Haifeng Huo; Xian Wen

First passage risk probability optimality for continuous time Markov decision processes

Kybernetika, Vol. 55 (2019), No. 1, 114–133

Persistent URL: http://dml.cz/dmlcz/147708

KYBERNETIKA — VOLUME 55 (2019), NUMBER 1, PAGES 114–133

FIRST PASSAGE RISK PROBABILITY OPTIMALITY FOR CONTINUOUS TIME MARKOV DECISION PROCESSES

Haifeng Huo and Xian Wen

In this paper, we study continuous time Markov decision processes (CTMDPs) with a denumerable state space, a Borel action space, unbounded transition rates and nonnegative reward function. The optimality criterion to be considered is the first passage risk probability criterion. To ensure the non-explosion of the state processes, we first introduce a so-called drift condition, which is weaker than the well known regular condition for semi-Markov decision processes (SMDPs). Furthermore, under some suitable conditions, by value iteration recursive approximation technique, we establish the optimality equation, obtain the uniqueness of the value function and the existence of optimal policies. Finally, two examples are used to illustrate our results.

Keywords: continuous time Markov decision processes, first passage time, risk probability criterion, optimal policy

Classification: 90C40, 60E20

1. INTRODUCTION

This paper consider the risk probability optimality in first passage for continuous time Markov decision processes with a denumerable state space, a Borel action space, unbounded transition rates and nonnegative reward function.

As is well known, there are a large number of works on the Markov decision processes (MDPs), see [4, 2, 10, 8, 9, 22, 5, 11, 19, 20, 27, 24, 17], the horizon of MDPs with either finite or infinite. However, many practical situations such as ruin problems[24, 17], reliability[17], maintenance[17] are involved in a random horizon. Inspired by these situations, the first passage performance criteria are introduced into the MDPs. The literature on the first passage optimality problems for MDPs can be classified into two groups: (*i*) One focuses on the first passage expected criterion (see, for instance [5, 11, 19, 26, 12, 6, 15]), which means that the expected total rewards during a rand time that the state process first enters a given target set. This criterion can usually be regarded as a generalization of the standard criterion [3, 4, 8, 9, 22]. (*ii*) The other is first passage risk probability criterion, which usually refers to the probability of the total rewards does not exceed a reward level (profit goal) during a first passage time that the state process first enters a target set. This paper belongs to the second group for MDPs, which have

DOI: 10.14736/kyb-2019-1-0114

been discussed in [10, 14, 18, 21, 20, 27]. More precisely, [21] consider risk minimizing problems in discrete time Markov decisions processes (DTMDPs) with a target set. They show that the value function is the unique solution to an optimality equation, and that there exists an optimal stationary policy. Huang and Guo [10] consider the first passage risk probability problem for semi-Markov decisions processes (SMDPs), and obtain the optimality equation and the existence of optimal policies by using a successive approximation technique. Furthermore, Huang, Zou and Guo [14] investigate the minimum risk probability with loss rates for SMDPs. They establish the optimality equation, give suitable conditions to prove the existence of optimal policies, and develop an algorithm for computing ε-optimal policies. To the best of our knowledge, all of these existing literatures on the first passage risk probability problem only focus on the SMDPs or the DTMDPs. However, CTMDPs with the first passage risk probability criterion is considered in the less known article [18], where the authors discussions are restricted to the stationary policies and bounded transition rates. This paper is an attempt to investigate this criterion for CTMDPs with unbounded transition rates and history-dependent policies, and point out the gap between CTMDPs and SMDPs.

A common feature to the risk probability criterion (see, [10, 14, 20, 27, 25, 16]) is that the decision maker considers the reward levels as well as the system states when making decisions, which is different from the classical expected criterion (see [3, 4, 6, 22, 23]) and average criterion (see [4, 22, 28]) for CTMDPs. Therefore, we can not directly use the results of the classical standard criterions for CTMDPs. Actually we need to introduce a class of history-dependent policies, which depends not only on the usual states but also on reward levels, and establish a new probability space (see Section 2). Secondly, motivated by many practical problems, such as queueing control and population processes, where the transition rates are unbounded. We will consider in this paper the case when the transition rates are unbounded. To deal with this case, we first use the drift condition (see Theorem 3.3) to ensure the non-explosion of the state processes $\{x_t, t \geq 0\}$, which is weaker than the well known regular condition for SMDPs in [10, 12, 13, 14], see Remark 3.4. Furthermore, under some suitable conditions, we not only establish the first passage risk probability optimality equation and show the existence of optimal policies, but also use a value iteration technique to calculate the value function (see Theorem 3.10). Finally, we illustrate our results with two examples. The first one is used to verify our conditions for CTMDPs with unbounded transition rates, the second one for the numerical calculations of the value function and an optimal policy by value iteration techniques.

The rest of this paper is organized as follows. In Section 2 we introduce the control model for CTMDPs and the first passage risk probability optimality problem. We are concerned with the existence and the computation aspects of a risk probability optimal policy for CTMDPs, which are stated in Section 3. Finally, we illustrated our results with two examples in Section 4.

2. THE CONTROL MODEL

The model of continuous-time MDP consists of five components

$$\{E, (A(i) \subseteq A, i \in E), q(j \mid i, a), B, r(i, a), \} \tag{1}$$

with the following meaning: (1) The state space E is a nonempty denumerable set. (2) The action space A is a Borel space, endowed with a Borel σ-algebra $\mathcal{B}(A)$. $A(i) \in \mathcal{B}(A)$ is the set of admissible actions in state $i \in E$, which is assumed to be finite. Let $K := \{(i,a)|i \in E, a \in A(i)\}$ be the set of feasible pairs of states and actions. (3) The transition rate $q(j|i,a)$ satisfies $q(j|i,a) \geq 0$ for all $(i,a) \in K$ and $j \neq i$, which is assumed to be conservative (i. e. $\sum_{j\in S} q(j|i,a) = 0$) and stable (i. e., $q^*(i) := \sup_{a\in A(i)} q_i(a) < \infty$), where $q_i(a) := -q(i|i,a) \geq 0$ for all $(i,a) \in K$. (4) The target set B is a measurable subset of E. (5) The nonnegative measurable reward function $r(i,a)$ satisfies $r(i,a) > 0$ for each $i \in B^c, a \in A(i)$, where B^c denotes the complement of B.

The evolution of CTMDPs with the first passage risk probability criterion may be described as follows. When the system state is i_0 at the initial decision epoch $t_0 = 0$, there is a common reward level (profit goal) $\lambda_0 \in R^+ := [0, +\infty)$ in the mind of a decision maker, that is, the decision maker tries to control the reward no more than λ_0 before the system state first passage time falls into the target set B. Then, the decision marker chooses an action $a_0 \in A(i_0)$. Consequently, the system stays at i_0 until time t_1, the sojourn time $\theta_1 := t_1 - t_0$ following exponential distribution with parameter $q_{i_0}(a_0)(q_{i_0}(a_0) \neq 0)$, and then the next decision epoch comes. At time t_1, the following happen: (1) the system goes into a new state i_1 with the transition probability $\frac{q(i_1|i_0,a_0)}{q_i(a_0)}$. (2) The decision marker gets a reward $r(i_0,a_0)t_1$. There is a remaining profit goal $\hat{\lambda}_1 = [\lambda_0 - r(i_0,a_0)t_1]^+$ for the decision marker, where $[x]^+ = \max(x,0)$. Thus, the decision marker chooses an action $a_1 \in A(i_1)$ based on the current state i_1, reward level $\hat{\lambda}_1$ and the previous state i_0, reward level λ_0. The system is developed in this way until the system state falls into the target set B.

As described above, we know that t_k $(k \geq 0)$ is the kth decision epoch, i_k is the state of the process on $[t_k, t_{k+})$, a_k is the action of the process at time t_k, $\theta_{k+1} := t_{k+1} - t_k$ is the sojourn time at state i_k and $\hat{\lambda}_k$ is the reward level at time t_k,

$$\hat{\lambda}_{k+1} := [\hat{\lambda}_k - r(i_k,a_k)\theta_{k+1}]^+ := L(i_k, \hat{\lambda}_k, a_k, \theta_{k+1}), \quad \text{where } \hat{\lambda}_0 := \lambda_0. \tag{2}$$

Due to the decision maker choosing actions to be considered not only on the usual system states but also on the reward levels, we need to reconstruct a probability space. The sample space Ω is given by $\Omega := \Omega^0 \bigcup \{(i_0, \lambda_0, t_1, i_1, \lambda_1, \ldots, t_k, i_k, \lambda_k, \infty, \Delta, \infty, \ldots)|\ i_0 \in E, \lambda_0 \in [0, +\infty), i_l \in E, \lambda_l \in [0, +\infty), t_l \in (0, \infty)$, for each $1 \leq l \leq k, k \geq 1\}$, where $E_\Delta := E \bigcup \{\Delta\}$ (with some $\Delta \notin E$), $\Omega^0 := E \times [0, +\infty) \times ((0, +\infty] \times E \times [0, +\infty))^\infty$. Let $\mathscr{F}$ be the corresponding Borel σ-algebra on Ω. Then, we obtain a measurable space $(\Omega, \mathscr{F})$.

For each $k \geq 0$, $e := (i_0, \lambda_0, t_1, i_1, \lambda_1, \ldots, t_k, i_k, \lambda_k, \ldots) \in \Omega$, let $h_0(e) := (i_0, \lambda_0)$,$h_k(e)$ $:= (i_0, \lambda_0, t_1, i_1, \lambda_1, \ldots, t_k, i_k, \lambda_k)$ denote the k-component internal history, and define the measurable mappings X_k, Λ_k, T_k on Ω as follows:

$$X_k(e) := i_k, \quad \Lambda_k(e) := \lambda_k, \quad T_0(e) := t_0 = 0, \quad T_k(e) := t_k, \quad T_\infty := \lim_{k\to\infty} T_k(e).$$

For simplicity, we often omit the argument e. Moreover, define the state process $\{x_t, t \geq 0\}$ by

$$x_t := \sum_{k\geq 0} I_{\{T_k \leq t < T_{k+1}\}} i_k + \Delta I_{\{t \geq T_\infty\}}, \tag{3}$$

where I_E denotes the indicator function on any set E. The controlled process after moment T_∞ is considered to be absorbed in the isolated state $x_\infty := \Delta \notin E$. Then, let $q(\cdot|\Delta, a_\Delta) :\equiv 0$, $r(\Delta, a_\Delta) :\equiv 0$, $A(\Delta) := \{a_\Delta\}$, $A_\Delta := A \cup \{a_\Delta\}$, where a_Δ is an isolated point.

Definition 2.1. A *deterministic history-dependent policy* $\pi(e,t)$ is defined by a sequence $(f_k, k \geq 0)$ such that $f_k(h_k(e))$ is a Borel measurable function from Ω onto A_Δ, for each $e = (i_0, \lambda_0, t_1, i_1, \lambda_1, \ldots, t_k, i_k, \lambda_k, \ldots) \in \Omega$,

$$\pi(e,t) = I_{\{t=0\}} f_0(i_0, \lambda_0) + \sum_{k\geq 0} I_{\{T_k < t \leq T_{k+1}\}} f_k(h_k(e)) + I_{\{t \geq T_\infty\}} \delta_{a_\Delta}(da), \tag{4}$$

where $\delta_{a_\Delta}(da)$ is the Dirac measure on A_Δ at the point a_Δ. Such a policy is denoted by $\pi = \{f_0, f_1, \ldots\}$ for simplicity.

A policy $\pi = \{f_0, f_1, \ldots\} \in \Pi$ is called *Markov* if there are some measurable functions f_k^M on A_∞ given $E_\Delta \times [0, \infty)$ such that

$$\pi(e,t) = I_{\{t=0\}} f_0^M(i_0, \lambda_0) + \sum_{k\geq 0} I_{\{T_k < t \leq T_{k+1}\}} f_k^M(i_k, \lambda_k) + I_{\{t \geq T_\infty\}} \delta_{a_\Delta}(da). \tag{5}$$

If there is a deterministic Markov policy $\pi = \{f_0, f_1, \ldots\} \in \Pi_m$ such that $f_k(k \geq 0)$ is independent of k, then this policy is called *stationary*. Such a stationary policy is denoted as f for simplicity.

We denote by Π, Π_m, Π_s the set of all deterministic history-dependent, deterministic Markov and deterministic stationary policies respectively. It is clear that $\Pi_s \subset \Pi_m \subset \Pi$.

For any policy $\pi \in \Pi$, employing [7, 16], the jumps intensity function of the process $\{x_t, t \geq 0\}$ is defined as follows:

$$m^\pi(j|e,t) \quad = \quad I_{\{t=0\}} m_0^\pi(j|i_0, \lambda_0) + \sum_{k\geq 0} I_{\{T_k < t \leq T_{k+1}\}} m_k^\pi(j|h_k(e)), \tag{6}$$

where $m_0^\pi(j|i_0, \lambda_0) := q(j|i_0, f_0(i_0, \lambda_0)) I_{\{j \neq i_0\}}$, $m_k^\pi(j|h_k(e)) := q(j|i_k, f(h_k(e))) I_{\{j \neq i_k\}}$.

For any initial distribution ν on $E \times R$ and policy $\pi = \{f_0, f_1, \ldots\} \in \Pi$, due to the changes of the reward levels, we construct the measure P_ν^π on the measurable space $(\Omega, \mathscr{F})$ as follows. Let $H_0 = E \times R^+$ and $H_k = (E \times R^+) \times ((0, \infty] \times E_\Delta \times R^+)^k, k = 1, 2, \ldots$. The measure P_ν^π on $H_k (k \geq 0)$ is given by $P_{\nu,0}^\pi(i, d\lambda) = \nu(i, d\lambda)$ for $(i, d\lambda) \in E \times \mathcal{B}(R^+)$,

$$\begin{aligned} P_{\nu,1}^\pi(\nu \times (dt_1, \mathrm{d}\lambda_1, j)) \quad &:= \quad \int_\Gamma P_{\nu,0}^\pi(i_0, \lambda_0) m_0^\pi(j|i_0, \lambda_0) \\ &\quad \times \exp\{-m_0^\pi(E|i_0, \lambda_0) t_1\} \\ &\quad \times \delta_{L(i_0, \lambda_0, f_0(i_0, \lambda_0), t_1)}(\mathrm{d}\lambda_1)\, \mathrm{d}t_1, \end{aligned} \tag{7}$$

$$\begin{aligned} P_{\nu,k+1}^\pi(\nu \times (dt_{k+1}, d\lambda_{k+1}, j)) \quad &:= \quad \int_\Gamma P_{\nu,k}^\pi(dh_k) I_{\{\theta_k < \infty\}} m_k^\pi(j|h_k(e)) \\ &\quad \times \exp\{-m_k^\pi(E|h_k(e))(t_{k+1} - t_k)\} \\ &\quad \times \delta_{L(i_k, \lambda_k, f_k(h_k(w)), t_{k+1} - t_k)}(\mathrm{d}\lambda_{k+1})\, \mathrm{d}t_{k+1}, \end{aligned} \tag{8}$$

$$P_\nu^\pi(\nu \times (\infty, \infty, \Delta)) := \int_\Gamma P_\nu^\pi(dh_k) I_{\{\theta_k = \infty\}} + I_{\{\theta_k < \infty\}} \times \exp\{-\int_0^\infty m_k^\pi(E|h_k(w))\, \mathrm{d}v\},$$

where $\Gamma \in \mathcal{B}(H_k), m_k^\pi(E|h_k(e)) := -q(i_k|i_k, f(h_k(e)))$, $\mathcal{B}(X)$ is the σ-algebra on X.

For any initial distribution ν on $E \times R^+$ and policy $\pi \in \Pi$, according to the extension of the well-known Ionescu Tulcea theorem (e. g., Proposition 7.45 in [1]), there exists a unique probability space $(\Omega, \mathscr{F}, P_\nu^\pi)$ such that the probability measure P_ν^π has a projection onto H_k satisfying (7). Let $\mathbb{E}_\nu^\pi$ be its corresponding expectation operator. In particular, $\mathbb{E}_\nu^\pi$ and P_ν^π will be respectively written as $\mathbb{E}_{(i,\lambda)}^\pi$ and $P_{(i,\lambda)}^\pi$ when the initial distribution ν is concentrated on state (i, λ).

Let the random variable τ_B be the first passage time into the target set B of the state process $\{x_t, t \geq 0\}$.

$$\tau_B = \begin{cases} \inf\{t \geq 0 : x_t \in B\}, & \text{if } \{t \geq 0 : x_t \in B\} \neq \emptyset; \\ +\infty, & \text{otherwise.} \end{cases}$$

For each $(i, \lambda) \in E \times R^+, \pi \in \Pi$, we define the first passage risk probability criterion $F^\pi(i, \lambda)$ as follows:

$$F^\pi(i, \lambda) := P_{(i,\lambda)}^\pi\Big(\int_0^{\tau_B} r(x_t, \pi_t)\, \mathrm{d}t \leq \lambda\Big) \tag{9}$$

where $r(x_t, \pi_t)(e) := r(x_t(e), \pi(e, t))$ for all $e \in \Omega$ and $t \geq 0$, which measures the risk of the system that the total rewards incurred during a first passage time that the state process first enters a target set B and does not exceed the reward level λ when using policy π.

Definition 2.2. A policy $\pi^* \in \Pi$ is said to be risk probability optimal if

$$F^{\pi^*}(i, \lambda) = \inf_{\pi \in \Pi} F^\pi(i, \lambda) := F^*(i, \lambda) \tag{10}$$

for all $(i, \lambda) \in E \times R^+$. The function $F^*(i, \lambda)$ is called the value function.

Remark 2.3. By the definition of τ_B, we know that $\tau_B = 0$ for all initial state $i \in B$, and thus $F^*(i, \lambda) = F^\pi(i, \lambda) = 1$ for each $(i, \lambda) \in B \times R^+$ and $\pi \in \Pi$. Below, we limit our arguments to the case $(i, \lambda) \in B^c \times R^+$.

The main objective of this paper is to give some conditions for the existence of an optimal policy among the deterministic history-dependent policies, and to provide an algorithm for computing the optimal policy.

3. MAIN RESULTS

Notation: Let us denote by

$$\mathcal{G}_m := \{F : B^c \times R^+ \to [0, 1] |\ \ F(\cdot, \cdot) \text{ is Borel measurable function}\}.$$

For $(i,\lambda) \in B^c \times R^+, f \in \Pi_s$ and $a \in A(i)$, the operators H^f, H on $\mathcal{G}_m$ are given by

$$\begin{aligned} H^f F(i,\lambda) &:= \sum_{j\in B} \frac{q(j|i,f)}{q_i(f)}\Big(1 - e^{-q_i(f)\frac{\lambda}{r(i,f)}}\Big) \\ &\quad + \sum_{j\neq i, j\in B^c} \int_0^{\frac{\lambda}{r(i,f)}} F\Big(j, \lambda - r(i,f)u\Big) e^{-q_i(f)u} q(j|i,f)\,\mathrm{d}u, \qquad (11) \\ H^a F(i,\lambda) &:= \sum_{j\in B} \frac{q(j|i,a)}{q_i(a)}\Big(1 - e^{-q_i(a)\frac{\lambda}{r(i,a)}}\Big) \\ &\quad + \sum_{j\neq i, j\in B^c} \int_0^{\frac{\lambda}{r(i,a)}} F\Big(j, \lambda - r(i,a)u\Big) e^{-q_i(a)u} q(j|i,a)\,\mathrm{d}u, \qquad (12) \\ HF(i,\lambda) &:= \min_{a\in A(i)} H^a F(i,\lambda). \qquad (13) \end{aligned}$$

with $q_i(f) := -q(i|i, f(i,\lambda))$, $q(j|i,f) := q(j|i, f(i,\lambda))$.

Hence, the operators $(H^f)^n, H^n$ are defined by

$$(H^f)^1 F = H^f F, (H^f)^{n+1} F = H^f((H^f)^n F), H^1 F = HF, H^{n+1} F = H(H^n F), n \geq 1.$$

The operators have the following important properties.

Lemma 3.1. The following results hold.

(a) If $F, G \in \mathcal{G}_m$, and $F \geq G$, then $H^a F(i,\lambda) \geq H^a G(i,\lambda)$, $HF(i,\lambda) \geq HG(i,\lambda)$, for any $a \in A(i), (i,\lambda) \in B^c \times R^+$.

(b) If $F \in \mathcal{G}_m$, then $HF \in \mathcal{G}_m$, and there exists an $f \in \Pi_s$ such that $HF(i,\lambda) = H^f F(i,\lambda)$ for any $(i,\lambda) \in B^c \times R^+$.

P r o o f. (a) Part (a) follows directly from the definition of operator H.

(b) The finiteness of $A(i)$ for every $i \in B^c$ and the measurable selection theorem (proposition D.5 in [9]) imply that there exists an $f \in \Pi_s$ attaining the minimum in (13). □

To avoid the possibility of an infinite number of decision epochs during any finite horizon, we need the following basic assumption.

Assumption 3.2. For any $\pi \in \Pi, (i,\lambda) \in B^c \times R^+$, $P^\pi_{(i,\lambda)}(S_\infty = \infty) = 1$.

This assumption means that the states processes $\{x_t, t \geq 0\}$ is non-explosive. It follows from [4, 6, 7], we also give the following "drift condition" to verify Assumption 3.2.

Theorem 3.3. If there exist a measurable function $W \geq 1$ on E and some constants $c_0 > 0$, $b_0 \geq 0$, and $L_0 \geq 0$ such that

(a) $\sum_{j\in E} W(j) q(j \mid i,a) \leq c_0 W(i) + b_0$, for all $(i,a) \in K$; and

(b) $q^*(i) \le L_0 W(i)$ for all $i \in E$, with $q^*(i) = \sup_{a\in A(i)} q_i(a)$.

Then Assumption 3.2 holds.

P r o o f . It follows from Theorem 1 in [16]. □

Remark 3.4. (1) Theorem 3.3 is an extension of Condition 3.1 in [23] and Assumption 2.2 in [4] for CTMDPs with the classical expected criterion. When the transition rates are uniformly bounded (i. e. $\sup_{i\in E} q^*(i) < \infty$), Theorem 3.3 is satisfied by taking $W \equiv 1$.

(2) Theorem 3.3 is also called Lyapunov condition, which is weaker than the well known regular condition for SMDPs in [10, 12, 13, 14]. This is because the regular condition for SMDPs means that $Q(\delta, E \mid i, a) \le 1-\varepsilon$, for all $(i, a) \in K$ (for some δ and $\epsilon > 0$), where $Q(\delta, E \mid i, a)$ is the semi-Markov kernel. But compared with our model, the regular condition becomes there exist some constants $\delta > 0$ and $\varepsilon > 0$ such that $1 - e^{-q_i(a)\delta} < 1 - \varepsilon$ for all $(i, a) \in K$. This implies that $e^{-q_i(a)\delta} > \varepsilon$ for all $(i, a) \in K$ and thus the transition rates $q(j \mid i, a)$ must be bounded. However, in this paper we deal with the case when the transition rates are unbounded.

Due to the non-explosion of the state processes $\{x_t, t \ge 0\}$, the nonnegativity of the reward rate and the continuity of probability measures, for each $(i, \lambda) \in B^c \times R^+$ and $\pi \in \Pi$, $F^\pi(i, \lambda)$ is rewritten as follows:

$$\begin{aligned} F^\pi(i,\lambda) &= P^\pi_{(i,\lambda)}\Big(\int_0^{\tau_B} r(x_t,\pi_t)\,\mathrm{d}t \le \lambda\Big) \\ &= P^\pi_{(i,\lambda)}\Big(\sum_{m=0}^{\infty}\int_{T_m}^{T_{m+1}} I_{\{\tau_B>t\}} r(x_t,\pi_t)\,\mathrm{d}t \le \lambda\Big) \\ &= P^\pi_{(i,\lambda)}\Big(\sum_{m=0}^{\infty}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k}\in B^c)\}\}} r(x_t,\pi_t)\,\mathrm{d}t \le \lambda\Big) \\ &= \lim_{n\to\infty} P^\pi_{(i,\lambda)}\Big(\sum_{m=0}^{n}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k}\in B^c)\}\}} r(x_t,\pi_t)\,\mathrm{d}t \le \lambda\Big). \end{aligned}$$

Thus, a sequence $\{F^\pi_n(i,\lambda), n = -1, 0, 1, \ldots\}$ is given by

$$F^\pi_n(i,\lambda) := P^\pi_{(i,\lambda)}\Big(\sum_{m=0}^{n}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k}\in B^c)\}\}} r(x_t,\pi_t))\,\mathrm{d}t \le \lambda\Big), \quad \text{with } F^\pi_{-1}(i,\lambda) := 1,$$

for all $(i, \lambda) \in B^c \times R^+$.

Obviously, $F^\pi_n(i,\lambda) \ge F^\pi_{n+1}(i,\lambda), n \ge -1$ and $\lim_{n\to\infty} F^\pi_n(i,\lambda) = F^\pi(i,\lambda)$.

The following Lemma is the foundation of our main results, we will use it to establish the optimal equation.

Lemma 3.5. Suppose that Assumption 3.2 is satisfied, the following statements hold for any $(i, \lambda) \in B^c \times R^+, n \ge -1, \pi \in \Pi$.

(a) $F_n^{\pi}(i,\lambda) \in \mathcal{G}_m$ and $F^{\pi}(i,\lambda) \in \mathcal{G}_m$.

(b) $F_{n+1}^{\pi}(i,\lambda) = H^{f_0} F_n^{{}^1\pi}(i,\lambda)$ and $F^{\pi}(i,\lambda) = H^{f_0} F^{{}^1\pi}(i,\lambda)$, where ${}^1\pi := (\hat{f}_0, \hat{f}_1, ...)$ being the 1-shift policy of π, and $\hat{f}_k(t_1, i_1, \lambda_1, \ldots, t_{k+1}, i_{k+1}, \lambda_{k+1}) := f_{k+1}(i, \lambda, t_1, i_1, \lambda_1, \ldots, t_{k+1}, i_{k+1}, \lambda_{k+1}), k = 0, 1, \ldots$.

In particular, for $f \in \Pi_s$, $F^f(i,\lambda) = H^f F^f(i,\lambda)$.

P r o o f. (*a*) For any $(i,\lambda) \in B^c \times R^+, \pi \in \Pi$, $F_{-1}^{\pi}(i,\lambda) = 1 \in \mathcal{G}_m$. Suppose the statement is true for $n = k \geq -1$, then

$$
\begin{aligned}
F_{k+1}^{\pi}(i,\lambda) &= P_{(i,\lambda)}^{\pi}\Big(\sum_{m=0}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)\,\mathrm{d}t \leq \lambda \Big) \\
&= E_{(i,\lambda)}^{\pi}[I_{\{\sum_{m=0}^{k+1}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)\,\mathrm{d}t \leq \lambda\}}] \\
&= E_{(i,\lambda)}^{\pi}[E_{(i,\lambda)}^{\pi}[I_{\{\int_0^{T_1} r(x_t,\pi_t)\,\mathrm{d}t + \sum_{m=1}^{k+1}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=1}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)\,\mathrm{d}t \leq \lambda\}} \\
&\quad |x_{T_1}, T_1, \Lambda_1]] \\
&= \sum_{j\neq i}\int_0^{+\infty} P_{(i,\lambda)}^{\pi}\Big(\int_0^u r(x_t,\pi_t)\,\mathrm{d}t + \sum_{m=1}^{k+1}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=1}^{m}\{x_{T_k}\in B^c\}\}} \\
&\quad r(x_t,\pi_t)dt \leq \lambda | x_{T_1} = j, T_1 = u, \Lambda_1 = [\lambda - r(i,f_0)u]^+\Big) \\
&\quad \times e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u \\
&= \sum_{j\neq i}\int_0^{+\infty} P_{(i,\lambda)}^{\pi}\Big(\sum_{m=1}^{k+1}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=1}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)dt \\
&\quad \leq \lambda - r(i,f_0)u | x_{T_1} = j, T_1 = u, \Lambda_1 = [\lambda - r(i,f_0)u]^+\Big) \\
&\quad \times e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u \\
&= \sum_{j\neq i}\int_0^{+\infty} I_{\{\lambda - r(i,f_0)u \geq 0\}} P_{(j,\lambda - r(i,f_0)u)}^{{}^1\pi}\Big(\sum_{m=0}^{k}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k}\in B^c\}\}} \\
&\quad r(x_t,\pi_t)dt \leq \lambda - r(i,f_0)u\Big) e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u \\
&= \sum_{j\neq i, j\in B}\int_0^{+\infty} I_{\{\lambda - r(i,f_0)u \geq 0\}} e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u \\
&\quad + \sum_{j\neq i, j\in B^c}\int_0^{+\infty} I_{\{\lambda - r(i,f_0)u \geq 0\}} P_{(j,\lambda - r(i,f_0)u)}^{{}^1\pi}\Big(\sum_{m=0}^{k}\int_{T_m}^{T_{m+1}} \\
&\quad \times I_{\{\bigcap_{k=0}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)dt \leq \lambda - r(i,f_0)u\Big) e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u \\
&= \sum_{j\in B}\int_0^{\frac{\lambda}{r(i,f_0)}} e^{-q_i(f_0)u} q(j|i,f_0)\,\mathrm{d}u
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{j\neq i,j\in B^c}\int_0^{\frac{\lambda}{r(i,f_0)}} P^{1\pi}_{(j,\lambda-r(i,f_0)u)}\Big(\sum_{m=0}^{k}\int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k}\in B^c\}\}} r(x_t,\pi_t)\,\mathrm{d}t\\
&\le \lambda - r(i,f_0)u\Big)e^{-q_i(f_0)u}q(j|i,f_0)du\\
=\;& \sum_{j\in B}\frac{q(j|i,f_0)}{q_i(f_0)}\Big(1-e^{-q_i(f_0)\frac{\lambda}{r(i,f_0)}}\Big)\\
&+\sum_{j\neq i,j\in B^c}\int_0^{\frac{\lambda}{r(i,f_0)}} F_k^{1\pi}\Big(j,\lambda-r(i,f_0)u\Big)e^{-q_i(f_0)u}q(j|i,f_0)\,\mathrm{d}u\\
:=\;& H^{f_0}F_k^{1\pi}(i,\lambda)\in\mathcal{G}_m,
\end{aligned}
$$

where the third equality is due to the property of conditional expectation, the fourth equality follows from (7). Thus, using induction, we have $F_n^{\pi}(i,\cdot)\in\mathcal{G}_m$. Hence, the limit of a sequence of measurable functions is still measurable implying that $\lim_{n\to\infty}F_n^{\pi}(i,\lambda)=F^{\pi}(i,\lambda)\in\mathcal{G}_m$.

(*b*) By part (a), for any $(i,\lambda)\in B^c\times R^+, n\ge -1$, we know that

$$F_{n+1}^{\pi}(i,\lambda)=H^{f_0}F_n^{1\pi}(i,\lambda),$$

which together with the dominated convergence theorem gives $F^{\pi}(i,\lambda)=H^{f_0}F^{1\pi}(i,\lambda)$. Moreover, for $\pi=f\in\Pi_s$, $F^f(i,\lambda)=H^fF^f(i,\lambda)$. □

To show the existence and uniqueness of the solution to the equation $F^*=HF^*$, we need the following assumption.

Assumption 3.6. For any $(i,\lambda)\in B^c\times R^+, f\in\Pi_s, P^f_{(i,\lambda)}(\tau_B<+\infty)=1$.

To explain the meaning of Assumption 3.6, we need to introduce the following notation. For any given $f\in\Pi_s$, set $\widetilde{X}_n:=x_{T_n}, n=0,1,\ldots$, where T_n and $\{x_t,t\ge 0\}$ are the same as in Section 2. Thus, we obtain a discrete-time embedded chain $\{\widetilde{X}_n,n\ge 0\}$.

Remark 3.7. (1) Assumption 3.6 indicates that, no matter what the initial state is, what the reward level is, and what the policy is, the system states $\{x_t,t\ge 0\}$ will eventually arrive at B within finite time.

(2) Assumption 3.6 is equivalent to the following assertion. For every $(i,\lambda)\in B^c\times R$,

$$P^f_{(i,\lambda)}(\bigcup_{n=1}^{\infty}\{\widetilde{X}_n\in B\})=1 \;\; or \;\; P^f_{(i,\lambda)}(\bigcap_{n=1}^{\infty}\{\widetilde{X}_n\in B^c\})=0.$$

The proof of this assertion is shown as follows. For $(i,\lambda)\in B^c\times R, f\in\Pi_s$,

$$
\begin{aligned}
P^f_{(i,\lambda)}(\tau_B<+\infty) &= \sum_{n=1}^{\infty}P^f_{(i,\lambda)}(\widetilde{X}_k\in B^c, 1\le k\le n-1,\widetilde{X}_n\in B)\\
&= P^f_{(i,\lambda)}\Big(\bigcup_{n=1}^{\infty}\{\widetilde{X}_n\in B\}\Big)\\
&= 1. \qquad (14)
\end{aligned}
$$

To verify Assumption 3.6, it is desired to give a sufficient condition imposed on the data of our control model.

Proposition 3.8. If $\inf_{(i,a)\in B^c\times A(i)}\sum_{j\in B}\frac{q(j|i,a)}{q_i(a)} > 0$, Then, Assumption 3.6 is satisfied.

P r o o f. By Proposition 1 in [16], we obtain that this Proposition is true. □

Lemma 3.9. Suppose that Assumptions 3.2 and 3.6 hold.

(a) For any $(i,\lambda) \in B^c \times R^+, F, F^{'} \in \mathcal{G}_m, f \in \Pi_s$, if $F(i,\lambda) - F^{'}(i,\lambda) \leq H^f(F - F^{'})(i,\lambda)$, then $F(i,\lambda) \leq F^{'}(i,\lambda)$.

(b) For any $(i,\lambda) \in B^c \times R^+, f \in \Pi_s$, $F^f(i,\lambda)$ is the unique solution in $\mathcal{G}_m$ to the equation $F(i,\lambda) = H^f F(i,\lambda)$.

P r o o f. (a) For any $(i,\lambda) \in B^c \times R^+$, we first establish by induction that

$$(H^f)^n(F - F^{'})(i,\lambda) \leq P^f_{(i,\lambda)}(\bigcap_{k=1}^{n}\{\widetilde{X}_k \in B^c\}), n \geq 1. \tag{15}$$

Since $F(i,\lambda) - F^{'}(i,\lambda) \leq 1$, then,

$$\begin{aligned} H^f(F - F^{'})(i,\lambda) &= H^f F(i,\lambda) - H^f F^{'}(i,\lambda) \\ &= \sum_{j\neq i, j\in B^c}\int_0^{\frac{\lambda}{r(i,f)}}(F - F^{'})(j,\lambda - r(i,f)u) \\ &\quad \times e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u \\ &\leq \sum_{j\neq i, j\in B^c}\int_0^{+\infty} e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u \\ &= P^f_{(i,\lambda)}(\widetilde{X}_1 \in B^c). \end{aligned}$$

So the fact holds for $n = 1$. Assume the fact (15) is valid for $n = k$. Then,

$$\begin{aligned} (H^f)^{k+1}(F - F^{'})(i,\lambda) &= H^f(H^f)^k(F - F^{'})(i,\lambda) \\ &= \sum_{j\neq i, j\in B^c}\int_0^{\frac{\lambda}{r(i,f)}}(H^f)^k(F - F^{'})(j,\lambda - r(i,f)u) \\ &\quad \times e^{q_i(f)u}q(j|i,f)\,\mathrm{d}u \\ &\leq \sum_{j\neq i, j\in B^c}\int_0^{\frac{\lambda}{r(i,f)}} P^f_{(j,\lambda - r(i,f)u)}(\bigcap_{l=1}^{k}\{X_l \in B^c\}) \\ &\quad \times e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u. \end{aligned} \tag{16}$$

The inequality in (16) follows from the induction hypothesis. On the other hand,

$$\begin{aligned}
& P^f_{(i,\lambda)}(\bigcap_{l=1}^{k+1}\{\widetilde{X}_l \in B^c\}) \\
= \;& E^f_{(i,\lambda)}[I_{\{\bigcap_{l=1}^{k+1}\{\widetilde{X}_l\in B^c\}\}}] \\
= \;& E^f_{(i,\lambda)}[E^f_{(i,\lambda)}[I_{\bigcap_{l=1}^{k+1}\{\widetilde{X}_l\in B^c\}}|\widetilde{X}_1, T_1, \Lambda_1]] \\
= \;& \sum_{j\neq i}\int_0^{+\infty} P^f_{(i,\lambda)}\Big(\bigcap_{l=1}^{k+1}\{\widetilde{X}_l \in B^c\}|\widetilde{X}_1 = j, T_1 = u, \\
& \Lambda_1 = [\lambda - r(i,f)u]^+\Big)e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u \\
= \;& \sum_{j\neq i}\int_0^{\frac{\lambda}{r(i,f)}} P^f_{(i,\lambda)}\Big(\bigcap_{l=2}^{k+1}\{\widetilde{X}_l \in B^c\}, j \in B^c|\widetilde{X}_1 = j, T_1 = u, \\
& \Lambda_1 = \lambda - r(i,f)u\Big) \times e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u \\
= \;& \sum_{j\neq i, j\in B^c}\int_0^{\frac{\lambda}{r(i,f)}} P^f_{(j,\lambda-r(i,f)u)}\Big(\bigcap_{l=1}^{k}\{\widetilde{X}_l \in B^c\}\Big) \\
& \times e^{-q_i(f)u}q(j|i,f)\,\mathrm{d}u,
\end{aligned}$$

which together with (16) gives that $(H^f)^{k+1}(F - F^{'})(i,\lambda) \leq P^f_{(i,\lambda)}(\bigcap_{l=1}^{k+1}\{\widetilde{X}_l \in B^c\})$. Hence, by $F(i,\lambda) - F^{'}(i,\lambda) \leq H^f(F(i,\lambda) - F^{'}(i,\lambda))$, the fact (15) and induction , we obtain for all $n \geq 1$,

$$F(i,\lambda) - F^{'}(i,\lambda) \leq (H^f)^n(F(i,\lambda) - G(i,\lambda)) \leq P^f_{(i,\lambda)}\left(\bigcap_{k=1}^{n}\{\widetilde{X}_k \in B^c\}\right). \tag{17}$$

Letting $n \to \infty$ in (17), from Assumption 3.6 and (14), we have

$$F(i,\lambda) - F^{'}(i,\lambda) \leq \lim_{n\to\infty} P^f_{(i,\lambda)}\left(\bigcap_{k=1}^{n}\{\widetilde{X}_k \in B^c\}\right) = 1 - P^f_{(i,\lambda)}(\bigcup_{n=1}^{\infty}\{\widetilde{X}_k \in B\}) = 0,$$

which implies $F(i,\lambda) \leq F^{'}(i,\lambda)$.

(b) For any $(i,\lambda) \in B^c \times R^+$, by Lemma 3.1 (b), we know that $F^f(i,\lambda)(F^f(i,\lambda) \in \mathcal{G}_m)$ satisfies the equation $F(i,\lambda) = H^f F(i,\lambda)$. On the other hand, let $F^{'}(i,\lambda)$ be another solution in $\mathcal{G}_m$ to the equation $F(i,\lambda) = H^f F(i,\lambda)$ on $B^c \times R^+$, and thus $F^{'}(i,\lambda) - F^f(i,\lambda) = H^f(F^{'}(i,\lambda) - F^f(i,\lambda))$. Employing part (a), we have $F^{'}(i,\lambda) = F^f(i,\lambda)$. Then, the uniqueness of this problem is proved. □

Theorem 3.10. Under Assumptions 3.2 and 3.6, for any $(i,\lambda) \in B^c \times R^+$,

(a) set $F^*_{-1}(i,\lambda) := 1, F^*_{n+1}(i,\lambda) := HF^*_n(i,\lambda), n \geq -1$. Then, $\lim_{n\to\infty} F^*_n(i,\lambda) = F^*(i,\lambda)$.

(b) $F^*(i,\lambda)$ is the unique solution to the equation $F(i,\lambda) = HF(i,\lambda)$.

(c) There exists an $f^* \in \Pi_s$ such that $F^*(i,\lambda) = H^{f^*}F^*(i,\lambda)$ and $F^*(i,\lambda) = F^{f^*}(i,\lambda)$.

(d) Set $\tilde{f}_0(i,\lambda) := f^*(i,\lambda)$, and for $(i,\lambda,\, t_1, i_1, \lambda_1, \ldots, t_k, i_k, \lambda_k) \in H_k, k \geq 1$,

$$\tilde{f}_k^*(i,\lambda,t_1,i_1,\lambda_1,\ldots,t_k,i_k,\lambda_k) := f^*(i_k,\hat{\lambda}_k),$$

with $\hat{\lambda}_k = L(i_{k-1},\hat{\lambda}_{k-1}, f^*(i_{k-1},\hat{\lambda}_{k-1}),\theta_k), i_0 = i, \hat{\lambda}_0 = \lambda$, $\theta_k = s_k - s_{k-1}$ and L is given in (2). Then, the policy $\pi^* := (\tilde{f}_0^*, \tilde{f}_1^*, \ldots, \tilde{f}_k^*)$ is optimal.

P r o o f. (a) It follows from Lemma 3.1(a) that $F_n^*(i,\lambda) \geq F_{n+1}^*(i,\lambda), n \geq -1$ for any $(i,\lambda) \in B^c \times R^+$. Hence, by $0 \leq F_n^*(i,\lambda) \leq 1$, we get $\lim_{n\to\infty} F_n^*(i,\lambda) := \tilde{F}(i,\lambda)$ exists.

To prove $\tilde{F}(i,\lambda) \leq F^*(i,\lambda)$, we need to prove by induction that $F_n^*(i,\lambda) \leq F_n^\pi(i,\lambda)$, for all $\pi \in \Pi$ and $n \geq -1$. Since $F_{-1}^*(i,\lambda) = F_{-1}^\pi(i,\lambda) := 1$, it is obviously true for $n = -1$. Suppose that $F_k^*(i,\lambda) \leq F_k^\pi(i,\lambda)$ for all $\pi \in \Pi$ holds. Then,

$$F_{k+1}^*(i,\lambda) = HF_k^*(i,\lambda) \leq HF_k^{{}^1\pi}(i,\lambda) \leq H^{f^0}F_k^{{}^1\pi}(i,\lambda) = F_{k+1}^\pi(i,\lambda),$$

where the first inequality is due to the induction hypothesis, and the last equality follows from Lemma 3.5(b). Employing the induction, we have

$$F_n^*(i,\lambda) \leq F_n^\pi(i,\lambda) \tag{18}$$

for all $\pi \in \Pi$ and $n \geq -1$. Letting $n \to \infty$ in (18), we get $\tilde{F}(i,\lambda) = \lim_{n\to\infty} F_n^*(i,\lambda) \leq F^\pi(i,\lambda)$, for all $\pi \in \Pi$. The arbitrariness of π yields $\tilde{F}(i,\lambda) \leq F^*(i,\lambda)$.

Then, to show the converse i. e. $\tilde{F}(i,\lambda) \geq F^*(i,\lambda)$, it is suffices to show that there exists a policy $\theta \in \Pi_m$ such that $F_k^*(i,\lambda) = F_k^\theta(i,\lambda)$ for any $(i,\lambda) \in B^c \times R^+$. It is clear that $F_{-1}^*(i,\lambda) = 1 = F_{-1}^\pi(i,\lambda)$ for any $\pi \in \Pi_m$. Suppose that the fact is true for $n = k \geq -1$. By Lemma 3.1(b), we know that there exists an $f \in \Pi_s$ such that $HF_k^*(i,\lambda) = H^f F_k^*(i,\lambda)$. Letting $\eta = \{f,\theta\} \in \Pi_m$, the induction hypothesis and Lemma 3.5(b) give that $F_{k+1}^*(i,\lambda) = HF_k^*(i,\lambda) = H^f F_k^*(i,\lambda) = H^f F_k^\theta(i,\lambda) = F_{k+1}^\eta(i,\lambda)$. Then, there exists a policy $\theta \in \Pi_m$ such that $F_n^*(i,\lambda) = F_n^\theta(i,\lambda)$. This implies that $F_n^*(i,\lambda) = F_n^\theta(i,\lambda) \geq F^\theta(i,\lambda) \geq F^*(i,\lambda)$, thus $\lim_{n\to\infty} F_n^*(i,\lambda) = \tilde{F}(i,\lambda) \geq F^*(i,\lambda)$. This completes the proof.

(b) For any $(i,\lambda) \in B^c \times R^+$, by Lemma 3.5(b), we know that

$$F^\pi(i,\lambda) = H^{f_0}F^{{}^1\pi}(i,\lambda) \geq H^{f_0}F^*(i,\lambda) \geq HF^*(i,\lambda), \forall \pi \in \Pi.$$

Taking the infimum over all policies π yields $F^*(i,\lambda) \geq HF^*(i,\lambda)$.

On the other hand, for any $a \in A(i)$, employing part (a), we obtain

$$F_{n+1}^*(i,\lambda) = HF_n^*(i,\lambda) \leq H^a F_n^*(i,\lambda),$$

which together with the dominated convergence theorem gives $F^*(i,\lambda) \leq H^a F^*(i,\lambda)$. The arbitrariness of $a \in A(i)$ gives that $F^*(i,\lambda) \leq HF^*(i,\lambda)$. Thus, $F^* = HF^*$.

For any $(i,\lambda) \in B^c \times R^+$, since $F^*(i,\lambda)$ satisfies the equation $F(i,\lambda) = HF(i,\lambda)$, by Lemma 3.1(b), we know that there is an $f \in \Pi_s$ such that

$$F^*(i,\lambda) = H^f F^*(i,\lambda). \tag{19}$$

On the other hand, let $F^{'}(i,\lambda) \in \mathcal{G}_m$ be another solution to the equation $F(i,\lambda) = HF(i,\lambda)$. Similarly, by Lemma 3.1(b), we know that there exists an $f^{'} \in \Pi_s$ satisfying

$$F^{'}(i,\lambda) = H^{f^{'}} F^{'}(i,\lambda). \tag{20}$$

Comparing (19) with (20), we have

$$F^*(i,\lambda) = H^f F^*(i,\lambda) \le H^{f^{'}} F^*(i,\lambda)$$

and

$$F^{'}(i,\lambda) = H^{f^{'}} F^{'}(i,\lambda) \le H^f F^{'}(i,\lambda),$$

which imply $F^*(i,\lambda) - F^{'}(i,\lambda) \le H^{f^{'}}(F^{'} - F^*)(i,\lambda)$. Then, by Lemma 3.9(a), we obtain $F^*(i,\lambda) \le F^{'}(i,\lambda)$. Hence, reversing the role of $F^{'}$ and F^* gives $F^{'}(i,\lambda) = F^*(i,\lambda)$.

(c) For any $(i,\lambda) \in B^c \times R^+$, since $F^*(i,\lambda)$ satisfies the equation $F(i,\lambda) = HF(i,\lambda)$, by Lemma 3.1(b), we know that there is an $f^* \in \Pi_s$ such that $F^*(i,\lambda) = HF^*(i,\lambda) = H^{f^*}F^*(i,\lambda)$. Moreover, from Lemma 3.5 and Lemma 3.9 (b), we know that $F^{f^*}(i,\lambda)$ is the unique solution to the equation $F(i,\lambda) = H^{f^*}F(i,\lambda)$, which together with part (b) gives $F^*(i,\lambda) = F^{f^*}(i,\lambda)$.

(d) Since $\tilde{f}_k(i,\lambda,t_1,i_1,\lambda_1,\ldots,t_k,i_k,\lambda_k) := f^*(i_k,\tilde{\lambda}_k), \pi^* := (\tilde{f}_0,\tilde{f}_1,\ldots,\tilde{f}_k)$ for $k \ge 0$, which together with (2),(4) and (7) give $P^{f^*}_{\nu,k} = P^{\pi^*}_{\nu,k}$ for all $k \ge 0$. This implies that $P^{f^*}_{\nu} = P^{\pi^*}_{\nu}$, $P^{f^*}_{\nu}(\int_0^{\tau_B} r(x_s,\pi^*_s)\,\mathrm{d}s \ne \int_0^{\tau_B} r(x_s,f^*_s)\,\mathrm{d}s) = 0$ and $F^{\pi^*}(i,\lambda) = F^{f^*}(i,\lambda) = F^*(i,\lambda)$. Thus, π^* is optimal. □

Theorem 3.10 provides an value iteration algorithm for finding the value function and optimal policies. The algorithm procedure includes the following three steps.

The value iteration algorithm:

Step 1: Set $F^*_{-1}(i,\lambda) := 1$, for $n = -1, (i,\lambda) \in E \times R^+$.

Step 2: For all $n \ge 0, a \in A(i), (i,\lambda) \in E \times R^+$, by Theorem 3.10$(a)$, $H^a F^*_n(i,\lambda)$ and $F^*_{n+1}(i,\lambda)$ are calculated as follows:

$$\begin{aligned} H^a F^*_n(i,\lambda) &= \sum_{j\in B} \frac{q(j|i,a)}{q_i(a)}\Big(1 - e^{-q_i(a)\frac{\lambda}{r(i,a)}}\Big) \\ &\quad + \sum_{j\ne i, j\in B^c} \int_0^{\frac{\lambda}{r(i,a)}} F^*_n\Big(j, \lambda - r(i,a)u\Big) e^{-q_i(a)u} q(j|i,a)\,\mathrm{d}u \qquad (21) \\ &\approx \sum_{j\in B} \frac{q(j|i,a)}{q_i(a)}\Big(1 - e^{-q_i(a)\frac{\lambda}{r(i,a)}}\Big) \end{aligned}$$

$$+\sum_{j\neq i, j\in B^c}\sum_{k=1}^{m-1}\frac{1}{2}\Big[F_n^*\Big(j,\lambda-r(i,a)kh\Big)e^{-q_i(a)kh} \\ +F_n^*\Big(j,\lambda-r(i,a)(k+1)h\Big)e^{-q_i(a)(k+1)h}\Big]q(j|i,a)h. \tag{22}$$

$$F_{n+1}^*(i,\lambda) \approx \min_{a\in A(i)}\{H^a F_n^*(i,\lambda)\}, \tag{23}$$

where h is the step length, $k\leq m, k, m\in\mathcal{N}$ such that $mh=\frac{\lambda}{r(i,a)}$, $\mathcal{N}$ denotes the set of natural numbers.

Step 3: If $|F_{n+1}^*(i,\lambda,t)-F_n^*(i,\lambda,t)|<10^{-12}$, the iteration stops, and the value F_{n+1}^* is usually approximated as F^*. Hence, by Lemma 3.1, we know that there exists a policy f^* such that $HF^*=H^{f^*}F^*$, which together with theorem 3.10 implies that π^* is an optimal policy. Otherwise, increase n by 1 and go to step 2.

Remark 3.11. Using the trapezoidal integration method in [17], the formula (22) can be written as follows:

$$\int_a^b g(x)\,\mathrm{d}x\approx\sum_{k=0}^{m-1}\frac{g(a+kh)+g(a+(k+1)h)}{2}h, \tag{24}$$

where the step length h satisfies $a+mh=b, m\in\mathcal{N}$, [a,b] is the integration interval.

4. EXAMPLE

In this section, two examples are given to illustrate our results. We illustrate the verification of our conditions with the first example, which is a controlled queueing system. The second example for the numerical calculations of the value function and an optimal policy by value iteration techniques.

Example 4.1. (Optimal control of a queueing system) Consider a queueing system in which the state variable denotes the total number of waiting in the queue at time $t\geq 0$. There are natural arrival and service rates denoted by nonnegative constants α and β, respectively. There are two additional parameters h_1 and h_2, which are assumed to be controlled by the decision maker. When the system state is $i\in B^c:=\{1,2,\ldots\}\subseteq E=\{0,1,2,\ldots\}$, the decision maker takes an action a from a finite set $A(i)$ of available actions, which may admit $(h_k(i,a)\geq 0, k=0,1.)$ or reject $(h_k(i,a)\leq 0, k=0,1.)$ arriving jobs, and increase $(h_2(i,a)\geq 0)$ or decrease $(h_2(i,a)\leq 0)$ the service rate. This action results in a reward rate $r(i,a)\geq 0$. Moreover, we assume that some emergency situations reduce the number of waiting in the queue directly to 0 with a rate being $\alpha_0>0$.

To formulate this control problem as a CTMDP, we introduce the transition rates $q(j|i,a)$ as follows:

For $i=0$ and $a\in A(0)$,

$$r(0,a)=0, q(0|0,a)=q(j|0,a)=0 \quad for \quad j\geq 1. \tag{25}$$

For $i = 1$ and $a \in A(1)$,

$$q(j|1,a) = \begin{cases} \alpha_0 + h_0(1,a), & \text{if } j = 0, \\ -(\alpha_0 + \beta) - h_0(1,a) - h_2(1,a), & \text{if } j = 1, \\ \beta + h_2(1,a), & \text{if } j = 2, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

For $i \geq 2$ and $a \in A(i)$

$$q(j|i,a) = \begin{cases} \alpha_0 i + h_0(i,a), & \text{if } j = 0, \\ \alpha i + h_1(i,a), & \text{if } j = i-1, \\ -(\alpha_0 + \alpha + \beta)i - h_0(i,a) - h_1(i,a) - h_2(i,a), & \text{if } j = i, \\ \beta i + h_2(i,a), & \text{if } j = i+1, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

The aim in this example is to ensure the existence of a risk probability optimal policy. To do so, we assume that the following conditions:

B1. $\alpha_0 i + h_0(i,a) \geq 0, \alpha i + h_1(i,a) \geq 0$ and $\beta i + h_2(i,a) \geq 0$ for all $a \in A(i)$ and $i \geq 1$;

B2. $\|h_k\| := \sup_{(i,a)\in K} |h_k(i,a)| < \infty$ for $k = 0, 1, 2$.

Under these conditions, we obtain the following fact.

Proposition 4.2. Under **B1** and **B2**, the above queueing system satisfies Assumptions 3.2 and 3.6. Then, by Theorem 3.10, there exists an optimal policy.

Proof. First, we will verify Assumption 3.2. Set $W(i) := i$ for all $i \in E$, $L_0 := \alpha_0 + \alpha + \beta + \|h_0\| + \|h_1\| + \|h_2\|$. For any $i \in E$, by **B1**, **B2**, (25), (26) and (27), we have

$$q^*(i) = \sup_{a\in A(i)} q_i(a) \leq L_0 W(i). \tag{28}$$

This implies Assumption 3.2(b) holds.

On the other hand, for $i = 0$, $a \in A(0)$, from (25), we obtain

$$\sum_{j\in E} W(j)q(j|0,a) = 0 \leq (\alpha_0 + \alpha + \beta)V(0) + L_0. \tag{29}$$

For $i \geq 1$ and $a \in A(i)$, employing (27), we have

$$\sum_{j\in E} W(j)q(j|i,a) \leq (\alpha_0 + \alpha + \beta)V(i) + L_0, \tag{30}$$

which together with (29) gives that Assumption 3.2(a) is verified with $c_0 := (\alpha_0 + \alpha + \beta), b_0 := L_0$. Thus, Assumption 3.2 holds.

Hence, by (25), we know that $i = 0$ is a single absorbing state. Moreover, for $i \geq 1, a \in A(i)$, using (27) and (28), we have

$$\inf_{(i,a)\in B^c\times A(i)} \sum_{j\in B} \frac{q(j|i,a)}{q_i(a)} = \inf_{a\in A(0)} \frac{q(0|i,a)}{q_i(a)} \geq \inf_{a\in A(0)} \frac{\alpha_0 i + h_0(i,a)}{L_0 i} > 0,$$

which shows that Proposition 3.8 is trivially true. Thus, Assumption 3.6 holds. □

Example 4.3. (A management problem in an insurance company) Consider a car insurance company classifies its profit situation into three states 0, 1 and 2, which denote the bankruptcy, the medium and the profit state, respectively. The state 0 means that the company went bankrupt and had no income, i. e. $r(0, a_{01}) = 0, a_{01} \in A(0)$. In state 1, the decision maker may take an insurance policy a_{11} or take another insurance policy a_{12}, which leading in a reward rate $r(1, a_{11}) \geq 0$ or a reward rate $r(1, a_{12}) \geq 0$, respectively. In state 2, the decision maker can also choose an new insurance policy a_{21} resulting in a reward rate $r(2, a_{21}) \geq 0$ or choose another new insurance policy a_{22} to result a reward rate $r(2, a_{22}) \geq 0$. The evolution of of this system is described as follows. For each $i \in \{1, 2\}$, when the action $a \in A(i)$ is selected, the system stays at i with a random time satisfying the exponential distribution with the parameter $q_i(a)$, where $a \in A(i), q_i(a) \neq 0$. At this new decision epoch, the system state changes into a new state $j(j \neq i, j = 0, 1, 2.)$ with the transition probability $P(j|i, a) = \frac{q(j|i,a)}{q_i(a)}, a \in A(i)$.

We formulate this control system as a CTMDP, some parameters are given as follows. The state space $E = \{0, 1, 2\}$, the target set $B = \{0\}$; the action sets $A(1) = \{a_{11}, a_{12}\}, A(2) = \{a_{21}, a_{22}\}, A(0) = \{a_{01}\}$. The transition rates are given as follows:

$$\begin{array}{lll}
q(0|0, a_{01}) = 0, & q(1|0, a_{01}) = 0, & q(2|0, a_{01}) = 0, \\
q(0|1, a_{11}) = 0.138, & q(1|1, a_{11}) = -0.46, & q(2|1, a_{11}) = 0.322, \\
q(0|1, a_{12}) = 0.024, & q(1|1, a_{12}) = -0.06, & q(2|1, a_{12}) = 0.036, \\
q(0|2, a_{21}) = 0.036, & q(1|2, a_{21}) = 0.084, & q(2|2, a_{21}) = -0.12, \\
q(0|2, a_{22}) = 0.005, & q(1|2, a_{22}) = 0.045, & q(2|2, a_{22}) = -0.05,
\end{array} \tag{31}$$

and the reward rates are given by

$$r(0, a_{01}) = 0, \quad r(1, a_{11}) = 0.6, \quad r(1, a_{12}) = 0.5, \quad r(2, a_{21}) = 0.1, \quad r(2, a_{22}) = 0.2.$$

Employing (31), we have (i) the transition rates are uniformly bounded; (ii) The state 0 is absorbing, and Proposition 3.8 is trivially true. Then, Assumptions 3.2 and 3.6 hold. This imply the first passage risk probability optimal policy exists. Hence, using Theorem 3.10, the value function $F^*(1, \lambda)$ and $F^*(2, \lambda)$ are calculated by the value iteration algorithm as follows.

Step 1: For $\lambda \in [0, +\infty)$ and $i = 1, 2$, set $F^*_{-1}(i, \lambda) := 1$,

Step 2: For $i = 1$,

$$H^{a_{11}} F^*_n(1, \lambda) \quad = \quad 0.3 \times (1 - e^{-\frac{23\lambda}{30}}) + 0.7 \times 0.46 \times \int_0^{\frac{\lambda}{0.6}} F^*_n(2, \lambda - 0.6u) e^{-0.46t} \, \mathrm{d}t,$$

$$H^{a_{12}}F_n^*(1,\lambda) = 0.4\times(1-e^{-\frac{3\lambda}{25}})+0.6\times 0.06\times\int_0^{\frac{\lambda}{0.5}} F_n^*(2,\lambda-0.5u)e^{-0.06t}\,\mathrm{d}t,$$
$$F_{n+1}^*(1,\lambda) = \min\{H^{a_{11}}F_n^*(1,\lambda), H^{a_{12}}F_n^*(1,\lambda)\}.$$

For $i=2$,

$$H^{a_{21}}F_n^*(2,\lambda) = 0.3\times(1-e^{-1.2\lambda})+0.7\times 0.12\times\int_0^{\frac{\lambda}{0.1}} F_n^*(1,\lambda-0.1u)e^{-0.12t}\,\mathrm{d}t,$$
$$H^{a_{22}}F_n^*(1,\lambda) = 0.1\times(1-e^{-\frac{\lambda}{4}})+0.9\times 0.05\times\int_0^{\frac{\lambda}{0.2}} F_n^*(1,\lambda-0.2u)e^{-0.05t}\,\mathrm{d}t,$$
$$F_{n+1}^*(2,\lambda) = \min\{H^{a_{21}}F_n^*(2,\lambda), H^{a_{22}}F_n^*(2,\lambda)\}.$$

Step 3: if $|F_{n+1}^*(i,\lambda)-F_n^*(i,\lambda)|<10^{-12}$, go to step 4, the value F_{n+1}^* is usually approximated as F^*; otherwise, increase n by 1 and go to step 2.

Step 4: Plot out the graphs of these functions $H^{a_{11}}F_n^*(1,\lambda), H^{a_{12}}F_n^*(1,\lambda)$, $H^{a_{21}}F_n^*(2,\lambda)$, $H^{a_{22}}F_n^*(2,\lambda), F^*(1,\lambda)$ and $F^*(2,\lambda)$, see Figure 1 and Figure 2.

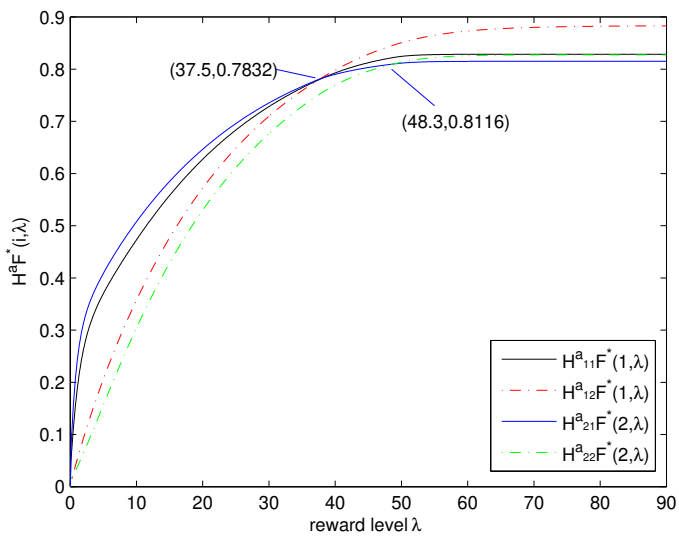

Fig. 1. The function $H^aF^*(i,\lambda)$.

From Figures 1–2 and the computational procedure, we have the following conclusions.

(a) From Figure 1, we see that in state 1, $H^{a_{12}}F^*(1,\lambda)$ is below $H^{a_{11}}F^*(1,\lambda)$ when $\lambda\in(0,37.5)$, and $H^{a_{11}}F^*(1,\lambda)$ is below $H^{a_{12}}F^*(1,\lambda)$ when $\lambda\in[37.5,90]$. This suggests that the decision maker should take the action a_{12} with lower risk rather than the action a_{11} when $\lambda\in(0,37.5)$, or take the action a_{11} with lower risk rather than the action a_{12}

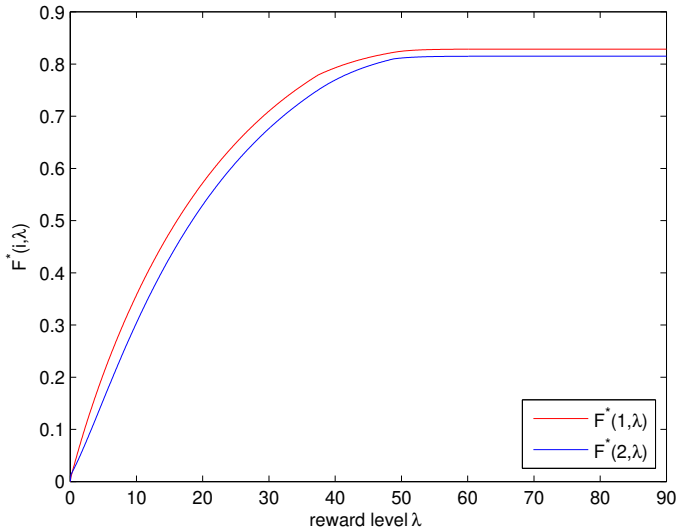

Fig. 2. The value function $F^*(i,\lambda)$.

when $\lambda \in [37.5, 90]$. Similarly, in state 2, the action a_{22} is with lower risk than the action a_{21} when $\lambda \in (0, 48.3)$, but the action a_{21} is with lower risk than the action a_{22} when $\lambda \in [48.3, 90]$.

(b) From Figures 1 – 2, we know that the optimal actions depend on the critical points $\lambda^*(i)$, and the optimal actions are given as follows:

$$f^*(1,\lambda) = \begin{cases} a_{12}, & 0 \le \lambda < 37.5; \\ a_{11}, & 37.5 \le \lambda \le 90. \end{cases}, \quad f^*(2,\lambda) = \begin{cases} a_{22}, & 0 \le \lambda < 48.3; \\ a_{21}, & 48.3 \le \lambda \le 90. \end{cases}$$

with $F^*(i,\lambda) = H^{f^*} F^*(i,\lambda)$.

This implies at time $t_0 = 0$, according to the system state i_0 and the initial reward level λ_0, the decision marker chooses an action $\tilde{f}_0^*(i_0,\lambda_0) := f^*(i_0,\lambda_0) \in A(i_0)$. Consequently, the system stays at i_0 until time t_1. At this point, the system goes into a new state i_1, and the decision marker gets a reward $r(i_0, \tilde{f}_0^*(i_0,\lambda_0))t_1$, and have a new profit goal $\hat{\lambda}_1 = L(i_0,\lambda_0,\tilde{f}_0^*(i_0,\lambda_0),\theta_1)$ for the decision marker. Then the next decision epoch comes, based on the current state i_1 and reward level $\hat{\lambda}_1$, the decision maker takes an action $\tilde{f}_1^*(i_0,\lambda_0,t_1,i_1,\lambda_1) := f^*(i_1,\hat{\lambda}_1) \in A(i_1)$. The decision maker takes action repeatedly in this way until the system state falls into the target set $B := \{0\}$. Then, by Theorem 3.10, we know that the policy $\pi^* = (\tilde{f}_0^*, \tilde{f}_1^*, \ldots)$ is optimal.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61773411; 11461008); The Basic Ability Improvement Project of Young and Middle aged Teachers in

Guangxi Institution of Higher Education (Grant No. KY2016YB844,KY2019YB0369). PhD research startup foundation of Guangxi University of Science and Technology(Grant No.18Z06). The authors also thank the Associate Editor and the referee for many valuable comments and suggestions which have improved this paper.

(Received January 15, 2018)

REFERENCES

[1] D. Bertsekas and S.Shreve: Stochastic Optimal Control: The Discrete-Time Case. Academic Press Inc 1996.

[2] N. Bauerle and U. Rieder: Markov Decision Processes with Applications to Finance. Springer, Heidelberg 2011. DOI:10.1007/978-3-642-18324-9

[3] E. Feinberg: Continuous time discounted jump Markov decision processes: a discrete-event approach. Math. Operat. Res. *29* (2004), 492–524. DOI:10.1287/moor.1040.0089

[4] X. P. Guo and O. Hernández-Lerma: Continuous-Time Markov Decision Process: Theorey and Applications. Springer-Verlag, Berlin 2009.

[5] X. P. Guo, A. Hernández-Del-Valle, and O. Hernández-Lerma: First passage problems for nonstationary discrete-time stochastic control systems. Europ. J. Control *18* (2012), 528–538. DOI:10.3166/ejc.18.528-538

[6] X. P. Guo, X. Y. Song and Y. Zhang: First passage optimality for continuous time Markov decision processes with varying discount factors and history-dependent policies. IEEE Trans. Automat. Control *59* (2014), 163–174. DOI:10.1109/tac.2013.2281475

[7] X. P. Guo, X. X. Huang, and Y. H. Huang: Finite-horizon optimality for continuous-time Markov decision processs with unbounded transition rates. Adv. Appl. Prob. *47* (2015), 1064–1087. DOI:10.1017/s0001867800049016

[8] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Process: Basic Optimality Criteria. Springer-Verlag, New York 1996. DOI:10.1007/978-1-4612-0729-0

[9] O. Hernández-Lerma and J. B. Lasserre: Further Topics on Discrete-Time Markov Control Process. Springer-Verlag, New York 1999. DOI:10.1007/978-1-4612-0561-6

[10] Y. H. Huang and X. P. Guo: Optimal risk probability for first passage models in Semi-Markov processes. J. Math. Anal. Appl. *359* (2009), 404–420. DOI:10.1016/j.jmaa.2009.05.058

[11] Y. H. Huang and X. P. Guo: First passage models for denumberable Semi-Markov processes with nonnegative discounted cost. Acta. Math. Appl. Sinica *27* (2011), 177–190. DOI:10.1007/s10255-011-0061-2

[12] Y. H. Huang, Q. D. Wei, and X. P. Guo: Constrained Markov decision processes with first passage criteria. Ann. Oper. Res. *206* (2013), 197–219. DOI:10.1007/s10479-012-1292-1

[13] Y. H. Huang, X. P. Guo, and Z. F. Li: Minimum risk probability for finite horizon semi-Markov decision process. J. Math. Anal. Appl. *402* (2013), 378–391. DOI:10.1016/j.jmaa.2013.01.021

[14] X. X. Huang, X. L. Zou, and X. P. Guo: A minimization problem of the risk probability in first passage semi-Markov decision processes with loss rates. Sci. China Math. *58* (2015), 1923–1938. DOI:10.1007/s11425-015-5029-x

[15] X. X. Huang and Y. H. Huang: Mean-variance optimality for semi-Markov decision processes under first passage. Kybernetika *53* (2017), 59–81. DOI:10.14736/kyb-2017-1-0059

[16] H. F. Huo, X. L. Zou, and X. P. Guo: The risk probability criterion for discounted continuous-time Markov decision processes. Discrete Event Dynamic system: Theory Appl. *27* (2017), 675–699. DOI:10.1007/s10626-017-0257-6

[17] J. Janssen and R. Manca: Semi-Markov Risk Models For Finance, Insurance, and Reliability. Springer, New York 2006.

[18] Y. L. Lin, R. J. Tomkins, and C. L. Wang: Optimal models for the first arrival time distribution function in continuous time with a special case. Acta. Math. Appl. Sinica *10* (1994), 194–212. DOI:10.1007/bf02006119

[19] J. Y. Liu and K. Liu: Markov decision programming – the first passage model with denumerable state space. Systems Sci. Math. Sci. *5* (1992), 340–351.

[20] J. Y. Liu and S. M. Huang: Markov decision processes with distribution function criterion of first-passage time. Appl. Math. Optim. *43* (2001), 187–201. DOI:10.1007/s00245-001-0007-9

[21] Y. Ohtsubo: Optimal threshold probability in undiscounted Markov decision processes with a target set. Appl. Math. Anal. Comp. *149* (2004), 519–532. DOI:10.1016/s0096-3003(03)00158-9

[22] M. L. Puterman: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley, New York 1994.

[23] A. Piunovskiy and Y. Zhang: Discounted continuous-time Markov decision processes with unbounded rates: the convex analytic approach. SIAM J. Control Optim. *49* (2011), 2032–2061. DOI:10.1137/10081366x

[24] M. Schäl: Control of ruin probabilities by discrete-time investments. Math. Meth. Oper. Res. *70* (2005), 141–158. DOI:10.1007/s00186-005-0445-2

[25] C. B. Wu and Y. L. Lin: Minimizing risk models in Markov decision processes with policies depending on target values. J. Math. Anal. Appl. *231* (1999), 47–57. DOI:10.1006/jmaa.1998.6203

[26] X. Wu and X. P. Guo: First passage optimality and variance minimization of Markov decision processes with varying discount factors. J. Appl. Prob. *52* (2015), 441–456. DOI:10.1017/s0021900200012560

[27] S. X. Yu, Y. L. Lin, and P. F. Yan: Optimization models for the first arrival target distribution function in discrete time. J. Math. Anal. Appl. *225* (1998), 193–223. DOI:10.1006/jmaa.1998.6015

[28] X. L. Zou, and X. P. Guo: Another set of verifiable conditions for average Markov decision processes with Borel spaces. Kybernetika *51* (2015), 276–292. DOI:10.14736/kyb-2015-2-0276

Haifeng Huo, Corresponding author. School of Science, Guangxi University of Science and Technology, Liuzhou, 545006. P. R. China.
e-mail: xiaohuo08ok@163.com

Xian Wen, 1. School of Science, Guangxi University of Science and Technology, Liuzhou, 545006, P. R. China˙ 2. Lushan College of Guangxi University of Science and Technology, Liuzhou, 5450616. P. R. China.
e-mail: wenxian879@163.com