A(i)$, so $F^*(i, w) \leq \min_{a \in A(i)} L^a F^*(i, w) = UF^*(i, w)$. Therefore F^* satisfies the equation $F^* = UF^*$.

We have $F^* \in \mathcal{F}_r$ from Theorem 3.1(iii). Thus Lemma 3.2 leads that there exists a rightcontinuous decision rule $f \in \Delta_M$ such that $F^* = L^f F^*$.

Let $F \in \mathcal{F}$ be another solution to equation $F = UF$ with $F = F^* = I_{[0,\infty)}$ on $B \times \mathbb{R}$. Since $A(i)$ is finite, we can define $\hat{f} \in \Delta_M$ by $UF(i, w) = L^{\hat{f}} F(i, w)$ for all $w \in \mathbb{R}$. Thus we have $F^* = L^f F^* \leq L^{\hat{f}} F^*$ and $F = L^{\hat{f}} F \leq L^f F$ on $B^c \times \mathbb{R}$. Hence we have $F - F^* \leq L^f (F - F^*)$ and $F^* - F \leq L^{\hat{f}} (F^* - F)$ on $B^c \times \mathbb{R}$. Lemma 3.4(i) implies that $F = F^*$, which completes the uniqueness of F^*.

Let $\pi = \{f\}^\infty$ for a right continuous decision rule $f \in \Delta_M$ such that $F^* = L^f F^*$. It follows from Lemma 3.4(ii) that F^π is a unique solution in $\mathcal{F}$ to equation $F = L^f F$ with $F = I_{[0,\infty)}$ on $B \times \mathbb{R}$. Thus we have $F^* = F^\pi$ and hence it follows that π is optimal. □

4. Value and Policy Iteration Methods

In this section we consider several value iterations and a policy space iteration.

We see from Theorem 3.1 that a value iteration is given by $F^* = \lim_n U^n F_0^*$ where $F_0^*(i, w) = I_{[0,\infty)}(w)$ for each $(i, w) \in S_R$. We give other value iteration.

Theorem 4.1. *Let* $K \in \mathcal{F}$ *be a function satisfying* $K \geq F^*$. *Then* $\{U^n K\}$ *converges and* $\lim_{n\to\infty} U^n K = F^*$.

Proof. Since $K \in \mathcal{F}$, We have $F_0^* \geq K$. Hence $U^n F_0^* \geq U^n K^*$, which leads the inequality $F^* = \lim_{n\to\infty} F_0^* \geq \limsup_n U^n K$. Conversely, since $K \geq F^*$ and $F^* = UF^*$, we have $U^n K \geq U^n F^* = F^*$, and hence $\liminf_n U^n K \geq F^*$. Therefore, combining with the previous inequality, we have $\lim_{n\to\infty} U^n K = F^*$. □

Corollary 4.1. *For any policy* $\pi \in C$, $\lim_{n\to\infty} U^n F^\pi = F^*$.

Proof. The corollary is an immediate result of Theorem 4.1, since $F^{\pi} \geq F^*$ and $F^{\pi} \in \mathcal{F}$. □

Next we consider a policy space iteration. White in [20] proposes a policy space iteration for discounted Markov decision processes, and Ohtsubo in [16] envelopes it for undiscounted Markov decision processes.

The policy space procedure in our model is as follows:

(i) Select an initial policy $\pi_0 = \{f_0\}^{\infty} \in C_S$.

(ii) At step n, assume that we have a policy $\pi_n = \{f_n\}^{\infty} \in C_S$ and solve the equation $F = L^{f_n}F$ with $F = I_{[0,\infty)}$ on $B \times \mathbb{R}$ to give a function $F^{\pi_n} \in \mathcal{F}$.

(iii) If $L^{f_n}F^{\pi_n} = UF^{\pi_n}$, stop the procedure. If $L^{f_n}F^{\pi_n} \neq UF^{\pi}$, go the next step.

(iv) Find a new policy $\pi_{n+1} = \{f_{n+1}\}^{\infty} \in C_S$ by $L^{f_{n+1}}F^{\pi_n} = UF^{\pi_n}$.

(v) Return to step (ii) replacing n by $n+1$.

We can, from Lemma 3.4(ii), uniquely solve the equations in $\mathcal{F}$ at step (ii). We have the following convergence theorem.

Theorem 4.2. *(i) The sequence $\{F^{\pi_n}\}_{n\geq 0}$ is nonincreasing and converges F^*.*

(ii) If $L^{f_n}F^{\pi_n} = UF^{\pi_n}$, then F^{π_n} is the optimal value and $\pi_n = \{f_n\}^{\infty} \in C_S$ is an optimal policy.

Proof. (i) Since $F^{\pi_n} = L^{f_n}F^{\pi_n}$ for each $n \geq 0$ by Lemma 3.4(ii), we have $F^{(f_{n+1},\pi_n)} = L^{f_{n+1}}F^{\pi_n} = UF^{\pi_n} \leq L^{f_n}F^{\pi_n} = F^{\pi_n}$. Thus Lemma 3.4(iii) implies that $F^{\pi_n} \geq F^{\pi_{n+1}}$. Hence the sequence $\{F^{\pi_n}\}$ is nonincreasing. Thus $\{F^{\pi_n}\}_{n\geq 0}$ tends to a function $\tilde{F} \in \mathcal{F}$. We show that $F^* = \tilde{F}$. Since $F^* \leq F^{\pi_n}$ for all $n \geq 0$ we have $F^* \leq \tilde{F}$. Next is follows that for each $n \geq 2$,

$$\tilde{F} \leq F^{\pi_n} = L^{f_n}F^{\pi_n} \leq L^{f_n}F^{\pi_{n-1}} = UF^{\pi_{n-1}}.$$

Also we similarly have $F^{\pi_{n-1}} \leq UF^{\pi_{n-2}}$. Thus we obtain $\tilde{F} \leq U^2F^{\pi_{n-2}}$ and hence $\tilde{F} \leq U^nF^{\pi_0}$ for all $n \geq 0$ by induction. From Corollary 4.1, it follows that

$$\tilde{F} \leq \lim_{n\to\infty} U^nF^{\pi_0} = F^*.$$

Therefore we have $\tilde{F} = F^*$.

(ii) Let $L^{f_n}F^{\pi_n} = UF^{\pi_n}$. Then, since $F^{\pi_n} = L^{f_n}F^{\pi_n} = UF^{\pi_n} = L^{f_{n+1}}F^{\pi_n}$ and $F^{\pi_{n+1}} = L^{f_{n+1}}F^{\pi_{n+1}}$, we have $F^{\pi_n} - F^{\pi_{n+1}} = L^{f_{n+1}}(F^{\pi_n} - F^{\pi_{n+1}})$. Thus Lemma 3.4(ii) implies that $F^{\pi_{n+1}} = F^{\pi_n}$. Hence we have $L^{f_{n+1}}F^{\pi_{n+1}} = L^{f_{n+1}}F^{\pi_n} = UF^{\pi_n} = UF^{\pi_{n+1}}$. From the same discussion as the above we

have $F^{\pi_{n+2}} = F^{\pi_{n+1}}$. By induction, we have $F^{\pi_n} = F^{\pi_k}$ for all $k \geq n$. Letting $k \to \infty$ and using the statement (i), we obtain $F^{\pi_n} = F^*$ and hence π_n is optimal. □

5. Numerical Examples

In this section, we give several numerical examples. We first give a policy $\pi \in C$ such that $F^\pi \notin \mathcal{F}_r$. In the second example, we notice that, when Assumption 2.3 does not hold, the optimal value is not necessarily a probability distribution function. We finally give an example in which we get the optimal value by a policy space method. We give some notations for these examples. For every $i \in S$, when $G^a_{ij}(t)$ is the same distribution function as for all $j \in S$ and $a \in A$, we simply note by $G_i(t)$ the function. Let G_1 and G_2 be the distribution function. We denote by $G_1 * G_2$ the convolution operation for G_1 and G_2, that is, $G_1 * G_2(t) = \int_{\mathbb{R}} G_1(t-y)G_2(dy)$. Thus we define $G_1^{(1)*} = G_1$ and denote by $G_1^{(n)*}$ the n-fold convolution of G_1 with itself, that is, $G_1^{(n+1)*} = G_1^{(n)*} * G_1$. For $\lambda > 0$ and $\mu \in \mathbb{R}$, exponential distribution function with parameter (λ, μ) is given by $G(t) = 1 - \exp(-\lambda(t-\mu))$ if $t \geq \mu$ and 0 otherwise, let us use the shorthand $\mathcal{E}(\lambda, \mu)$ for it.

Example 5.1. Let $S = \{1, 2\}$ be a state space and $\{2\}$ be a target set. We assume that the state 2 is recurrent (absorbing) and reward-free. Also let $A = \{a_1, a_2\}$ be an action space. Let transition probabilities be $p^{a_1}(2|1) = 1$ and $p^{a_2}(1|1) = p^{a_2}(2|1) = 1/2$. Let distributions of the sojourn time be $G_1(t)$ and $G_2(t)$ as $\mathcal{E}(1, 1)$ and $\mathcal{E}(2, 0)$, respectively. Let a reward function be $r(1, a_1, s) = 1/2$, $r(1, a_2, s) = 1$ and $r(2, a_1, s) = r(2, a_2, s) = 0$ for every $s \in \mathbb{R}^+$.

We define a policy $\pi = \{f\}^\infty \in C_S$ by $f(1, w) = a_1$ if $w \leq 1$ and $f(1, w) = a_2$ otherwise. Then if $w \leq 1$ then $f(1, w) = a_1$, and have

$$F^\pi_{n+1}(1, w) = \int_{t\in\mathbb{R}} F^\pi_n\Big(2, w - \frac{t}{2}\Big)G_1(dt) = \int_{t\in\mathbb{R}} I_{(-\infty, 2w]}(t)G_1(dt)$$
$$= \begin{cases} 1 - e^{-2\left(w-\frac{1}{2}\right)} & \text{if } w \geq \frac{1}{2}, \\ 0 & \text{if } w < \frac{1}{2}, \end{cases}$$

for all $n \geq 0$.

If $2 > w > 1$ then $f(1, w) = a_2$. When $t > 1$, we see that $w - \int_0^t r(1, a_2, s)ds = w - t \leq 1$. Hence we have

$$F_n^\pi\left(1, w - \int_0^t r(1, a_2, s)ds\right) = F_n^\pi(1, w-t)$$
$$= \begin{cases} 1 - e^{-2\left(w-t-\frac{1}{2}\right)} & \text{if } w-t \geq \frac{1}{2}, \\ 0 & \text{if } w-t < \frac{1}{2}, \end{cases}$$

for all $n \geq 1$. Therefore we easily see that

$$F_{n+1}^\pi(1, w) = \begin{cases} \frac{1}{2} - \frac{3}{2}e^{-w+1} - \frac{1}{2}e^{-w+2} + \frac{1}{2}e^{-2w+\frac{7}{2}} & \text{if } w \geq \frac{3}{2}, \\ \frac{1}{2}(1 - e^{-(w-1)}) & \text{if } w < \frac{3}{2}, \end{cases}$$

for all $n \geq 1$. Thus we have $\lim_{w \downarrow 1} F^\pi(1, w) = 0$, $F^\pi(1, 1) = 1 - e^{-1}$. Therefore we see that F^π is not nondecreasing and not right continuous, so $F^\pi \notin \mathcal{F}_r$.

In the following system, we can take a policy π such that $P_{(i,w)}^\pi(\zeta = \infty) = 1$, that is, $P_{(i,w)}^\pi(Z(t) \in B^c \text{ for all } t \in \mathbb{R}^+) = 1$, against Assumption 2.3. Then we can easily see that $P_{(i,w)}^\pi(V = \infty) = 1$, if there exists a positive constant M such that $\inf\{r(i, a, s) \mid i \in B^c, a \in A, s \in \mathbb{R}^+\} \geq M$. Thus It is clear that the system obtains $F^* = 0$ on $B^c \times \mathbb{R}$. In the case $\{F_n^*\}$ tends to 0.

Example 5.2. Let $S = \{1, 2, 3\}$ be a state space and $\{3\}$ be a target set. We assume that the state 3 is recurrent (absorbing) and reward-free. Also let $A = \{a_1, a_2\}$ be an action space. Let transition probability be $p^{a_1}(2|1) = p^{a_1}(3|2) = p^{a_2}(1|2) = 1$, $p^{a_2}(2|1) = 2/3$ and $p^{a_2}(3|1) = 1/3$. Let the distributions of the sojourn time be $G_1(t)$ as $\mathcal{E}(1, 0)$, and $G_2(t) = G_1 * G_1(t)$. Let a reward function be $r(i, a, s) = 1$ for every $i \in S$, all $a \in A$ and any $s \in \mathbb{R}^+$. By induction, we have

$$F_n^*(2, w) = \begin{cases} G_2 * (G_1 * G_2)^{(2m-1)*}(w) & \text{if } n = 2m+1, \\ (G_1 * G_2)^{(m)*}(w) & \text{if } n = 2m, \end{cases}$$
$$F_n^*(1, w) = \begin{cases} G_1 * (G_1 * G_2)^{(2m-1)*}(w) & \text{if } n = 2m+1, \\ F_n^*(2, w) & \text{if } n = 2m, \end{cases}$$

for $m \geq 1$. Since $\lim_{m\to\infty}(G_1 * G_2)^{(m)*}(w) = 0$, $F^*(i, w) = 0$ for every $(i, w) \in B^c \times \mathbb{R}$.

Example 5.3. Let $S = \{1, 2, 3\}$ be a state space and $\{3\}$ be a target set. We assume that the state 3 is recurrent (absorbing) and reward-free. Also let $A = \{a_1, a_2\}$ be an action space. Let transition probabilities be $p^{a_1}(2|1) = p^{a_1}(3|2) = 1$, $p^{a_2}(2|2) = p^{a_2}(3|2) = 1/2$, $p^{a_2}(2|1) = 1/3$ and $p^{a_2}(3|1) = 2/3$. Let distributions of the sojourn time be $G_1(t)$ and $G_2(t)$ as $\mathcal{E}(1, 0)$ and $\mathcal{E}(1, 2)$, respectively. We denote by $\tilde{G}_2(t)$ as $\mathcal{E}(1/2, 4)$. Let a reward function be $r(1, a_1, s) = r(1, a_2, s) = r(2, a_1, s) = 1$ and $r(2, a_2, s) = 2$ for any

$s \in \mathbb{R}^+$.

We consider a policy space procedure to give an optimal policy. Let $\pi_0 = \{f_0\}^\infty$ be an initial policy such that $f_0(i,w) = a_1$ for every $(i,w) \in S_R$. Solving the equation $F = L^{f_0}F$ with $F(3,w) = I_{[0,\infty)}(w)$ for every $w \in \mathbb{R}$, we have

$$F^{\pi_0}(2,w) = \tilde{G}_2(w), \quad F^{\pi_0}(1,w) = G_1 * \tilde{G}_2(w).$$

We now see that $L^{f_0}F^{\pi_0} \neq UF^{\pi_0}$, since

$$\begin{aligned} UF^{\pi_0}(2,w) &= \tilde{G}_2(w)I_{(-\infty,w^*)}(w) \\ &+ \left(\frac{1}{2}G_2 * \tilde{G}_2(w) + \frac{1}{2}G_2(w)\right)I_{[w^*,\infty)}(w), \\ UF^{\pi_0}(1,w) &= G_1 * \tilde{G}_2(w), \end{aligned}$$

where $4 < w^* < 6$ which is the unique solution to $L^{f_0}F^{\pi_0}(2,w) = UF^{\pi_0}(2,w)$. Using $L^{f_1}F^{\pi_0} = UF^{\pi_0}$, we give a policy $\pi_1 = \{f_1\}^\infty \in C_M$ by

$$\begin{cases} f_1(3,w) = a_1, \\ f_1(2,w) = a_1 I_{(-\infty,w^*)}(w) + a_2 I_{[w^*,\infty)}(w), \\ f_1(1,w) = a_1. \end{cases}$$

By solving $F = L^{f_1}F$, F^{π_1} is given by

$$F^{\pi_1}(2,w) = \begin{cases} \tilde{G}_2(w) & \text{if } w < w^*, \\ D^{(n)}(w) & \text{if } w^* + 2n \le w < w^* + 2(n+1), \end{cases}$$

where, for $n \ge 0$,

$$\begin{aligned} D^{(n+1)}(w) &= \frac{1}{2}\Bigg\{ \int_{w-w^*}^{w-4} \tilde{G}_2(w-t)G_2(dt) \\ &+ \sum_{k=0}^{n} \int_{-w-w^*-2(k+1)}^{w-w^*-2k} D^{(k)}(w-t)G_2(dt) \Bigg\} + \frac{1}{2}G_2(w), \end{aligned}$$

and

$$D^{(0)}(w) = \frac{1}{2}G_2 * \tilde{G}_2(w) + \frac{1}{2}G_2(w).$$

Then it follows that $F^{\pi_1}(1,w) = F^{\pi_0}(1,w)$ and $F^{\pi_1}(2,w) = UF^{\pi_1}(2,w)$. Thus we have $L^{f_1}F^{\pi_1} = UF^{\pi_1}$, and hence we stop the procedure. From Theorem 4.2(ii) we obtain the optimal value F^{π_1} and an optimal policy $\pi_1 = \{f_1\}^\infty$.

References

[1] R.N. Bhattacharya, M. Majumdar, Controlled semi-Markov models – The discouted case, *J. Statist. Plan. Infer.*, **21** (1989), 365-381.

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

[3] D. Blackwell, Positive dynamic programming, In: *Proc. 5-th Berkeley Symp. on Math. Statist. Prob.*, Volume 1, University of California Press, Berkeley (1967), 415-418.

[4] C. Derman, On sequential decisions and Markov chains, *Manage. Sci.*, **9** (1962/63), 16-24.

[5] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York (1970).

[6] J.H. Eaton, L.A. Zadeh, Optimal pursuit strategies in discrete-state probabilistic systems, *Trans. ASME Ser. D, J. Basic Eng.*, **84** (1962), 23-29.

[7] A. Federgruen, A. Hordijk, H.C. Tijms, Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion, *Stochastic Process. Appl.*, **9** (1979), 223-235.

[8] A. Federgruen, P.J. Schweitzer, H.C. Tijms, Denumerable undiscounted semi-Markov decision processes with unbounded rewards, *Math. Oper. Res.*, **8** (1983), 298-313.

[9] E.A. Feinberg, Constrained discouted semi-Markov decision processes, In: *Markov Processes and Contolled Markov Chains* (Ed-s: Z. Hou et al), Kluwer Academic Publishers, Netherlands (2002), 233-244.

[10] O. Hernández-Lerma, J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes,* Springer, New York (1999).

[11] R.A. Howard, *Dynamic Programming and Markov Processes*, The M.I.T. Press, Massachusetts, (1960).

[12] R.A. Howard, Research in semi-Markovian decision structures, *J. Opns. Res. Soc. Japan*, **6** (1964), 163-199.

[13] Y. Ohtsubo, K. Toyonaga, Optimal policy for minimizing risk models in Markov decision processes, *J. Math. Anal. Appl.*, **271** (2002), 66-81.

[14] Y. Ohtsubo, K. Toyonaga, Equivalence classes for minimizing risk models in Markov decision processes, *Math. Method. Oper. Res.*, **60** (2004), 239-250.

[15] Y. Ohtsubo, Minimizing risk models in stochastic shortest path problems, *Math. Meth. Oper. Res.,* **57** (2003), 79-88.

[16] Y. Ohtsubo, Optimal threshold probability in undiscounted Markov decision processes with a target set, *Appl. Math. Comput.*, **149** (2004), 519-532.

[17] S.M. Ross, Average cost semi-Markov decision processes, *J. Appl. Prob.*, **7** (1970), 649-656.

[18] A.F. Veinott, Jr., Discrete dynamic programming with sensitive discount optimality criteria, *Ann. Math. Statist.*, **40** (1969), 1635-1660.

[19] D.J. White, *Markov Decision Processes*, Wiley, New York, (1993).

[20] D.J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.*, **173** (1993), 634-646.

[21] 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.

[22] M. Yasuda, Semi-Markov decision processes with countable state space and compact action space, *Bull. Math. Statist.*, **18** (1978/79), 35-54.