Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Optimal risk probability for first passage models in semi-Markov decision processes [☆]

Yonghui Huang, Xianping Guo [*]

School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 18 June 2008
Available online 2 June 2009
Submitted by C.E. Wayne

Keywords:
Semi-Markov decision processes
Target set
First passage time
Risk probability
Optimal policy

ABSTRACT

This paper studies the risk minimization problem in semi-Markov decision processes with denumerable states. The criterion to be optimized is the risk probability (or risk function) that a first passage time to some target set doesn't exceed a threshold value. We first characterize such risk functions and the corresponding optimal value function, and prove that the optimal value function satisfies the optimality equation by using a successive approximation technique. Then, we present some properties of optimal policies, and further give conditions for the existence of optimal policies. In addition, a value iteration algorithm and a policy improvement method for obtaining respectively the optimal value function and optimal policies are developed. Finally, two examples are given to illustrate the value iteration procedure and essential characterization of the risk function.

1. Introduction

In the field of Markov decision processes (MDPs) researchers have considered the optimization problem $\sup_\pi P_i^\pi(\tau_B > \lambda)$, where i is an initial state, π is a policy, τ_B is a first passage time to a given target set B, and λ is a threshold value. Such optimization problems arise from the background of reliability engineering and risk analysis, in which the target set B usually corresponds to the set of failure states of a system, and the probability $P_i^\pi(\tau_B > \lambda)$ assesses the reliability of the system that the working life would be more than λ time units. According to the time parameter, the existing works on this optimization problem can be roughly classed into two cases: the discrete-time MDPs (DTMDPs) and the continuous-time MDPs (CTMDPs). For the DTMDPs, Liu and Huang [12] establish the optimality equation and present some properties of several kinds of optimal policies. They also show existence results and algorithms for these optimal policies. For the CTMDPs, Lin, Tomkins and Wang [10] consider the equivalent risk minimization problem $\inf_\pi P_i^\pi(\tau_B \leqslant \lambda)$, and give some necessary and sufficient conditions for the existence of optimal policies.

In this paper, we devote ourselves to the risk minimization problem $\inf_\pi P_i^\pi(\tau_B \leqslant \lambda)$ in continuous-time semi-Markov decision processes (SMDPs) with denumerable states. As is known, since in the SMDPs the time between decision epochs may follow arbitrary distributions, the SMDPs [3,7,8] are a generalization of both CTMDPs [4,5,10] and DTMDPs [2,6,12]. Therefore, the model in this paper is more general than those in [10,12]; see Remark 3.5 and Example 6.2 for details.

As mentioned above, the risk minimization problem for the continuous-time case has been considered in [10]. However, in [10] both the state and action spaces are assumed to be *finite*, the *additional* conditions for the existence of optimal policies, which require that (finite or infinite) countably many intersections of certain policy sets are nonempty, are not easy to verify, and moreover, the optimality equation has not been established yet. In addition, it is worth noting that the

[☆] This work was supported by NSFC (60874004, 60736028).

[*] Corresponding author.

E-mail addresses: hyhui@mail2.sysu.edu.cn (Y. Huang), mcsgxp@mail.sysu.edu.cn (X. Guo).

0022-247X/$ – see front matter Crown Copyright © 2009 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jmaa.2009.05.058

analytic methods for the discrete-time models in [12] may be invalid for the continuous-time ones. For instance, we can not use the existence of ε-optimal policies to establish the optimality equation for the continuous-time models because these ε-optimal policies may not exist in continuous-time models. All of these, which largely motivate this paper, may be due to the fact that each policy in [10] is independent of threshold values λ.

In this paper, however, our analytic method used is different from those in [10,12]. In fact, our method is similar to those in Ohtsubo and Toyonaga [13], Ohtsubo [14], and Wu and Lin [17] for another minimizing risk problem $\inf_\pi P_i^\pi(V \leqslant \lambda)$ in DTMDPs, where V denotes a total discounted reward over an infinite horizon. In the spirit of the analytic technique in [13,14,17], we introduce the class of policies which include not only the usual states and actions but also threshold values λ for SMDPs, and then establish the optimality equation (Theorem 3.2) and the existence of optimal policies (Theorem 4.2) under suitable conditions much weaker than those in [10].

More precisely, using a successive approximation technique, we prove that the optimal value function is nondecreasing and right continuous of threshold values and satisfies the optimality equation, and thus establish the optimality equation. Then, we further show that there exists an optimal stationary policy and any stationary policy realizing the minimum in the optimality equation is optimal under some suitable conditions (see Assumptions 2.1 and 4.1). We also give sufficient and necessary conditions for existence of optimal policies which are independent of threshold values (see Theorem 4.3). Besides, our works still include: (1) some properties of optimal policies (see Theorem 4.1), (2) a value iteration algorithm and a policy improvement method for computing respectively the optimal value function and optimal policies, and (3) two examples used to illustrate the value iteration procedure and some essential characterization for our model. These results are new for continuous-time models, which include SMDPs and CTMDPs.

The rest of this paper is as follows. In Section 2, we formulate our control model. In Section 3, basic properties of risk functions and optimal value function are studied, and the optimality equation is established. In Section 4, some properties of optimal policies and conditions for the existence of two kinds of optimal policies are discussed. After giving a policy improvement method as well as a value iteration algorithm in Section 5, we illustrate the value iteration algorithm by two examples in Section 6.

2. The control model

In this section we introduce the model of SMDPs

$$\bigl\{S, B, A, \bigl(A(i),\ i \in S\bigr), Q(t, j \mid i, a)\bigr\}, \tag{1}$$

where S is a state space and A is an action set, which are assumed to be *denumerable*, respectively; $B \subset S$ is a given target set; $A(i) \subset A$ is a set of admissible actions at state $i \in S$, which is assume to be *finite*. Let $K := \{(i, a) \mid i \in S,\ a \in A(i)\}$ be the set of feasible state-action pairs, $B^c := S - B$, $R_+ := [0, +\infty)$ and $R := (-\infty, +\infty)$. The function $Q(t, j \mid i, a)$ in (1) is the semi-Markov decision kernel which satisfies:

1. $Q(\cdot, j \mid i, a)$, for any fixed $j \in S$, $(i, a) \in K$, is a nondecreasing, right continuous real function on R_+ such that $Q(0, j \mid i, a) = 0$;
2. $Q(t, \cdot \mid \cdot, \cdot)$, for every $t \in R_+$, is a sub-stochastic kernel on S given K such that $D(t \mid i, a) := \sum_{j \in S} Q(t, j \mid i, a) \leqslant 1$ for all $(i, a) \in K$;
3. $P(\cdot \mid \cdot, \cdot) := Q(\infty, \cdot \mid \cdot, \cdot)$ is a stochastic kernel on S given K such that $\sum_{j \in S} P(j \mid i, a) = 1$ for all $(i, a) \in K$.

Remark 2.1. Note that our model here is slightly different from the usual SMDPs because we have not considered a reward or cost structure; see Lippman [11], Puterman [15] and Ross [16].

In the SMDPs, if the system occupies state $i \in S$ at some decision epoch and the decision-maker chooses an action $a \in A(i)$, the following consequence occurs (regardless of the previous history of the system): the system state stays at i within t units of time, then changes to some $j \in S$ with probability $Q(t, j \mid i, a)$, and the next decision epoch follows.

We now describe the evolution of a SMDP and how a decision-maker chooses his or her actions. At time t_0, which is the initial decision epoch, the system occupies state i_0, and the decision-maker has a goal (threshold value) λ_0 in mind, that is, he should try to avoid the risk that the system state falls in the target set B within λ_0 time units. Therefore, the decision-maker chooses an action a_0 according to the current state i_0 and his goal λ_0. As a consequence of this action choice, the system remains in i_0 until time t_1, at which point the system state changes to i_1 and the next decision epoch occurs. At time t_1, the decision-maker has a new goal (threshold value) $\lambda_1 := \lambda_0 - (t_1 - t_0)$, that is, he now should try to avoid the risk that the system state falls in the target set B within λ_1 time units. According to the current state i_1 and the new goal λ_1 as well as the previous state i_0 and the previous goal λ_0, the decision-maker chooses an action a_1 and the same sequence of events occur. The decision process evolves in this way and thus we obtain an admissible history h_n of the SMDP up to the nth decision epoch, i.e.,

$$h_n = (t_0, i_0, \lambda_0, a_0, t_1, i_1, \lambda_1, a_1, \ldots, t_{n-1}, i_{n-1}, \lambda_{n-1}, a_{n-1}, t_n, i_n, \lambda_n),$$

where $t_{m+1} \geqslant t_m \geqslant 0$, $(i_m, a_m) \in K$, $\lambda_0 \in R$, $\lambda_{m+1} := \lambda_m - (t_{m+1} - t_m)$ for $m = 0, 1, \ldots, n-1$, and $i_n \in S$. Let H_n denote the set of all admissible histories h_n of the system up to the nth decision epoch, where H_n is endowed with the Borel σ-algebra.

Remark 2.2.

(1) The history h_n here generalizes the one for the usual SMDPs (see Lippman [11], Puterman [15, p. 533]) by taking into account the decision-maker's goals (threshold values) λ_n as well as the decision epochs t_n, the states i_n and actions a_n.
(2) Note that the goals (threshold values) λ_m may be negative for some $m \geqslant 0$, i.e., $\lambda_m < 0$. In this case, the controlled system is thought to be risk-free on behalf of the decision-maker at the mth decision epoch.

Now we are in a position to introduce the concept of a policy.

Definition 2.1. A randomized history-dependent policy is a sequence $\pi = \{\pi_n,\ n = 0, 1, \ldots\}$ of stochastic kernels π_n on A given H_n satisfying

$$\pi_n\big(A(i_n)\ \big|\ h_n\big) = 1, \quad \forall h_n \in H_n,\ n \geqslant 0.$$

The set of all randomized history-dependent policies is denoted by Π.

Remark 2.3. Note that a policy here is similar to those in Ohtsubo and Toyonaga [13], Ohtsubo [14] and Wu and Lin [17] for another criterion in DTMDPs, which include not only the usual states and actions but also the goals (threshold values) λ_n. In fact, both the criterion in this paper and those in [13,14,17] are risk-sensitive, and so it is natural to consider the goals (threshold values) λ for the decision-maker when making decisions.

Let Φ represent the set of all stochastic kernels φ on A given $S \times R$ such that $\varphi(A(i) \mid i, \lambda) = 1$ for all $(i, \lambda) \in S \times R$, and $\mathbb{F}$ denote the set of all measurable functions $f : S \times R \to A$ such that $f(i, \lambda)$ is in $A(i)$ for each $(i, \lambda) \in S \times R$. The functions in $\mathbb{F}$ are called decision functions. A policy $\pi = \{\pi_n\}$ is said to be *randomized Markov* if there is a sequence $\{\varphi_n\}$ of $\varphi_n \in \Phi$ such that $\pi_n(\cdot \mid h_n) = \varphi_n(\cdot \mid i_n, \lambda_n)$ for every $h_n \in H_n$ and $n \geqslant 0$. We write such policies as $\pi = \{\varphi_n\}$. A randomized Markov policy $\pi = \{\varphi_n\}$ is said to be *randomized stationary* if φ_n are independent of n. In this case, we write $\pi = \{\varphi, \varphi, \ldots\}$ as φ for simplicity. Moreover, a randomized Markov policy $\pi = \{\varphi_n\}$ is said to be *deterministic Markov* if there is a sequence $\{f_n\}$ of $f_n \in \mathbb{F}$ such that $\varphi_n(\cdot \mid i, \lambda)$ is the Dirac measure at $f_n(i, \lambda)$ for every $(i, \lambda) \in S \times R$ and $n \geqslant 0$. Similarly, we denote such policies by $\pi = \{f_n\}$. In particular, a deterministic Markov policy $\pi = \{f_n\}$ is said to be *stationary* if f_n are independent of n. For simplicity, we write $\pi = \{f, f, \ldots\}$ as f. We denote by Π_{RM}, Π_{RS}, Π_{DM}, and Π_{DS} the families of all randomized Markov, randomized stationary, deterministic Markov, and stationary policies, respectively. Obviously, $\Pi_{RS} \subset \Pi_{RM} \subset \Pi$ and $\Pi_{DS} \subset \Pi_{DM} \subset \Pi$.

Let Π_0 denote the set of all policies which are independent of threshold values. If a policy $\pi = \{f_n\} \in \Pi_{DM} \cap \Pi_0$, then $f_n(i, \lambda) \equiv f_n(i)$ for all $(i, \lambda) \in S \times R$ and $n \geqslant 0$. Obviously, $\Pi_0 \subset \Pi$. Moreover, for a policy $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$ and $m \geqslant 1$, let ${}^{(m)}\pi := \{\varphi_m, \varphi_{m+1}, \ldots\}$ denote the m-remainder policy of π.

For each $(s, i, \lambda) \in R_+ \times S \times R$ and $\pi \in \Pi$, by the well-known Tulcea's theorem, there exist a unique probability measure space $(\Omega, \mathcal{F}, P^\pi_{(s,i,\lambda)})$ and a stochastic process $\{S_n, J_n, \lambda_n, A_n,\ n \geqslant 0\}$ such that, for each $t \in R_+$, $j \in S$, $a \in A$ and $n \geqslant 0$,

$$P^\pi_{(s,i,\lambda)}(S_0 = s,\ J_0 = i,\ \lambda_0 = \lambda) = 1, \tag{2}$$

$$P^\pi_{(s,i,\lambda)}(A_n = a \mid h_n) = \pi_n(a \mid h_n), \tag{3}$$

$$P^\pi_{(s,i,\lambda)}(S_{n+1} - S_n \leqslant t,\ J_{n+1} = j \mid h_n, a_n) = Q(t, j \mid i_n, a_n), \tag{4}$$

where S_n, J_n, $\lambda_n := \lambda_{n-1} - (S_n - S_{n-1})$ and A_n denote the nth decision epoch, the state, the threshold value and the action chosen at the nth decision epoch, respectively. The expectation operator with respect to $P^\pi_{(s,i,\lambda)}$ is denoted by $E^\pi_{(s,i,\lambda)}$. For simplicity, $P^\pi_{(0,i,\lambda)}$ and $E^\pi_{(0,i,\lambda)}$ is denoted by $P^\pi_{(i,\lambda)}$ and $E^\pi_{(i,\lambda)}$. Without loss of generality, we always set the initial decision epoch $S_0 = 0$ and omit it.

Remark 2.4.

(1) The construction of the probability measure space $(\Omega, \mathcal{F}, P^\pi_{(s,i,\lambda)})$ and the above properties (2)–(4) of the stochastic process $\{S_n, J_n, \lambda_n, A_n,\ n \geqslant 0\}$ follow from those in Limnios and Oprisan [9, p. 33] and Puterman [15, pp. 534–535].
(2) Let $X_0 := 0$, $X_n := S_n - S_{n-1}$ $(n \geqslant 1)$ denote the sojourn times between two successive decision epochs. In this setting, the stochastic process $\{S_n, J_n, \lambda_n, A_n,\ n \geqslant 0\}$ may be equivalently rewritten as $\{X_n, J_n, \lambda_n, A_n,\ n \geqslant 0\}$, where $\lambda_n := \lambda_{n-1} - X_n$.

In applications, it is natural to avoid the possibility of an infinite number of decision epochs within a finite amount of time. To this end, we impose the following assumption *throughout this paper*.

Assumption 2.1. *There exist $\delta > 0$ and $\epsilon > 0$ such that*

$$D(\delta \mid i, a) \leqslant 1 - \epsilon, \quad \forall (i, a) \in K.$$

Remark 2.5.

(1) Assumption 2.1 asserts that (with probability one) only a finite number of decision epochs are made in a finite amount of time, i.e.,

$$P^{\pi}_{(i,\lambda)}\Bigl(\Bigl\{\lim_{n\to\infty} S_n = \infty\Bigr\}\Bigr) = 1, \quad \forall (i,\lambda) \in S \times R \text{ and } \pi \in \Pi. \tag{5}$$

(2) In fact, Assumption 2.1 is an extension of Assumption 11.1.1 in Puterman [15], Assumption 2 in Lippman [11] and Condition 1 in Ross [16] to the case that the *additional* variables λ_n have been considered here.

Under Assumption 2.1, we define an underlying continuous-time state-action process $\{Z(t), A(t),\ t \in R_+\}$ corresponding to the discrete-time process $\{S_n, J_n, A_n\}$ by

$$Z(t) = J_n, \qquad A(t) = A_n, \quad \text{for } S_n \leqslant t < S_{n+1},\ t \in R_+ \text{ and } n \geqslant 0.$$

Definition 2.2. The stochastic process $\{Z(t), A(t),\ t \geqslant 0\}$ is called a (continuous-time) semi-Markov decision process.

For the target set $B \subset S$, we introduce the random variable

$$\tau_B := \inf\bigl\{t \geqslant 0 \bigm| Z(t) \in B\bigr\} \quad (\text{with } \inf\emptyset := \infty)$$

which is the first passage time into the target set B of the process $\{Z(t),\ t \geqslant 0\}$. Obviously, $\tau_B = 0$ when $Z(0) \in B$, and $\tau_B \geqslant S_1$ when $Z(0) \in B^c$.

For every $(i,\lambda) \in S \times R$ and $\pi \in \Pi$, we define the risk probability (risk function) by

$$F^{\pi}(i,\lambda) := P^{\pi}_{(i,\lambda)}(\tau_B \leqslant \lambda)$$

and the corresponding optimal value function by

$$F^*(i,\lambda) := \inf_{\pi \in \Pi} F^{\pi}(i,\lambda).$$

Definition 2.3. A policy $\pi^* \in \Pi$ is called optimal if

$$F^{\pi^*}(i,\lambda) = F^*(i,\lambda), \quad \forall (i,\lambda) \in S \times R.$$

Remark 2.6.

(1) It is easy to see that $F^{\pi}(i,\lambda)$ is well defined. In Section 3, we will show that $F^{\pi}(i,\lambda)$ is measurable with respect to threshold values λ, and $F^*(i,\lambda)$ is nondecreasing and right continuous in λ. However, $F^{\pi}(i,\lambda)$ may be neither nondecreasing nor right continuous in λ, and thus $F^{\pi}(i,\lambda)$ may not be a distribution function of λ (see Example 6.1), whereas $F^{\pi}(i,\lambda)$ are always distribution functions of λ in [10,12] because only the policy class Π_0 is considered there.
(2) Let B represent the set of failure states of a system. Then τ_B denotes the working life of the system and $F^{\pi}(i,\lambda)$ denotes the risk probability that the system eventually fails within λ time units when using policy π. Roughly, our aim is to find an optimal policy which minimizes $F^{\pi}(i,\lambda)$ in the policy class Π. Hence, the background of our model is an optimization problem in the field of risk analysis.

Note that, for every $(i,\lambda) \in B \times R$ and $\pi \in \Pi$, we have $F^{\pi}(i,\lambda) = \mathbb{1}_{[0,\infty)}(\lambda)$, where $\mathbb{1}_D$ is the indicator function on a set D. To avoid this trivial case, in the following we restrict our arguments about $F^{\pi}(i,\lambda)$ to the case $(i,\lambda) \in B^c \times R$. Moreover, we limit ourselves to *randomized Markov* policies below, since, as shown in Theorem 11.1.1 in [15], we also have $F^*(i,\lambda) = \inf_{\pi \in \Pi_{RM}} F^{\pi}(i,\lambda)$.

3. On the optimal value function and optimality equation

In this section, we characterize the optimal value function F^* as well as the risk functions F^{π}, and prove that F^* is a solution to the optimality equation.

To characterize F^{π} and F^*, we introduce some function sets and operators. Let $\mathcal{F}_m$ be the set of functions $F : B^c \times R \to [0,1]$ such that $F(i,\lambda) = 0$ if $\lambda < 0$ and $F(i,\cdot)$ is Borel measurable on R for every $i \in B^c$; and $\mathcal{F}_r$ the set of functions $F \in \mathcal{F}_m$ such that $F(i,\cdot)$ is monotone nondecreasing and right continuous on R for each $i \in B^c$.

We define operators T^a, T^φ, T from $\mathcal{F}_m$ into $\mathcal{F}_m$ as follows: for $F \in \mathcal{F}_m$, $i \in B^c$, $a \in A(i)$ and $\varphi \in \Phi$, if $\lambda \geqslant 0$,

$$T^a F(i,\lambda) := Q(\lambda, B \mid i, a) + \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i, a) F(j, \lambda - t), \tag{6}$$

$$T^\varphi F(i,\lambda) := \sum_{a \in A(i)} \varphi(a \mid i, \lambda) T^a F(i,\lambda), \tag{7}$$

$$TF(i,\lambda) := \min_{a \in A(i)} T^a F(i,\lambda), \tag{8}$$

where $Q(\lambda, B \mid i, a) := \sum_{j \in B} Q(\lambda, j \mid i, a)$, and $T^a F(i,\lambda) = T^\varphi F(i,\lambda) = TF(i,\lambda) := 0$ for $\lambda < 0$.

Remark 3.1.

(1) It follows from the definition of Q and Fubini's theorem that T^a is a map from $\mathcal{F}_m$ into $\mathcal{F}_m$. Moreover, by the definition of φ and the finiteness of $A(i)$, T^φ and T are also maps from $\mathcal{F}_m$ into $\mathcal{F}_m$.
(2) These operators T^a, T^φ and T are based on the characterization of our model in (1), and used to analyze the properties of F^π and F^*; see Lemma 3.3 and Theorems 3.1 and 3.2.

We next give two lemmas about T^a and T.

Lemma 3.1.

(a) *If $G \in \mathcal{F}_r$, then $T^a G$ and TG are both in $\mathcal{F}_r$ for any $a \in A(\cdot)$.*
(b) *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. (a) From the definitions of T^a, T and $G \in \mathcal{F}_r$, it is obvious that $T^a G(i,\cdot)$ and $TG(i,\cdot)$ are nondecreasing on R. Furthermore, since $Q(\cdot, j \mid i, a)$ and $G(i,\cdot)$ are right continuous for each $j \in S$, $i \in B^c$ and $a \in A(i)$, the dominated convergence theorem gives the right continuity of $T^a G(i,\cdot)$ and $TG(i,\cdot)$.

(b) Note that $G_n \geqslant G_{n+1} \geqslant 0$ for each $n \geqslant 0$, and so the limit $G := \lim_{n\to\infty} G_n$ exists. We easily see that $G(i,\cdot)$ is nondecreasing on R. To prove $G \in \mathcal{F}_r$, it remains to show that $G(i,\cdot)$ is right continuous on R. Indeed, for every $\lambda \in R$ and any sequence $\{\lambda_k\}$ in R such that $\lambda_k \downarrow \lambda$, we have $G(i,\lambda_k) \leqslant G_n(i,\lambda_k)$ for any $n,k \geqslant 0$. Hence, $\limsup_{\lambda_k \downarrow \lambda} G(i,\lambda_k) \leqslant \lim_{\lambda_k \downarrow \lambda} G_n(i,\lambda_k) = G_n(i,\lambda)$ for any $n \geqslant 0$ and thus $\limsup_{\lambda_k \downarrow \lambda} G(i,\lambda_k) \leqslant G(i,\lambda)$. On the other hand, since $G(i,\lambda_k) \geqslant G(i,\lambda)$, we obtain $\liminf_{\lambda_k \downarrow \lambda} G(i,\lambda_k) \geqslant G(i,\lambda)$, which together with the previous inequality yields $\lim_{\lambda_k \downarrow \lambda} G(i,\lambda_k) = G(i,\lambda)$. Therefore, $G \in \mathcal{F}_r$ and the proof is completed. □

Lemma 3.2. *For each $F \in \mathcal{F}_m$, there exists a decision function $f \in \mathbb{F}$ such that $T^f F = TF$.*

Proof. By the finiteness of $A(i)$ and the measurable selection theorem (see Bertsekas and Shreve [1, Proposition 7.33, p. 153]), there exists a measurable mapping f from $B^c \times R$ to A such that $f(i,\lambda) \in A(i)$ and $T^f F(i,\lambda) = TF(i,\lambda)$ for each $(i,\lambda) \in B^c \times R$. □

To guarantee that $T^a F^\pi$ and TF^π are well defined (for any $\pi \in \Pi_{RM}$), it is required that F^π is in $\mathcal{F}_m$, that is, $F^\pi(i,\cdot)$ is measurable on R for each $i \in B^c$. To do this, we define some functions F_n^π as below.

Note that for each $(i,\lambda) \in B^c \times R_+$ and $\pi \in \Pi_{RM}$, we have

$$\begin{aligned} F^\pi(i,\lambda) &= P^\pi_{(i,\lambda)}(\tau_B \leqslant \lambda) \\ &= 1 - P^\pi_{(i,\lambda)}(\tau_B > \lambda) \\ &= 1 - P^\pi_{(i,\lambda)}\big(Z(t) \in B^c,\ \forall t \in [0,\lambda]\big) \\ &= 1 - \sum_{m=0}^\infty P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m\big) \\ &= \lim_{n\to\infty}\left[1 - \sum_{m=0}^n P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m\big)\right], \end{aligned} \tag{9}$$

where the fourth equality follows from Assumption 2.1 and the others are obvious. Basing on (9), we define $F^\pi_{-1}(i,\lambda) := \mathbb{1}_{[0,\infty)}(\lambda)$, and $F^\pi_n(i,\lambda) := 1 - \sum_{m=0}^n P^\pi_{(i,\lambda)}(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m)$ if $\lambda \geqslant 0$ and $F^\pi_n(i,\lambda) := 0$ otherwise for $i \in B^c$ and $n \geqslant 0$. Clearly, $\lim_{n\to\infty} F^\pi_n(i,\lambda) = F^\pi(i,\lambda)$, and $F^\pi_n(i,\lambda) \geqslant F^\pi_{n+1}(i,\lambda)$ for every $(i,\lambda) \in B^c \times R$ and $n \geqslant -1$.

Lemma 3.3. *Let $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$ be arbitrary.*

(a) *For each $n \geqslant -1$, $F_n^\pi \in \mathcal{F}_m$ and $F^\pi \in \mathcal{F}_m$.*

(b) *For each $n \geqslant -1$, $F^\pi_{n+1} = T^{\varphi_0} F_n^{{}^{(1)}\pi}$ and $F^\pi = T^{\varphi_0} F^{{}^{(1)}\pi}$, where ${}^{(1)}\pi = \{\varphi_1, \varphi_2, \ldots\} \in \Pi_{RM}$. In particular, $F^\varphi = T^\varphi F^\varphi$ when $\pi = \varphi \in \Pi_{RS}$.*

Proof. (a) To show that $F_n^\pi \in \mathcal{F}_m$, it suffices to prove that $F_n^\pi(i,\cdot)$ is measurable on R for each $i \in B^c$. We do this by induction. When $n = -1$, it is obviously true. Now assume $F_n^\pi(i,\cdot)$ is measurable for some n and every $\pi \in \Pi_{RM}$. It then follows from the property of T^φ that for any $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$,

$$T^{\varphi_0} F_n^{{}^{(1)}\pi}(i,\lambda) = \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda)\left[Q(\lambda, B \mid i,a) + \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i,a) F_n^{{}^{(1)}\pi}(j, \lambda - t)\right]$$

is well defined and measurable in λ, where ${}^{(1)}\pi = \{\varphi_1, \varphi_2, \ldots\} \in \Pi_{RM}$. On the other hand, for $\lambda < 0$, it is clear that $F^\pi_{n+1}(i,\lambda) = T^{\varphi_0} F_n^{{}^{(1)}\pi}(i,\lambda) = 0$, and for $\lambda \geqslant 0$, we have

$$\begin{aligned}
F^\pi_{n+1}(i,\lambda) &= 1 - \sum_{m=0}^{n+1} P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m\big) \\
&= 1 - P^\pi_{(i,\lambda)}(S_1 > \lambda) - \sum_{m=1}^{n+1} P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m\big) \\
&= P^\pi_{(i,\lambda)}(S_1 \leqslant \lambda) - E^\pi_{(i,\lambda)}\Bigg[\sum_{m=1}^{n+1} P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m \,\big|\, S_0, J_0, \lambda_0, \\
&\qquad A_0, S_1, J_1, \lambda_1 = \lambda_0 - (S_1 - S_0)\big)\Bigg] \\
&= P^\pi_{(i,\lambda)}(S_1 \leqslant \lambda) - E^\pi_{(i,\lambda)}\Bigg[\sum_{m=1}^{n+1} \mathbb{1}_{\{J_0 \in B^c, J_1 \in B\}} P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1},\ J_k \in B^c,\ k = 2,\ldots,m \,\big|\, S_0, J_0, \lambda_0, \\
&\qquad A_0, S_1, J_1, \lambda_1 = \lambda_0 - (S_1 - S_0)\big)\Bigg] \\
&= P^\pi_{(i,\lambda)}(S_1 \leqslant \lambda) - \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda) \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i,a)\Bigg[\sum_{m=1}^{n+1} P^\pi_{(i,\lambda)}\big(S_m \leqslant \lambda < S_{m+1}, \\
&\qquad J_k \in B^c,\ k = 2,\ldots,m \,\big|\, S_0 = 0,\ J_0 = i,\ \lambda_0 = \lambda,\ A_0 = a,\ S_1 = t,\ J_1 = j,\ \lambda_1 = \lambda - t\big)\Bigg] \\
&= \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda) D(\lambda \mid i,a) - \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda) \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i,a) \\
&\qquad \times \sum_{m=0}^{n} P^{{}^{(1)}\pi}_{(j,\lambda-t)}\big(S_m \leqslant \lambda - t < S_{m+1},\ J_0 = j,\ J_k \in B^c,\ k = 1,\ldots,m\big) \\
&= \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda)\Bigg[D(\lambda \mid i,a) - \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i,a) \\
&\qquad \times \sum_{m=0}^{n} P^{{}^{(1)}\pi}_{(j,\lambda-t)}\big(S_m \leqslant \lambda - t < S_{m+1},\ J_k \in B^c,\ k = 0,1,\ldots,m\big)\Bigg] \\
&= \sum_{a \in A(i)} \varphi_0(a \mid i,\lambda)\Bigg[Q(\lambda, B \mid i,a) + \sum_{j \in B^c} \int_0^\lambda Q(dt, j \mid i,a)
\end{aligned}$$

$$\times\left[1-\sum_{m=0}^{n}P^{{}^{(1)}\pi}_{(j,\lambda-t)}\big(S_m\leqslant\lambda-t<S_{m+1},\ J_k\in B^c,\ k=0,1,\dots,m\big)\right]\Bigg]$$

$$=\sum_{a\in A(i)}\varphi_0(a\mid i,\lambda)\left[Q(\lambda,B\mid i,a)+\sum_{j\in B^c}\int_0^{\lambda}Q(dt,j\mid i,a)F_n^{{}^{(1)}\pi}(j,\lambda-t)\right],$$

where the fifth equality follows from the properties (2)–(4), the sixth equality is due to the Markov property of policy π and also the properties (2)–(4), and the others are straightforward calculations. Hence, $F^{\pi}_{n+1}(i,\lambda)=T^{\varphi_0}F_n^{{}^{(1)}\pi}(i,\lambda)$ for $\lambda\in R$ and thus $F^{\pi}_{n+1}(i,\cdot)$ is measurable. Therefore, by induction, $F^{\pi}_n(i,\cdot)$ is measurable for every $n\geqslant-1$. Furthermore, since a (pointwise) limit of measurable functions is still measurable, we have $F^{\pi}=\lim_{n\to\infty}F^{\pi}_n\in\mathcal{F}_m$.

(b) From the proof of (a), we have $F^{\pi}_{n+1}=T^{\varphi_0}F_n^{{}^{(1)}\pi}$. Letting $n\to\infty$, since $A(i)$ is finite, we obtain $F^{\pi}=T^{\varphi_0}F^{{}^{(1)}\pi}$ by the dominated convergence theorem. The last statement is obvious. □

Remark 3.2. Lemma 3.3 is mainly based on the characterization of SMDPs such as the properties (2)–(4) and Assumption 2.1.

Next result provides an approximation to F^*.

Theorem 3.1.

(a) *For each $(i,\lambda)\in B^c\times R$ and $n\geqslant-1$, let $F^*_n(i,\lambda):=\inf_{\pi\in\Pi_{RM}}F^{\pi}_n(i,\lambda)$, then $F^*_n\in\mathcal{F}_r$ and $\{F^*_n,\ n\geqslant-1\}$ satisfy equations:*

$$F^*_{-1}=\mathbb{1}_{[0,\infty)},\qquad F^*_{n+1}=TF^*_n,\quad n\geqslant-1.$$

(b) *For each $n\geqslant-1$, there exists a policy $\pi\in\Pi_{DM}$ such that $F^*_n=F^{\pi}_n$.*

(c) $\lim_{n\to\infty}F^*_n=F^*$ *and $F^*\in\mathcal{F}_r$.*

Proof. We prove (a) and (b) together by induction. When $n=-1$, we see that $F^*_{-1}=\mathbb{1}_{[0,\infty)}=F^{\pi}_{-1}\in\mathcal{F}_r$ for any policy $\pi\in\Pi_{DM}$. Now assume that (a) and (b) are true for some $k\geqslant-1$. Thus, $F^*_k\in\mathcal{F}_r$ and there exists a policy $\theta\in\Pi_{DM}$ such that $F^*_k=F^{\theta}_k$. Since $F^*_k\in\mathcal{F}_r$ (hence F^*_k is in $\mathcal{F}_m$), it follows from Lemma 3.2 that there exists a decision function $f\in\mathbb{F}$ such that $TF^*_k=T^fF^*_k$. Therefore, for $\pi=\{f,\theta\}\in\Pi_{DM}$, we have

$$F^*_{k+1}\leqslant F^{\pi}_{k+1}=T^fF^{\theta}_k=T^fF^*_k=TF^*_k,\tag{10}$$

where the inequality is due to the definition of F^*_{k+1}, and the first equality follows from Lemma 3.3(b) with ${}^{(1)}\pi=\theta$. On the other hand, we see that for any $\eta=\{\eta_0,\eta_1,\dots\}\in\Pi_{RM}$,

$$F^{\eta}_{k+1}=T^{\eta_0}F_k^{{}^{(1)}\eta}\geqslant T^{\eta_0}F^*_k\geqslant TF^*_k,\tag{11}$$

where the first equality follows from Lemma 3.3(b) with ${}^{(1)}\eta=\{\eta_1,\eta_2,\dots\}$, and the first inequality is due to the definition of F^*_k again. Since η is arbitrary, taking infimum over Π_{RM} in (11) yields $F^*_{k+1}\geqslant TF^*_k$, which together with (10) and Lemma 3.1(a) gives $F^*_{k+1}=TF^*_k=F^{\pi}_{k+1}\in\mathcal{F}_r$. By induction, the proof of (a) and (b) is achieved.

(c) For any $\pi\in\Pi_{RM}$, we have $F^{\pi}_n\geqslant F^{\pi}_{n+1}\geqslant F^{\pi}$, hence $F^*_n\geqslant F^*_{n+1}\geqslant F^*$, and thus $\lim_{n\to\infty}F^*_n\geqslant F^*$. On the other hand, for arbitrary $\theta\in\Pi_{RM}$, we have $F^{\theta}=\lim_{n\to\infty}F^{\theta}_n\geqslant\lim_{n\to\infty}F^*_n$. Hence, by the arbitrariness of θ, $F^*\geqslant\lim_{n\to\infty}F^*_n$. Therefore, $\lim_{n\to\infty}F^*_n=F^*$. Moreover, since $F^*_n\in\mathcal{F}_r$ and $F^*_n\geqslant F^*_{n+1}$ for all $n\geqslant-1$, it follows from Lemma 3.1(b) that $F^*\in\mathcal{F}_r$. □

Remark 3.3. Indeed, Theorem 3.1 gives a value iteration algorithm for computing the optimal value function F^*, i.e., $F^*(i,\lambda)=\lim_{n\to\infty}T^nF^*_{-1}(i,\lambda)$ with $F^*_{-1}(i,\lambda)=\mathbb{1}_{[0,\infty)}(\lambda)$ for every $(i,\lambda)\in B^c\times R$.

We now state our main result in this section, which establishes the so-called optimality equation.

Theorem 3.2.

(a) *F^* satisfies the optimality equation $F^*=TF^*$.*

(b) *There exists a decision function $f\in\mathbb{F}$ such that $F^*=T^fF^*$.*

Proof. (a) It follows from Lemma 3.3(b) that $F^{\pi}=T^{\varphi_0}F^{{}^{(1)}\pi}\geqslant T^{\varphi_0}F^*\geqslant TF^*$ for any $\pi=\{\varphi_n\}\in\Pi_{RM}$. Since π is arbitrary, we have $F^*\geqslant TF^*$.

It remains to show the reverse inequality. In fact, it follows from Theorem 3.1(a) that $F^*_n(i,\lambda)=TF^*_{n-1}(i,\lambda)\leqslant T^aF^*_{n-1}(i,\lambda)$ for every $(i,\lambda)\in B^c\times R$ and $a\in A(i)$. By Theorem 3.1(c) and dominated convergence theorem, we have $F^*(i,\lambda)\leqslant T^aF^*(i,\lambda)$ for any $a\in A(i)$, and so $F^*(i,\lambda)\leqslant TF^*(i,\lambda)$. Therefore we obtain $F^*=TF^*$.

(b) It is a straightforward result of (a) and Lemma 3.2. □

Remark 3.4. In general, a policy $f \in \Pi_{DS}$ satisfying $F^* = T^f F^*$ may not be optimal. However, a sufficient condition for such a policy to be optimal is given in Theorem 4.2 below.

Remark 3.5. (1) Suppose $P = \{p(j \mid i,a),\ j \in S,\ (i,a) \in K\}$ is a stochastic kernel on S given K. If the semi-Markov decision kernel Q is of the form

$$Q(t, j \mid i, a) = \begin{cases} p(j \mid i,a), & t \geqslant 1, \\ 0, & \text{otherwise}, \end{cases}$$

for every $j \in S$, $(i,a) \in K$, then Theorem 3.2 gives

$$F^*(i,\lambda) = \min_{a \in A(i)} \left[\sum_{j \in B} p(j \mid i,a) + \sum_{j \in B^c} p(j \mid i,a) F^*(j, \lambda - 1) \right], \quad \forall \lambda \geqslant 1,$$

which coincides with the optimality equation for DTMDPs; see Theorem 2.2 in Liu and Huang [12]. In fact, as an extended case, our results in this paper are all true for DTMDPs, and accord with those in [12] in some sense although their analytic method is different.

(2) Suppose that $\widetilde{Q} = \{q(j \mid i,a),\ j \in S,\ (i,a) \in K\}$ is a Q-matrix on S given K. We assume that $\widetilde{Q}$ is conservative, i.e., $\sum_{j \in S} q(j \mid i,a) = 0$ for all $(i,a) \in K$; and stale, i.e., $\sup_{a \in A(i)} q_i(a) < \infty$ for all $i \in S$, where $q_i(a) = -q(i \mid i,a) \geqslant 0$ for each $(i,a) \in K$. If the semi-Markov decision kernel Q is of the form

$$Q(t, j \mid i, a) = \begin{cases} (1 - e^{-q_i(a)t}) \frac{q(j|i,a)}{q_i(a)}, & j \neq i,\ t \geqslant 0, \\ 0, & \text{otherwise}, \end{cases}$$

and $\sup_{(i,a) \in K} q_i(a) < \infty$ (verifies Assumption 2.1), then for $(i,\lambda) \in B^c \times R_+$ and $f \in \Pi_{DS}$, Lemma 3.3(b) gives

$$F^f(i,\lambda) = \sum_{j \in B} \frac{q(j \mid i,f)}{q_i(f)} \big(1 - e^{-q_i(f)\lambda}\big) + \sum_{j \in B^c, j \neq i} \frac{q(j \mid i,f)}{q_i(f)} \int_0^\lambda F^f(j, \lambda - t)\, d\big(1 - e^{-q_i(f)t}\big),$$

which coincides with the one for CTMDPs; see Theorem 1 in Lin, Tomkins and Wang [10].

Note that the optimality equation hasn't been established in Lin, Tomkins and Wang [10]. However, we do this in Theorem 3.2 above and obtain a relation between the optimality equation and an optimal policy in Theorem 4.2. In fact, as an extended case, our results in this paper are all valid for CTMDPs and more general than those in [10] because we consider policies in Π much larger than Π_0 in [10].

The following are some characterization of F^* with respect to T.

Theorem 3.3.

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

Proof. (a) Let $G \in \mathcal{F}_m$ be a fixed point of T. Then we have $G \leqslant F^*_{-1} = \mathbb{1}_{[0,\infty)}$ and $G = T^n G$. Hence it follows from Theorem 3.1(c) that

$$G = \lim_{n\to\infty} T^n G \leqslant \lim_{n\to\infty} T^n F^*_{-1} = \lim_{n\to\infty} F^*_{n-1} = F^*.$$

(b) If G is a function in $\mathcal{F}_m$, then $G \leqslant F^*_{-1}$ and hence $T^n G \leqslant T^n F^*_{-1} = F^*_{n-1}$. Thus, we have $\limsup_n T^n G \leqslant \lim_n F^*_n = F^*$. On the other hand, since $G \geqslant F^*$, we have $T^n G \geqslant T^n F^* = F^*$, and so $\liminf_n T^n G \geqslant F^*$, which together with the previous inequality yields $\lim_{n\to\infty} T^n G = F^*$. □

Corollary 3.1.

(a) *For any policy $\pi \in \Pi_{RM}$, $\lim_{n\to\infty} T^n F^\pi = F^*$.*
(b) *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 $G = F^*$.*
(c) *If there exists a policy $\pi \in \Pi_{RM}$ such that $F^\pi = TF^\pi$, then π is optimal.*

Proof. Statement (a) is an immediate result of Theorem 3.3(b) since $F^\pi \geqslant F^*$, statement (b) is due to Theorem 3.3(a), and finally, statement (c) follows from statement (b). □

4. Properties and existence of optimal policies

This section is devoted to properties and existence of optimal policies. More precisely, we show that there exists an optimal stationary policy and any stationary policy realizing the minimum in the optimality equation is optimal under suitable assumptions. Also, we give sufficient and necessary conditions for the existence of optimal policies which are independent of threshold values.

To characterize optimal policies, for $(i,\lambda)\in B^c\times R$, we define optimal action sets by

$$A^*(i,\lambda):=\bigl\{a\in A(i)\bigm| F^*(i,\lambda)=T^aF^*(i,\lambda)\bigr\},\qquad A^*(i):=\bigcap_{\lambda\in R}A^*(i,\lambda).\tag{12}$$

Then by the finiteness of $A(i)$ and Theorem 3.2(a), $A^*(i,\lambda)\neq\emptyset$. However, $A^*(i)$ may be empty.

We now state some properties of optimal policies.

Theorem 4.1. *Let $\pi=\{\varphi_0,\varphi_1,\ldots\}\in\Pi_{RM}$ be optimal.*

(a) *For each $(i,\lambda)\in B^c\times R$, $A_{\varphi_0}(i,\lambda)\subset A^*(i,\lambda)$ and $\varphi_0(A^*(i,\lambda)\mid i,\lambda)=1$, where*

$$A_{\varphi_0}(i,\lambda)=\bigl\{a\in A(i)\bigm|\varphi_0(a\mid i,\lambda)>0\bigr\}.$$

(b) *If a decision function $f\in\mathbb{F}$ satisfies $f(i,\lambda)\in A_{\varphi_0}(i,\lambda)$ for every $(i,\lambda)\in B^c\times R$, then $f^{(1)}\pi:=\{f,\varphi_1,\varphi_2,\ldots\}$ is optimal.*
(c) *If $\varphi\in\Phi$ is a stochastic kernel such that $F^*=T^\varphi F^*$, then $\{\varphi,\pi\}$ is optimal.*

Proof. (a) Since $\pi=\{\varphi_0,\varphi_1,\ldots\}$ is optimal, it follows from Lemma 3.3(b) and Theorem 3.2(a) that for each $(i,\lambda)\in B^c\times R$,

$$F^*(i,\lambda)=F^\pi(i,\lambda)=T^{\varphi_0}F^{^{(1)}\pi}(i,\lambda)\geqslant T^{\varphi_0}F^*(i,\lambda)\geqslant TF^*(i,\lambda)=F^*(i,\lambda).\tag{13}$$

This means that the terms in (13) are all equal, and thus $T^{\varphi_0}F^*(i,\lambda)=F^*(i,\lambda)$, that is

$$\sum_{a\in A(i)}\varphi_0(a\mid i,\lambda)\bigl[T^aF^*(i,\lambda)-F^*(i,\lambda)\bigr]=0,$$

which together with $T^aF^*(i,\lambda)\geqslant TF^*(i,\lambda)=F^*(i,\lambda)$ implies the desired results in (a).

(b) We show that $F^{f^{(1)}\pi}(i,\lambda)=F^*(i,\lambda)$ for all $(i,\lambda)\in B^c\times R$. If this does not hold, then there exists some $(i,\lambda)\in B^c\times R$ such that $F^{f^{(1)}\pi}(i,\lambda)>F^*(i,\lambda)$. Since $f(i,\lambda)\in A_{\varphi_0}(i,\lambda)$ and $F^{f^{(1)}\pi}(i,\lambda)>F^*(i,\lambda)$, it follows from Lemma 3.3(b) and Theorem 3.2(a) that

$$\begin{aligned}
F^\pi(i,\lambda)&=\sum_{a\in A(i)}\varphi_0(a\mid i,\lambda)T^aF^{^{(1)}\pi}(i,\lambda)\\
&=\varphi_0\bigl(f(i,\lambda)\bigm| i,\lambda\bigr)T^fF^{^{(1)}\pi}(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a\mid i,\lambda)T^aF^{^{(1)}\pi}(i,\lambda)\\
&=\varphi_0\bigl(f(i,\lambda)\bigm| i,\lambda\bigr)F^{f^{(1)}\pi}(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a\mid i,\lambda)T^aF^{^{(1)}\pi}(i,\lambda)\\
&>\varphi_0\bigl(f(i,\lambda)\bigm| i,\lambda\bigr)F^*(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a\mid i,\lambda)T^aF^*(i,\lambda)\\
&\geqslant\varphi_0\bigl(f(i,\lambda)\bigm| i,\lambda\bigr)F^*(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a\mid i,\lambda)TF^*(i,\lambda)\\
&=\varphi_0\bigl(f(i,\lambda)\bigm| i,\lambda\bigr)F^*(i,\lambda)+\sum_{a\in A(i)-f(i,\lambda)}\varphi_0(a\mid i,\lambda)F^*(i,\lambda)\\
&=\sum_{a\in A(i)}\varphi_0(a\mid i,\lambda)F^*(i,\lambda)\\
&=F^*(i,\lambda),
\end{aligned}$$

which contradicts the optimality of π, and therefore $F^{f^{(1)}\pi}(i,\lambda)=F^*(i,\lambda)$ for all $(i,\lambda)\in B^c\times R$, i.e., $f^{(1)}\pi$ is optimal.

(c) By Lemma 3.3(b) and the optimality of π, we obtain $F^{\{\varphi,\pi\}}=T^\varphi F^\pi=T^\varphi F^*=F^*$ and so $\{\varphi,\pi\}$ is optimal. □

In the following, we discuss the existence of optimal policies. To ensures the existence of optimal policies, we need Assumption 4.1 below.

Assumption 4.1. *For every $(i,\lambda)\in B^c\times R$ and $f\in\Pi_{DS}$, $P^f_{(i,\lambda)}(\tau_B<\infty)=1$.*

Remark 4.1. (1) Assumption 4.1 implies that no matter what the initial state is, what the goal for the decision-maker is, and what the stationary policy used is, the controlled system will eventually fail within a finite horizon, that is, the working life of the controlled system is finite. In this setting, Assumption 4.1 is reasonable and mild.

(2) For each $(i,\lambda)\in B^c\times R$, $f\in\Pi_{DS}$ and $t\in R_+$, it follows from the proof of Theorem 3.3 in Limnios and Oprisan [9] that

$$P^f_{(i,\lambda)}(\tau_B\leqslant t)=\sum_{n=1}^{\infty}P^f_{(i,\lambda)}\big(J_k\in B^c,\ 1\leqslant k\leqslant n-1,\ J_n\in B,\ S_n\leqslant t\big),$$

which leads to

$$P^f_{(i,\lambda)}(\tau_B<\infty)=\sum_{n=1}^{\infty}P^f_{(i,\lambda)}\big(J_k\in B^c,\ 1\leqslant k\leqslant n-1,\ J_n\in B\big)=P^f_{(i,\lambda)}\left(\bigcup_{k=1}^{\infty}\{J_k\in B\}\right). \tag{14}$$

Hence, Assumption 4.1 is equivalent to the following form:

$$P^f_{(i,\lambda)}\left(\bigcup_{k=1}^{\infty}\{J_k\in B\}\right)=1,\quad\text{or}\quad P^f_{(i,\lambda)}\left(\bigcap_{k=1}^{\infty}\{J_k\in B^c\}\right)=0,\quad\text{for every }(i,\lambda)\in B^c\times R\text{ and }f\in\Pi_{DS}.$$

(3) The equality (14) brings the SMDPs $\{Z(t),A(t)\}$ into relation with the discrete-time decision processes $\{J_n,A_n\}$, where the transition probabilities are given by $P(j\mid i,a):=Q(\infty,j\mid i,a)$. This implies that we can impose some sufficient conditions on the processes $\{J_n,A_n\}$ to verify Assumption 4.1. Indeed, using Corollary 3.2 in Liu and Huang [12] for DTMDPs, we may have the following: if there exists a real number $\alpha>0$ such that

$$P(B\mid i,a):=\sum_{j\in B}P(j\mid i,a)\geqslant\alpha,\quad\text{for all } i\in B^c,\ a\in A(i), \tag{15}$$

then $P^f_{(i,\lambda)}(\bigcup_{k=1}^{\infty}\{J_k\in B\})=1$, and so Assumption 4.1 holds.

(4) Note that we cannot conclude from Assumption 4.1 that $\lim_{\lambda\to\infty}P^f_{(i,\lambda)}(\tau_B<\lambda)=1$ since these policies defined in this paper depend on threshold values.

We now give a lemma which is key to the existence of optimal policies. To begin with, we first introduce some notations. Let $\widehat{\mathcal{F}}_m$ be the set of functions $F:B^c\times R\to[-1,1]$ such that $F(i,\lambda)=0$ if $\lambda<0$ and $F(i,\cdot)$ is Borel measurable on R for every $i\in B^c$; and $\widehat{T}^f$ an operator from $\widehat{\mathcal{F}}_m$ into $\widehat{\mathcal{F}}_m$ as follows: for each $f\in\mathbb{F}$, $F\in\widehat{\mathcal{F}}_m$ and $(i,\lambda)\in B^c\times R$, if $\lambda\geqslant 0$,

$$\widehat{T}^fF(i,\lambda):=\sum_{j\in B^c}\int_0^{\lambda}Q\big(dt,j\,\big|\,i,f(i,\lambda)\big)F(j,\lambda-t),$$

and $\widehat{T}^fF(i,\lambda):=0$ for $\lambda<0$.

Lemma 4.1. *Suppose that Assumption* 4.1 *holds and $f\in\Pi_{DS}$.*

(a) *Let $F,G\in\mathcal{F}_m$. If $F-G\leqslant\widehat{T}^f(F-G)$, then $F\leqslant G$.*
(b) *F^f is the unique solution in $\mathcal{F}_m$ to equation $F=T^fF$.*

Proof. (a) For each $(i,\lambda)\in B^c\times R$, using the properties (2)–(4) and the fact that $F-G\leqslant 1$, we see that

$$\begin{aligned}\widehat{T}^f(F-G)(i,\lambda)&=\sum_{j\in B^c}\int_0^{\lambda}Q\big(dt,j\,\big|\,i,f(i,\lambda)\big)(F-G)(j,\lambda-t)\\&\leqslant\sum_{j\in B^c}\int_0^{\lambda}Q\big(dt,j\,\big|\,i,f(i,\lambda)\big)\\&=Q\big(\lambda,B^c\,\big|\,i,f(i,\lambda)\big)\\&=P^f_{(i,\lambda)}\big(S_1\leqslant\lambda,\ J_1\in B^c\big)\\&\leqslant P^f_{(i,\lambda)}\big(J_1\in B^c\big).\end{aligned}$$

Now assume that $(\widehat{T}^f)^n(F-G)(i,\lambda) \leqslant P^f_{(i,\lambda)}(\bigcap_{k=1}^n\{J_k \in B^c\})$ for any $(i,\lambda) \in B^c \times R$ and some $n \geqslant 1$. Then we have

$$\begin{aligned}\big(\widehat{T}^f\big)^{n+1}(F-G)(i,\lambda) &= \widehat{T}^f\big(\widehat{T}^f\big)^n(F-G)(i,\lambda)\\ &= \sum_{j\in B^c}\int_0^\lambda Q\big(dt,j\,\big|\,i,f(i,\lambda)\big)\big(\widehat{T}^f\big)^n(F-G)(j,\lambda-t)\\ &\leqslant \sum_{j\in B^c}\int_0^\lambda Q\big(dt,j\,\big|\,i,f(i,\lambda)\big)P^f_{(j,\lambda-t)}\left(\bigcap_{k=1}^n\big\{J_k\in B^c\big\}\right)\\ &= P^f_{(i,\lambda)}\left(\bigcap_{k=1}^{n+1}\big\{J_k\in B^c\big\}\right),\end{aligned}$$

where the last equality follows from the properties (2)–(4). By induction, we obtain

$$(F-G)(i,\lambda) \leqslant \big(\widehat{T}^f\big)^n(F-G)(i,\lambda) \leqslant P^f_{(i,\lambda)}\left(\bigcap_{k=1}^n\big\{J_k\in B^c\big\}\right), \quad \text{for every } (i,\lambda)\in B^c\times R \text{ and all } n\geqslant 1. \tag{16}$$

Note that Assumption 4.1 implies that $P^\pi_{(i,\lambda)}(\bigcap_{k=1}^\infty\{J_k \in B^c\}) = 0$. Letting $n \to \infty$ in inequality (16), we obtain $(F-G)(i,\lambda) \leqslant 0$ for every $(i,\lambda) \in B^c \times R$, which completes the proof of (a).

(b) Let $F \in \mathcal{F}_m$ be a solution to $F = T^f F$. Since F^f satisfies $F^f = T^f F^f$ (by Lemma 3.3(b)), we have $F - F^f = \widehat{T}^f(F - F^f)$, and so statement (a) implies that $F = F^f$. □

Remark 4.2. Note that Lemma 4.1 is slightly similar to Lemma 3.4 in Ohtsubo [14] for another criterion in DTMDPs. However, our method is mainly based on the characterization of SMDPs such as properties (2)–(4).

Now we present another main result in this paper, which shows that any stationary policy realizing the minimum in the optimality equation is optimal and there exists an optimal stationary policy.

Theorem 4.2. *Suppose that Assumption* 4.1 *holds. Then*

(a) *Any policy $f \in \Pi_{DS}$ such that $F^* = T^f F^*$ is optimal.*
(b) *There exists an optimal stationary policy.*

Proof. (a) Since Assumption 4.1 holds and $F^* = T^f F^*$, it follows from Lemma 4.1(b) that $F^f = F^*$ and thus f is optimal.

(b) It is an immediate result of Theorem 3.2(b) and statement (a). □

We next give sufficient and necessary conditions for the existence of optimal policies which are in Π_0. First, we need the lemma below.

Lemma 4.2. *For any fixed policy $f \in \Pi_{DS} \cap \Pi_0$, if $\sup_{i\in B^c} Q(t, B^c \mid i, f(i)) < 1$ for some $t > 0$, then F^f is the unique solution in $\mathcal{F}_m$ to equation $F = T^f F$.*

Proof. It immediately follows from Lemma 3.3(b) that F^f satisfies $F^f = T^f F^f$. Suppose that $F \in \mathcal{F}_m$ is another solution to equation $F = T^f F$. Then we have $F^f - F = \widehat{T}^f(F^f - F)$. To prove the uniqueness, we need only to show that $G = \widehat{T}^f G$, for $G \in \mathcal{F}_m$, implies $G = 0$. We do this as below.

For $(i,t) \in B^c \times R_+$, taking the Laplace transform in $G(i,t) = \widehat{T}^f G(i,t)$ with respect to t yields

$$\widetilde{G}(i,\lambda) = \sum_{j\in B^c}\widetilde{Q}\big(\lambda,j\,\big|\,i,f(i)\big)\widetilde{G}(j,\lambda), \tag{17}$$

where $\widetilde{G}(i,\lambda) := \int_0^\infty e^{-\lambda t}G(i,t)\,dt$, and $\widetilde{Q}(\lambda, j \mid i, f(i)) := \int_0^\infty e^{-\lambda t}Q(dt, j \mid i, f(i))$. Iterating (17) n times, we obtain

$$\widetilde{G}(i,\lambda) = \sum_{j_1\in B^c}\widetilde{Q}\big(\lambda,j_1\,\big|\,i,f(i)\big)\sum_{j_2\in B^c}\widetilde{Q}\big(\lambda,j_2\,\big|\,j_1,f(j_1)\big)\dots\sum_{j_n\in B^c}\widetilde{Q}\big(\lambda,j_n\,\big|\,j_{n-1},f(j_{n-1})\big)\widetilde{G}(j_n,\lambda). \tag{18}$$

On the other hand, suppose that $Q(\delta, B^c \mid i, f(i)) = 1 - \epsilon < 1$ for any $i \in B^c$ and some $\delta > 0$. Then, taking the Laplace transform, for any $\lambda > 0$,

$$\begin{aligned}
\widetilde{Q}\big(\lambda, B^c \bigm| i, f(i)\big) &= \int_0^\infty e^{-\lambda t} Q\big(dt, B^c \bigm| i, f(i)\big)\\
&= e^{-\lambda t} Q\big(t, B^c \bigm| i, f(i)\big)\Big|_0^\infty + \int_0^\infty \lambda e^{-\lambda t} Q\big(t, B^c \bigm| i, f(i)\big)\, dt\\
&= \int_0^\delta \lambda e^{-\lambda t} Q\big(t, B^c \bigm| i, f(i)\big)\, dt + \int_\delta^\infty \lambda e^{-\lambda t} Q\big(t, B^c \bigm| i, f(i)\big)\, dt\\
&\leqslant (1-\epsilon)\int_0^\delta \lambda e^{-\lambda t}\, dt + \int_\delta^\infty \lambda e^{-\lambda t}\, dt\\
&= 1 - \epsilon\big(1 - e^{-\lambda\delta}\big)\\
&= 1 - \beta_\lambda,
\end{aligned} \tag{19}$$

where $\beta_\lambda := \epsilon(1 - e^{-\lambda\delta})$.

Now let $\lambda > 0$ be fixed. It follows from (18) and (19) that $\widetilde{G}(i,\lambda) \leqslant (1-\beta_\lambda)^n \lambda^{-1}$. Note that $(1-\beta_\lambda) < 1$ and thus $\widetilde{G}(i,\lambda) \leqslant \lim_{n\to\infty}(1-\beta_\lambda)^n\lambda^{-1} = 0$, that is,

$$\widetilde{G}(i,\lambda) = \int_0^\infty e^{-\lambda t} G(i,t)\, dt \leqslant 0, \quad \text{for all } i \in B^c. \tag{20}$$

For each $i \in B^c$, since $G(i,t) \geqslant 0$ for all $t \in R_+$, it follows from (20) that

$$G(i,t) = 0, \quad \text{a.e. for } t \in R_+,$$

which together with $G(i,t) = \widehat{T}^f G(i,t)$ implies $G(i,t) = 0$ for all $t \in R_+$. Therefore, we have $G(i,t) = 0$ for all $(i,t) \in B^c \times R$, and the proof is achieved. □

Theorem 4.3. *Suppose that* $\sup_{i\in B^c}\sup_{a\in A(i)} Q(t, B^c \mid i, a) < 1$ *for some* $t > 0$. *Then there exists an optimal policy* $\pi \in \Pi_0$ *if and only if* $A^*(i) \neq \emptyset$ *for all* $i \in B^c$.

Proof. $\Rightarrow$. Let $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_0$ be optimal. By Theorem 4.1(a), $A_{\varphi_0}(i) \equiv A_{\varphi_0}(i,\lambda) \subset A^*(i,\lambda)$ for all $(i,\lambda) \in B^c \times R$, from which it follows that $A_{\varphi_0}(i) \subset \bigcap_{\lambda\in R} A^*(i,\lambda) = A^*(i)$. Since $A_{\varphi_0}(i) \neq \emptyset$, and so $A^*(i) \neq \emptyset$ for all $i \in B^c$.

$\Leftarrow$. Let $A^*(i) \neq \emptyset$ for all $i \in B^c$. Then the measurable selection theorem (see Bertsekas and Shreve [1, Proposition 7.33, p. 153]) gives the existence of $f : B^c \times R \to A$ such that $f(i,\lambda) \equiv f(i) \in A^*(i)$ for all $(i,\lambda) \in B^c \times R$. Therefore, $f \in \Pi_0$ and $F^* = T^f F^*$. By Lemma 4.2 we have $F^f = F^*$, and hence f is optimal. □

5. Value iteration and policy improvement methods

In this section we present a value iteration and a policy improvement methods, which are used to compute the optimal value function and optimal policies, respectively.

From Theorem 3.1 we see that a value iteration is given by $F^*(i,\lambda) = \lim_{n\to\infty} T^n F^*_{-1}(i,\lambda)$, where $F^*_{-1}(i,\lambda) = \mathbb{1}_{[0,\infty)}(\lambda)$ for each $(i,\lambda) \in B^c \times R$. We illustrate this iteration by some examples in Section 6.

Next we consider a policy improvement method under Assumption 4.1. The policy improvement procedure is as below:

(1) Take $n = 0$ and an initial policy $f_n \in \Pi_{DS}$.
(2) At step n, solve the equation $F = T^{f_n} F$ to obtain a function $F^{f_n} \in \mathcal{F}_m$.
(3) If $T^{f_n} F^{f_n} = T F^{f_n}$, stop. If $T^{f_n} F^{f_n} \neq T F^{f_n}$, go to the next step.
(4) Choose a new policy $f_{n+1} \in \Pi_{DS}$ such that $T^{f_{n+1}} F^{f_n} = T F^{f_n}$.
(5) Return to step (2) by replacing n with $n + 1$.

From Lemma 4.1(b) we can uniquely solve the equations in $\mathcal{F}_m$ at step (2). The following is the convergence theorem for this policy improvement procedure.

Theorem 5.1.

(a) *The sequence* $\{F^{f_n}\}$ *is nonincreasing and converges to* F^*.
(b) *If* $T^{f_n} F^{f_n} = T F^{f_n}$, *then* F^{f_n} *is the optimal value function and* f_n *is an optimal policy.*

Proof. (a) From the policy improvement procedure above, we have

$$F^{f_{n+1}} - F^{f_n} = T^{f_{n+1}} F^{f_{n+1}} - T^{f_n} F^{f_n} \leqslant T^{f_{n+1}} F^{f_{n+1}} - T^{f_{n+1}} F^{f_n} = \widehat{T}^{f_{n+1}}\big(F^{f_{n+1}} - F^{f_n}\big),$$

which together with Lemma 4.1(a) gives $F^{f_{n+1}} \leqslant F^{f_n}$. Hence $\{F^{f_n}\}$ is nonincreasing and converges to some function $\widetilde{F} \in \mathcal{F}_m$. We now show that $F^* = \widetilde{F}$. On the one hand, since $F^* \leqslant F^{f_n}$ for all n, we have $F^* \leqslant \widetilde{F}$. On the other hand, it follows that for each $n \geqslant 2$,

$$F^{f_n} = T^{f_n} F^{f_n} \leqslant T^{f_n} F^{f_{n-1}} = T F^{f_{n-1}}.$$

Similarly, we have $F^{f_{n-1}} \leqslant T F^{f_{n-2}}$ and so $F^{f_n} \leqslant T^2 F^{f_{n-2}}$. By induction, we obtain that $F^{f_n} \leqslant T^n F^{f_0}$. Hence, $\widetilde{F} \leqslant F^{f_n} \leqslant T^n F^{f_0}$ for all $n \geqslant 0$. From Corollary 3.1(a), it follows that

$$\widetilde{F} \leqslant \lim_{n\to\infty} T^n F^{f_0} = F^*.$$

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

(b) If $T^{f_n} F^{f_n} = T F^{f_n}$, we see that $f^k = f^n$ for every $k \geqslant n$. Hence it follows from (a) that $F^{f_n} = \widetilde{F} = F^*$, which shows f_n is an optimal policy. □

Remark 5.1. Note that the policy improvement procedure and Theorem 5.1 are slightly similar to those in Ohtsubo [14] for another criterion in DTMDPs.

6. Examples

In this section, we give two examples. One is for the characterization of the risk function, and another is to show how to obtain both the optimal value function and an optimal stationary policy.

The first example shows that there exists a policy $\pi \in \Pi_{DS}$ such that $F^\pi \notin \mathcal{F}_r$.

Example 6.1. Let $S = \{1, 2\}$ be a state space and $B = \{2\}$ a target set. Let $A(1) = \{a_{11}, a_{12}\}$, $A(2) = \{a_{21}\}$, and $A = A(1) \cup A(2)$. Suppose that the semi-Markov decision kernel $Q(t, j \mid i, a)$ is given by

$$Q(t, j \mid 1, a_{11}) = \begin{cases} 1/2, & \text{if } t \geqslant 1,\ j = 1, 2, \\ 0, & \text{otherwise;} \end{cases} \qquad Q(t, j \mid 1, a_{12}) = \begin{cases} 1, & \text{if } t \geqslant 2,\ j = 2, \\ 0, & \text{otherwise;} \end{cases}$$

$$Q(t, j \mid 2, a_{21}) = \begin{cases} 1 - e^{-2t}, & \text{if } t \geqslant 0,\ j = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Note that Assumptions 2.1 holds with $\eta = 1/2$ and $\epsilon = e^{-1}$ in this example. We now define a policy d as follows:

$$d(1, \lambda) = \begin{cases} a_{12}, & \lambda \leqslant 2, \\ a_{11}, & \lambda > 2. \end{cases}$$

Then for $n \geqslant -1$, it follows from Lemma 3.3(b) that

$$F^d_{-1}(1, \lambda) = \mathbb{1}_{[0,\infty)}(\lambda),$$

$$F^d_{n+1}(1, \lambda) = Q\big(\lambda, 2 \bigm| 1, d(1, \lambda)\big) + \int_0^\lambda Q\big(dt, 1 \bigm| 1, d(1, \lambda)\big) F^d_n(1, \lambda - t)$$

for $\lambda \geqslant 0$, and $F^d_{n+1}(1, \lambda) = 0$ otherwise. For all $n \geqslant 0$, we find that $F^d_n(1, \lambda) = 0$ if $\lambda < 2$, and $F^d_n(1, \lambda) = 1$ when $\lambda = 2$. Moreover, if $2 < \lambda < 3$, we have

$$\begin{aligned} F^d_{n+1}(1, \lambda) &= Q(\lambda, 2 \mid 1, a_{11}) + \int_0^\lambda Q(dt, 1 \mid 1, a_{11}) F^d_n(1, \lambda - t) \\ &= \frac{1}{2} + \frac{1}{2} \times F^d_n(1, \lambda - 1), \end{aligned}$$

where $F^d_n(1, \lambda - 1) = 0$ for all $n \geqslant 0$ since $1 < \lambda - 1 < 2$, and thus $F^d_{n+1}(1, \lambda) = \frac{1}{2}$ for all $n \geqslant 0$. Then, using the fact that $F^d(1, \lambda) = \lim_{n\to\infty} F^d_n(1, \lambda)$ yields

$$F^d(1, \lambda) = \begin{cases} 0, & \lambda < 2, \\ 1, & \lambda = 2, \\ 1/2, & 2 < \lambda < 3, \end{cases}$$

which shows that $F^d(1, \lambda)$ is neither nondecreasing nor right continuous in λ, and hence $F^d \notin \mathcal{F}_r$.

Example 6.2. Let $S = \{1, 2, 3\}$ be a state space and $B = \{3\}$ a target set, where states 1, 2 and 3 represent the good state, the medium and the failure ones of a controlled system, respectively. Let $A(1) = \{a_{11}, a_{12}\}$, $A(2) = \{a_{21}, a_{22}\}$, $A(3) = \{a_{31}\}$, and $A = A(1) \cup A(2) \cup A(3)$. The semi-Markov decision kernel Q is of the form: $Q(t, j \mid i, a) = H(t \mid i, a)p(j \mid i, a)$ for every $t \in R_+$, $j \in S$, $(i, a) \in K$, in which $H(t \mid i, a)$ and $p(j \mid i, a)$ denote the distribution functions of the sojourn time and the transition probabilities, respectively. Suppose that the distribution functions $H(t \mid i, a)$ are given by

$$H(t \mid 1, a_{11}) = \begin{cases} 1/25, & t \in [0, 25], \\ 1, & t > 25; \end{cases} \qquad H(t \mid 1, a_{12}) = 1 - e^{-0.08t}, \quad t \in R_+;$$

$$H(t \mid 2, a_{21}) = \begin{cases} 1/20, & t \in [0, 20], \\ 1, & t > 20; \end{cases} \qquad H(t \mid 2, a_{22}) = 1 - e^{-0.15t}, \quad t \in R_+;$$

$$H(t \mid 3, a_{31}) = 1 - e^{-0.2t}, \quad t \in R_+;$$

and the transition probabilities $p(j \mid i, a)$ are given by

$$p(1 \mid 1, a_{11}) = 0, \qquad p(2 \mid 1, a_{11}) = \frac{9}{20}, \qquad p(3 \mid 1, a_{11}) = \frac{11}{20};$$

$$p(1 \mid 1, a_{12}) = 0, \qquad p(2 \mid 1, a_{12}) = \frac{1}{2}, \qquad p(3 \mid 1, a_{12}) = \frac{1}{2};$$

$$p(1 \mid 2, a_{21}) = \frac{1}{5}, \qquad p(2 \mid 2, a_{21}) = 0, \qquad p(3 \mid 2, a_{21}) = \frac{4}{5};$$

$$p(1 \mid 2, a_{22}) = \frac{1}{4}, \qquad p(2 \mid 2, a_{22}) = 0, \qquad p(3 \mid 2, a_{22}) = \frac{3}{4}; \qquad p(3 \mid 3, a_{31}) = 1.$$

First of all, in this example Assumption 2.1 holds with $\eta = 1$ and $\epsilon = e^{-0.2}$. Moreover, condition (15) in Remark 4.1(3) holds with $\alpha = 1/2$, and thus Assumption 4.1 is also fulfilled.

We now compute the optimal value function $F^*(i, \lambda)$ by the value iteration (see Theorem 3.1): $F^*_{-1}(i, \lambda) = \mathbb{1}_{[0,\infty)}(\lambda)$, $F^*_{n+1}(i, \lambda) = TF^*_n(i, \lambda)$ for $i = 1, 2$, $\lambda \in R$ and $n \geqslant -1$. The detailed computation procedure is as below, in which, without loss of generality, a finite interval $[0, 80]$ for λ is considered instead of R.

Value Iteration Procedure (VI-Procedure).

Step 1 (Initialization). Let $n = -1$, and $F^*_n(i, \lambda) = \mathbb{1}_{[0,\infty)}(\lambda)$ for $i = 1, 2$ and $\lambda \in [0, 80]$.

Step 2 (Iteration). Compute these functions $T^a F^*_n(i, \lambda)$ and $F^*_{n+1}(i, \lambda)$ for $a \in A(i)$, $i = 1, 2$ and $\lambda \in [0, 80]$. For $i = 1$,

$$T^{a_{11}} F^*_n(1, \lambda) = \frac{11}{20} \times \frac{\lambda}{25} + \frac{9}{20} \times \frac{1}{25} \times \int_0^{\lambda} F^*_n(2, \lambda - t)\, dt, \quad \lambda \in [0, 25];$$

$$T^{a_{11}} F^*_n(1, \lambda) = \frac{11}{20} + \frac{9}{20} \times \frac{1}{25} \times \int_0^{25} F^*_n(2, \lambda - t)\, dt, \quad \lambda \in (25, 80];$$

$$T^{a_{12}} F^*_n(1, \lambda) = \frac{1}{2} \times \left(1 - e^{-0.08\lambda}\right) + \frac{1}{2} \times 0.08 \times \int_0^{\lambda} e^{-0.08t} F^*_n(2, \lambda - t)\, dt, \quad \lambda \in [0, 80];$$

$$F^*_{n+1}(1, \lambda) = \min\left\{T^{a_{11}} F^*_n(1, \lambda), T^{a_{12}} F^*_n(1, \lambda)\right\}, \quad \lambda \in [0, 80].$$

For $i = 2$,

$$T^{a_{21}} F^*_n(2, \lambda) = \frac{4}{5} \times \frac{\lambda}{20} + \frac{1}{5} \times \frac{1}{20} \times \int_0^{\lambda} F^*_n(1, \lambda - t)\, dt, \quad \lambda \in [0, 20];$$

$$T^{a_{21}} F^*_n(2, \lambda) = \frac{4}{5} + \frac{1}{5} \times \frac{1}{20} \times \int_0^{20} F^*_n(1, \lambda - t)\, dt, \quad \lambda \in (20, 80];$$

$$T^{a_{22}} F^*_n(2, \lambda) = \frac{3}{4} \times \left(1 - e^{-0.15\lambda}\right) + \frac{1}{4} \times 0.15 \times \int_0^{\lambda} e^{-0.15t} F^*_n(1, \lambda - t)\, dt, \quad \lambda \in [0, 80];$$

$$F^*_{n+1}(2, \lambda) = \min\left\{T^{a_{21}} F^*_n(2, \lambda), T^{a_{22}} F^*_n(2, \lambda)\right\}, \quad \lambda \in [0, 80].$$

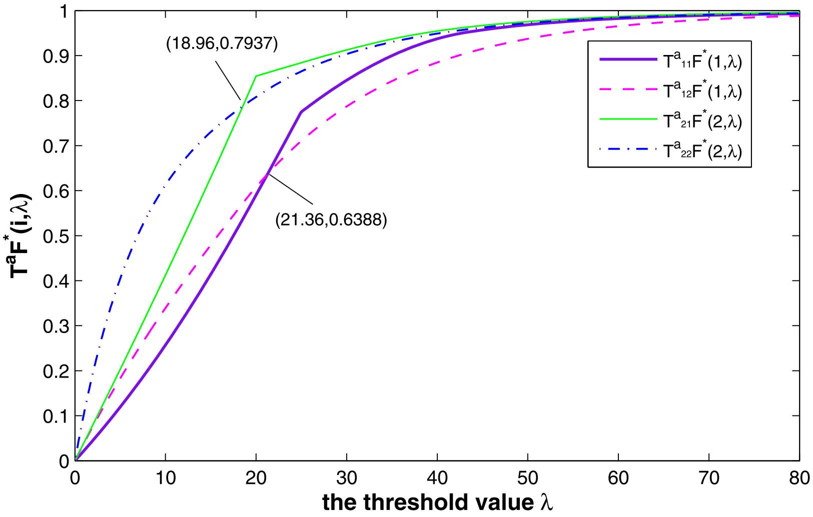

Fig. 6.1. The functions $T^a F^*(i,\lambda)$.

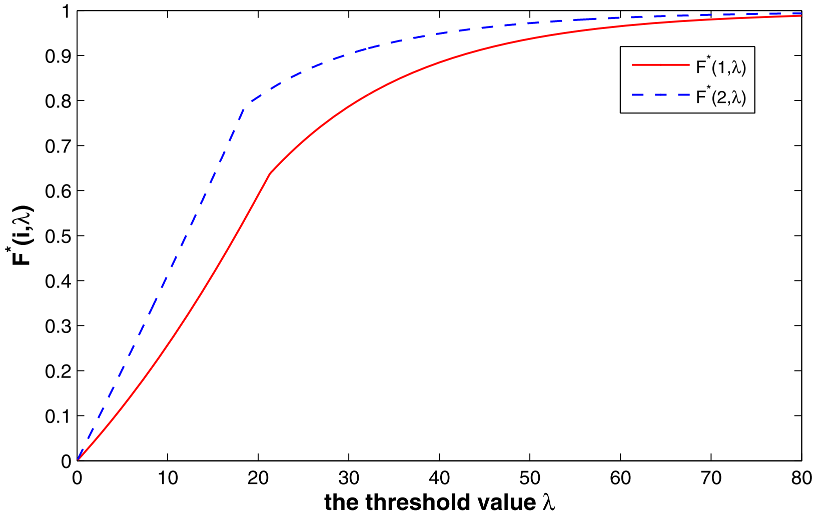

Fig. 6.2. The optimal value function $F^*(i,\lambda)$.

Step 3 (Accuracy control). If $F_n^*(i,\lambda) - F_{n+1}^*(i,\lambda) \leqslant 10^{-12}$ for $i = 1, 2$ and $\lambda \in [0, 80]$, go to Step 4; otherwise, go to Step 2 by replacing n with $n+1$.

Step 4 (Plotting). Plot out the graphs of these functions $T^{a_{11}} F_n^*(1,\lambda)$, $T^{a_{12}} F_n^*(1,\lambda)$, $T^{a_{21}} F_n^*(2,\lambda)$, $T^{a_{22}} F_n^*(2,\lambda)$, $F_{n+1}^*(1,\lambda)$ and $F_{n+1}^*(2,\lambda)$ for $\lambda \in [0, 80]$; see Figs. 6.1 and 6.2 above.

Remark 6.1.

(1) Note that we have not computed the optimal value function $F^*(3,\lambda)$ because it is clear that $F^*(3,\lambda) = \mathbb{1}_{[0,\infty)}(\lambda)$ for each $\lambda \in R$.

(2) This procedure is implemented in *Matlab*, which stops if the condition "$F_n^* - F_{n+1}^* \leqslant 10^{-12}$" is satisfied. Indeed, F_n^* and F_{n+1}^* are thought to be identical, and both equal to F^* when there is some n (sufficiently large) such that $F_n^* - F_{n+1}^* \leqslant 10^{-12}$. This together with the iteration method $F_{n+1}^* = T F_n^*$ shows that F^* satisfies the optimality equation $F^* = T F^*$.

From Figs. 6.1 and 6.2 and the VI-procedure, we have the following.

(a) In Fig. 6.1, $T^{a_{11}} F^*(1,80) = T^{a_{12}} F^*(1,80) = 1$. Using the fact that (by Lemma 3.1(a)) $T^a F^*(1,\lambda)$ is nondecreasing in λ yields that $T^{a_{11}} F^*(1,\lambda) = T^{a_{12}} F^*(1,\lambda) = 1$ for all $\lambda > 80$. Obviously, $T^{a_{11}} F^*(1,\lambda) = T^{a_{12}} F^*(1,\lambda) = 0$ for any $\lambda < 0$. Similar conclusions can be obtained for the functions $T^{a_{21}} F^*(2,\lambda)$ and $T^{a_{22}} F^*(2,\lambda)$.

(b) In Fig. 6.1, $T^{a_{11}}F^*(1,\lambda)$ is below $T^{a_{12}}F^*(1,\lambda)$ for every $\lambda \in (0, 21.36)$, but $T^{a_{11}}F^*(1,\lambda)$ is above $T^{a_{12}}F^*(1,\lambda)$ for $\lambda \in (21.36, 80)$. This implies that action a_{11} is with lower risk than action a_{12} when $\lambda \in (0, 21.36)$, and a_{12} is with lower risk than a_{11} if $\lambda \in (21.36, 80)$, which means that what action is optimal depends on the threshold value λ. Similarly, we have the same conclusion for the functions $T^{a_{21}}F^*(2,\lambda)$ and $T^{a_{22}}F^*(2,\lambda)$.

(c) In fact, Fig. 6.2 is obtained from Fig. 6.1 by using the optimality equation $F^*(i,\lambda) = \min_{a \in A(i)}\{T^a F^*(i,\lambda)\}$. More clearly, we have

$$F^*(1,\lambda) = \begin{cases} T^{a_{11}}F^*(1,\lambda) = T^{a_{12}}F^*(1,\lambda) = 0, & \lambda \leqslant 0, \\ T^{a_{11}}F^*(1,\lambda), & 0 < \lambda < 21.36, \\ T^{a_{11}}F^*(1,\lambda) = T^{a_{12}}F^*(1,\lambda), & \lambda = 21.36, \\ T^{a_{12}}F^*(1,\lambda), & 21.36 < \lambda < 80, \\ T^{a_{11}}F^*(1,\lambda) = T^{a_{12}}F^*(1,\lambda) = 1, & \lambda \geqslant 80, \end{cases} \tag{21}$$

$$F^*(2,\lambda) = \begin{cases} T^{a_{21}}F^*(2,\lambda) = T^{a_{22}}F^*(2,\lambda) = 0, & \lambda \leqslant 0, \\ T^{a_{21}}F^*(2,\lambda), & 0 < \lambda < 18.96, \\ T^{a_{21}}F^*(2,\lambda) = T^{a_{22}}F^*(2,\lambda), & \lambda = 18.96, \\ T^{a_{22}}F^*(2,\lambda), & 18.96 < \lambda < 80, \\ T^{a_{21}}F^*(2,\lambda) = T^{a_{22}}F^*(2,\lambda) = 1, & \lambda \geqslant 80. \end{cases} \tag{22}$$

(d) In Fig. 6.2, both $F^*(1,\lambda)$ and $F^*(2,\lambda)$ are monotone nondecreasing and continuous in λ, that is, $F^* \in \mathcal{F}_r$. Moreover, $F^*(1,\lambda)$ is below $F^*(2,\lambda)$ for every $\lambda > 0$, which means that the controlled system occupying state 1 is less likely to fail. This fact may be due to that state 1 is a good state whereas state 2 is a medium one.

(e) Optimal policies can be derived from (21) and (22) above. More precisely, if a policy f^* is defined by

$$f^*(1,\lambda) = \begin{cases} a_{12}, & \lambda \leqslant 0, \\ a_{11}, & 0 < \lambda \leqslant 21.36, \\ a_{12}, & 21.36 < \lambda \leqslant 80, \\ a_{11}, & \lambda > 80, \end{cases} \qquad f^*(2,\lambda) = \begin{cases} a_{22}, & \lambda \leqslant 0, \\ a_{21}, & 0 < \lambda \leqslant 18.96, \\ a_{22}, & 18.96 < \lambda \leqslant 80, \\ a_{21}, & \lambda > 80, \end{cases}$$

it then follows from (21) and (22) that $F^*(i,\lambda) = T^{f^*}F^*(i,\lambda)$ for $i = 1, 2$ and all $\lambda \in R$, and therefore (by Theorem 4.2(a)) f^* is an optimal stationary policy.

In addition, by the definition of $A^*(i,\lambda)$ in (12), we see that

$$A^*(1,\lambda) = \begin{cases} \{a_{11}, a_{12}\}, & \lambda \leqslant 0, \\ \{a_{11}\}, & 0 < \lambda < 21.36, \\ \{a_{11}, a_{12}\}, & \lambda = 21.36, \\ \{a_{12}\}, & 21.36 < \lambda < 80, \\ \{a_{11}, a_{12}\}, & \lambda \geqslant 80, \end{cases} \qquad A^*(2,\lambda) = \begin{cases} \{a_{21}, a_{22}\}, & \lambda \leqslant 0, \\ \{a_{21}\}, & 0 < \lambda < 18.96, \\ \{a_{21}, a_{22}\}, & \lambda = 18.96, \\ \{a_{22}\}, & 18.96 < \lambda < 80, \\ \{a_{21}, a_{22}\}, & \lambda \geqslant 80, \end{cases}$$

and so $A^*(1) = \bigcap_{\lambda \in R} A^*(1,\lambda) = \emptyset$, and $A^*(2) = \bigcap_{\lambda \in R} A^*(2,\lambda) = \emptyset$, which shows that (by Theorem 4.3) there is no optimal policy in Π_0.

Remark 6.2. Example 6.2 shows that we can not always find an optimal policy in Π_0, which are independent of threshold values. In fact, since a risk minimizing criterion is risk-sensitive, it is natural and necessary for the decision-makers to consider the threshold values as well as the states when making decisions. Hence, optimal policies usually depends on threshold values, whereas a policy independent of threshold values cannot be optimal in many applications.

References

[1] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control: The Discrete-Time Case, Athena Scientific, Belmont, MA, 1996.

[2] X.R. Cao, Stochastic Learning and Optimization: A Sensitivity-Based Approach, Springer, New York, 2007.

[3] E.A. Feinberg, Continuous time discounted jump Markov decision processes: A discrete-event approach, Math. Oper. Res. 29 (2004) 492–524.

[4] X.P. Guo, Constrained optimality for average cost continuous-time Markov decision processes, IEEE Trans. Automat. Control 52 (2007) 1139–1143.

[5] X.P. Guo, X.R. Cao, Optimal control of ergodic continuous-time Markov chains with average sample-path rewards, SIAM J. Control Optim. 44 (2005) 29–48.

[6] O. Hernández-Lerma, J.B. Lasserre, Discrete-Time Markov Control Processes: Basic Optimality Criteria, Springer-Verlag, New York, 1996.

[7] Y.H. Huang, X.P. Guo, First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs, Acta Math. Appl. Sin., in press.

[8] D. Klabjan, D. Adelman, Existence of optimal policies for semi-Markov decision processes using duality for infinite linear programming, SIAM J. Control Optim. 44 (2006) 2104–2122.

[9] N. Limnios, J. Oprisan, Semi-Markov Processes and Reliability, Birkhäuser, Boston, 2001.

[10] Y.L. Lin, R.J. Tomkins, C.L. Wang, Optimal models for the first arrival time distribution function in continuous time – with a special case, Acta Math. Appl. Sin. 10 (1994) 194–212.
[11] S.A. Lippman, Semi-Markov decision processes with unbounded rewards, Management Sci. 19 (1973) 717–731.
[12] J.Y. Liu, S.M. Huang, Markov decision processes with distribution function criterion of first-passage time, Appl. Math. Optim. 43 (2001) 187–201.
[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, Optimal threshold probability in undiscounted Markov decision processes with a target set, Appl. Math. Anal. Comp. 149 (2004) 519–532.
[15] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, 1994.
[16] S.M. Ross, Average cost semi-Markov decision processes, J. Appl. Probab. 7 (1970) 649–656.
[17] C.B. Wu, Y.L. Lin, Minimizing risk models in Markov decision processes with policies depending on target values, J. Math. Anal. Appl. 231 (1999) 47–67.