KYBERNETIKA — VOLUME *60* (2024), NUMBER 3, PAGES 357–378

MINIMIZING RISK PROBABILITY FOR INFINITE DISCOUNTED PIECEWISE DETERMINISTIC MARKOV DECISION PROCESSES

Haifeng Huo, Jinhua Cui and Xian Wen

The purpose of this paper is to study the risk probability problem for infinite horizon piecewise deterministic Markov decision processes (PDMDPs) with varying discount factors and unbounded transition rates. Different from the usual expected total rewards, we aim to minimize the risk probability that the total rewards do not exceed a given target value. Under the condition of the controlled state process being non-explosive is slightly weaker than the corresponding ones in the previous literature, we prove the existence and uniqueness of a solution to the optimality equation, and the existence of the risk probability optimal policy by using the value iteration algorithm. Finally, we provide two examples to illustrate our results, one of which explains and verifies our conditions and the other shows the computational results of the value function and the risk probability optimal policy.

Keywords: piecewise deterministic Markov decision processes, risk probability criterion, optimal policy, the value iteration algorithm

Classification: 90C40, 60E20

1. INTRODUCTION

Piecewise deterministic Markov decision processes (PDMDPs) are significant dynamic programming models that are widely used in many fields such as finance [5], communication networks [9, 15], neuro medicine [9, 15]. As is well known, the popularly used performance criteria in PDMDPs are the expected criteria, see, (i) The finite horizon expected criterion [5, 8, 9, 16]. Specifically, Davis [8, 9] considered the finite horizon optimality problem for the uncontrolled case with bounded cost functions and bounded jump rates, and established the relationship between the expected value and the optimality equation by exploiting an infinitesimal approach. Using the embedded chain technique, Bauerle and Rieder [5] transformed the optimal control problem of PDMDPs into discrete-time Markov decision processes (DTMDPs), and proved the existence of the optimal policy and the optimality equation under the compactness-continuity condition. Different from the literature [5], Huang and Guo [16] established the corresponding Hamilton-Jacobi-Bellman equation, and proved the existence of an optimal policy by using the infinitesimal approach. (ii)The expected average criterion [3, 10]. Costa et al. [3]

DOI: 10.14736/kyb-2024-3-0357

studied the expected average control problem and established the optimality equation by the discrete embedded chain technique. Dufour et al. [10] formulated the existence condition of an optimal average continuous control through the vanishing factor method. (iii) The discounted expected criterion [1, 4]. Almudevar [1] converted the infinite discounted optimization problem from PDMDPs to DTMDPs and exploited the dynamic programming method to compute the value function. Dufou and Costa [4] investigated the infinite discounted expectation problem by using the infinitesimal approach, and established the conditions for the existence and uniqueness to the solution of the optimality equation.

All existing works are focused on investigating the performance of PDMDPs in terms of the expected rewards. However, the expected criteria are risk-neutral and cannot effectively describe the control system's risk situation. Therefore, it is necessary to introduce the risk probability criterion, which measures the probability that the total rewards of the control system do not exceed a given value over a fixed period. The risk probability performance analysis plays a vital role in the field of risk control, which is widely applied to finance and insurance, such as ruin problems [5], reliability [17], and maintenance [21]. According to the characteristics of holding time distribution of the system states, the existing literature on the risk probability problems can be divided into three categories:

(i) discrete-time Markov decision processes (DTMDPs), where the sojourn time of a system state is a fixed constant; see, [23, 26, 27],

(ii) semi-Markov decision processes (SMDPs), where the sojourn time of a system state follows arbitrary probability distribution; see, [18, 17],

(iii) continuous-time Markov decision processes (CTMDPs), where the sojourn time of a system state follows an exponential distribution; see, [20, 19].

A common feature of all previous studies is that the system state remains unchanged between jumps. PDMDPs [5, 9, 15] are more generalized stochastic models in which the system states between jumps change according to a given flow function. On the other hand, in many economic and financial systems (i. e., uncertain interest rates), the discount factors are usually considered non-constant depending on the system states. These features motivate us to investigate the risk probability criterion for infinite discounted PDMDPs with the state-dependent factors.

In this paper, our goal is to establish the existence condition and the computational method of the risk probability optimal policies for infinite discounted PDMDPs. Different from considering only the system state in the expected case, to solve the risk probability optimality for PDMDPs, the reward levels need to be considered as a component of extended states. Thus, the existing results of the expected rewards for PDMDPs [3, 5, 8, 9, 10, 16] cannot be applied directly to our model. Since the choice of an action under any policy in PDMDPs depends on the reward (or cost) levels and past states as well as decision epochs, the history-dependent policies should be redefined with a k-component internal history, see Definition 2.1. However, the system states are determined by a given flow function and the transition rates, the theoretical results of the risk probability for DTMDPs [22, 23, 26], SMDPs [17, 18] and CTMDPs [19, 20] are not suitable for the model of PDMDPs, see Remark 2.3 and 3.1.

As a consequence, for any given redefined policy, initial system state and reward level, we need to expand the state space and reconstruct the probability space (7) by the well-known Ionescu Tulcea theorem. On account of the transition rate are unbounded, the state process may be explosive. Then, a generalized drift condition needs to be established to ensure that the state process is non-explosive, which is more general than those in [17, 18, 19, 20] for SMDPs and CTMDPs, see Lemma 3.4 and Remark 3.5. To assure the existence of the optimal policy of the risk probability based on the non-explosive condition, we first establish a new fact in Theorem 3.7. Secondly, we use the dynamic programming approach and the value iteration technique to solve the risk probability optimality problem, including establishing the optimality equation, and the existence and uniqueness of the solution for the optimality equation in Theorem 3.10, see Remark 3.11. Moreover, the value iteration algorithm is developed to approximate the risk probability optimal value, which together with Theorem 3.7 proves the existence of a risk probability optimal policy. Finally, we explain the main results through two examples, one of which is used to verify our conditions, and the other shows the effectiveness and feasibility of the value iteration algorithm to calculate the value function and the optimal policy.

Compared with the existing works [20, 25], there are three features in this paper: (i) Our model is a generalization of CTMDPs, where a given drift function determines the states between jumps, while the states between jumps remain unchanged in CTMDPs [20]. Moreover, the state-dependent discount factor in this paper is more suitable to the applications in the financial and engineering fields [13, 27], while only the constant discount factor is considered in [20], see Remark 2.3. (ii) We provided an iteration technique to establish the optimality equation, and prove that the value function is the unique solution to the corresponding optimality equation. Since our model is different from those in [20], we know that the optimality equation in Theorem 3.10 is different from those in [20], see Remark 3.1. (iii) Although our model is more general, our condition is weaker than that for CTMDPs with discounted risk probability [20]. To assure the existence of the optimal policy of the risk probability, we establish the new fact in Theorem 3.7 by only using the condition that the controlled state process is non-explosive. In [20, 25], the non-explosive and first arrival conditions are both used to assure the existence of optimal risk probability policies, as described in Remark 3.11.

The main structure of the present article comprises four parts. Section 2 describes the infinite discounted risk probability control model of PDMDPs. Section 3 introduces the solution to the optimality problem of the risk probability, including the existence and computation of both the value function and the optimal policy of the risk probability. Section 4 describes the use of a technique called value iteration to resolve the optimal risk probability investment problem.

2. THE CONTROL MODEL

The model of infinite discounted risk probability PDMDPs contains a six-tuple of the subsequent parts:

$$\{E, (A(x) \subseteq A, x \in E), q(dy|x,a), \phi(x,t), r(x,a), \alpha(x)\}, \tag{1}$$

(a) E represents a Borel state space endowed with a Borel σ-algebra $\mathcal{B}(E)$.

(b) A denotes a Borel action space endowed with a Borel σ-algebra $\mathcal{B}(A)$; $A(x) \in \mathcal{B}(A)$ denotes a set of actions that can be selected in state $x \in E$; $K := \{(x, a)|x \in E, a \in A(x)\}$ describes the admissible state-action pairs in the set.

(c) $q(\cdot|x, a)$ represents the transition rate,which is a signed kernel on $\mathcal{B}(E)$ given K satisfying $0 \leq q(D|x, a) \leq +\infty$ with $(x, a) \in K, x \notin D \in \mathcal{B}(E)$. Moreover, it is assumed that the transition rates are conservative (i. e., $q(E|x, a) = 0$) and stable (i. e., $q^*(x) := \sup_{a\in A(x)} q_x(a) < \infty$), where $q_x(a) := -q(\{x\}|x, a) \geq 0$ for all $(x, a) \in K$.

(d) $\phi(x, s)$ denotes a deterministic flow, which is a Borel-measurable function from $E \times R$ to E, and satisfies the following properties:

(i) for any $s, t \geq 0$ and $x \in E$,

$$\phi(x, t + s) = \phi(\phi(x, t), s); \tag{2}$$

(ii) for any $x \in E$, $\phi(x, \cdot)$ is continuous on R^+ where $R^+ := [0, +\infty)$.

In particular, for the case of $\phi(x, s) = x$ for any $s \geq 0$, our model reduces to the model of CTMDPs for [11, 12, 24].

(e) $r(x, a)$ denotes the reward function, called a nonnegative measurable function from K to R^+.

(f) $\alpha(x)$ denotes the discount factor, which is related to the state $x \in E$.

The risk probability discounted piecewise deterministic Markov decision process elaborates as the following: At the initial time $s_0 = 0$, the system state x_0 is observed by the decision maker. Meanwhile, there exists a reward level (goal) $\lambda_0 \in R^+$ for the decision-maker to attempt to control the total rewards of the system operation are not larger than the initial reward level λ_0. Based on the observation information (x_0, λ_0) of the system, the decision-maker selects the control action $a_0 \in A(x_0)$. Consequently, the system evolves in two ways: (i) The change of the system state is based on the flow $\phi(x_0, s)(s \in [s_0, s_1))$ up to the time s_1. Now, the system state jumps into a new state $x_1 \in E$, which is governed by transition rate $q(\mathrm{d}x_1|x_0, a_0)$. (ii) During the period $[s_0, s_1)$, the decision maker obtains the rewards $\int_0^{s_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0, s), a_0)\mathrm{d}s$. Then, a new decision-making moment s_1 arrives. Here, considering the influence of the varying discount factor, the remaining reward goal becomes $\hat{\lambda}_1 = e^{\int_0^{s_1} \alpha(\phi(x_0,t))\mathrm{d}t}(\lambda_0 - \int_0^{s_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0, s), a_0)\mathrm{d}s)$. In terms of the historical information $(x_0, \lambda_0, s_1, x_1, \hat{\lambda}_1)$, the decision maker selects a control action a_1 from the set $A(x_1)$. The evolution of the system state is repeated in a manner similar to (i) and (ii). At the nth decision time $s_n, n = 0, 1, \ldots$, the decision-maker observes a series of historical information $(x_0, \hat{\lambda}_0, \theta_1, x_1, \hat{\lambda}_1, \ldots, \theta_n, x_n, \hat{\lambda}_n)$ where $\hat{\lambda}_0 := \lambda_0$, x_{n+1} is the system's state after the jump time s_{n+1}; $\hat{\lambda}_n$ represents the reward goal, and it satisfies

$$\hat{\lambda}_{n+1} := e^{\int_0^{\theta_{n+1}} \alpha(\phi(x_n,t))\mathrm{d}t}(\hat{\lambda}_n - \int_0^{\theta_{n+1}} e^{-\int_0^s \alpha(\phi(x_n,t))\mathrm{d}t} r(\phi(x_n, s), a_n)\mathrm{d}s), \tag{3}$$

by considering the varying discount factor into consideration. $\theta_{n+1} := s_{n+1} - s_n$ denotes the time interval between two neighborly jumps, a_n denotes the action. Thus, the decision maker aims to minimize probability of the full rewards that cannot reach the reward goal, which is defined by (9) below.

Different from the classical expectation criterion, the choice of an action under any policy in risk probability PDMDPs depends on the additional reward (or cost) levels and past states as well as decision epochs. To ensure a reasonable model, we first establish the probability space. Now, the measurable space $(\Omega, \mathscr{F})$ is established as follows: the sample space $\Omega := \bigcup_{n=0}^{\infty} \Omega_n \bigcup (E \times R) \times ((0, +\infty) \times E \times R)^{\infty})$, and the corresponding Borel σ-algebra $\mathscr{F} := \mathscr{B}(\Omega)$, where $E_\Delta := E \bigcup \{\Delta\}$, $\Omega_n := \{(x_0, \lambda_0, \theta_1, x_1, \lambda_1, \ \ldots, \theta_n, x_n, \lambda_n, \theta_{n+1}, \ldots, \infty, \Delta, \ \infty, \ldots) |\ x_0 \in E, \lambda_0 \in R, x_l \in E, \lambda_l \in R, \theta_l \in (0, \infty)$, for each $1 \leq l \leq n+1, n \geq 0\}$ with an artificial cemetery state $\Delta \notin E$.

For any $n \geq 0$, $\omega := (x_0, \lambda_0, \theta_1, x_1, \lambda_1, \ldots, \theta_n, x_n, \lambda_n, \theta_{n+1}, \ \ldots) \in \Omega$, let $h_n(\omega) := (x_0, \lambda_0, \theta_1, x_1, \lambda_1, \ldots, \ \theta_n, x_n, \lambda_n)$ be the n-component history for $n \geq 1$ and $h_0(\omega) := (x_0, \lambda_0)$. Moreover, for $n \geq 0$, some random variables X_n, Λ_n, S_n on $(\Omega, \mathscr{F})$ are defined as follows:

$$S_0(\omega) := \theta_0 = 0, X_n(\omega) := x_n, \Lambda_n(\omega) := \lambda_n, S_{n+1}(\omega) := s_{n+1} = \sum_{l=1}^{n+1} \theta_l. \tag{4}$$

In the following sections, the parameter ω is omitted for convenience. Then, the state process $\{\xi_s, s \geq 0\}$ is defined by

$$\xi_s(\omega) := \sum_{n \geq 0} I_{\{S_n \leq s < S_{n+1}\}} \phi(X_n(\omega), s - S_n(\omega)) + \Delta I_{\{s \geq S_\infty\}}, \tag{5}$$

where $S_\infty := \lim_{n \to \infty} S_n$, I_D describes an indicator function on the set D. The controlled operation after the moment S_∞ is assumed to be embedded in the artificial cemetery state $\Delta \notin E$. Hence, it is recorded as $q(\cdot|\Delta, a_\Delta) := 0$, $r(\Delta, a_\Delta) := 0$, $A_\Delta := A \cup \{a_\Delta\}$ with an isolated point a_Δ.

The idea of more general policies is introduced to effectively represent the optimization problem.

Definition 2.1. A *history-dependent policy* $\pi(\omega, s)$ denotes a sequence $\pi = \{f_n, n \geq 0\}$ of Borel measurable mapping from Ω to A_Δ. For any $\omega = (x_0, \lambda_0, \theta_1, x_1, \lambda_1, \ldots, \theta_n, x_n, \lambda_n, \ldots) \in \Omega$ and $s \geq 0$,

$$\pi(\omega, s) = I_{\{s=0\}} f_0(x_0, \lambda_0) + \sum_{n \geq 0} I_{\{S_n < s \leq S_{n+1}\}} f_n(h_n(\omega)) + I_{\{s \geq S_\infty\}} \delta_{a_\Delta}(\mathrm{d}a), \tag{6}$$

where $\delta_{a_\Delta}(da)$ represents the Dirac measure at the point a_Δ. Π stands for the set of all deterministic history-dependent policies.

A policy $\pi = \{f_0, f_1, \ldots\} \in \Pi$ is called a *Markov* one, if $f_n(h_n(\omega)) = f_n^M(x_n, \lambda_n)$ $(n \geq 0)$ for some measurable mapping f_n^M from $E_\Delta \times R$ to A_Δ. The class of entire Markov policies is described by Π_m.

A Markov policy $\pi = \{f_0^M, f_1^M, \ldots\} \in \Pi_m$ is called to be *stationary*, if $f_n^M(x_n, \lambda_n) = f(x_n, \lambda_n) (n \geq 0)$ for a measurable function f from $E_\Delta \times R$ to A_Δ. This stationary policy is represented by f, and the set of all stationary policies is described as Π_s for simplicity. Clearly, $\Pi_s \subset \Pi_m \subset \Pi$.

The set of histories up to the time $s \geq 0$ is denoted by $H_0 = E \times R$ and $H_n = (E \times R) \times ((0, \infty] \times E_\Delta \times R)^n, n = 1, 2, \ldots$. For each initial probability measure γ on $E \times R$ and a policy $\pi = \{f_0, f_1, \ldots\} \in \Pi$, the unique probability P_γ^π on $(\Omega, \mathscr{F})$ is established by a theorem (Ionescu Tulcea) (e. g., Proposition 7.45 in [6]), which satisfies the following properties:

$$
\begin{aligned}
&P_\gamma^\pi(\Gamma \times (\mathrm{d}\theta_{n+1}, \mathrm{d}x_{n+1}, \mathrm{d}\lambda_{n+1})) := \int_\Gamma P_\gamma^\pi(\mathrm{d}h_n) I_{\{\theta_n < \infty\}} q(\mathrm{d}x_{n+1} | \phi(x_n, \theta_{n+1}), f_n(h_n)) \\
&\times \exp\{-\int_0^{\theta_{n+1}} q_{\phi(x_n,u)}(f_n(h_n))\mathrm{d}u\} \qquad (7) \\
&\times \delta_{[e^{\int_0^{\theta_{n+1}} \alpha(\phi(x_n,t))\mathrm{d}t}(\lambda_n - \int_0^{\theta_{n+1}} e^{-\int_0^u \alpha(\phi(x_n,t))\mathrm{d}t} r(\phi(x_n,u), f_n(h_n))\mathrm{d}u)]}(\mathrm{d}\lambda_{n+1})\, \mathrm{d}\theta_{n+1}, \\
&P_\gamma^\pi(\Gamma \times (\infty, \infty, \Delta)) := \int_\Gamma P_\gamma^\pi(\mathrm{d}h_n)\{I_{\{\theta_n = \infty\}} + I_{\{\theta_n < \infty\}} \qquad (8) \\
&\times \exp\{-\int_0^\infty q_{\phi(x_n,u)}(f_n(h_n))\, \mathrm{d}u\}\},
\end{aligned}
$$

for $\Gamma \in \mathcal{B}(H_n)$, the corresponding expectation operator is described as $\mathbb{E}_{(x,\lambda)}^\pi$ versus the probability measure $P_{(x,\lambda)}^\pi$. If the distribution γ depends on state (x, λ), $P_{(x,\lambda)}^\pi$ and $\mathbb{E}_{(x,\lambda)}^\pi$ are used instead of P_γ^π and $\mathbb{E}_\gamma^\pi$, respectively.

For any $(x, \lambda) \in E \times R, \pi \in \Pi$, we define the infinite discounted risk probability of PDMDPs:

$$
U^\pi(x, \lambda) := P_{(x,\lambda)}^\pi\Big(\int_0^{+\infty} e^{-\int_0^s \alpha(\xi_u)\mathrm{d}u} r(\xi_s, \pi_s)\mathrm{d}s \leq \lambda\Big), \qquad (9)
$$

where $r(\xi_s, \pi_s)(\omega) := r(\xi_s(\omega), \pi(\omega, s))$ for all $\omega \in \Omega$ and $t \geq 0$. This criterion can be used to measure the risk probability that the total rewards are no more than the reward goal λ under the policy π.

Definition 2.2. The value function $U^*(x, \lambda)$ of the infinite discounted risk probability optimization problem is defined as

$$
U^*(x, \lambda) = \inf_{\pi \in \Pi} U^\pi(x, \lambda) \quad \text{for each} \quad (x, \lambda) \in E \times R. \qquad (10)
$$

An optimal policy of risk probability policy is denoted by $\pi^* \in \Pi$, if

$$
U^{\pi^*}(x, \lambda) = U^*(x, \lambda). \qquad (11)
$$

Remark 2.3. In contrast to risk probability problems for CTMDPs [20], the infinite discounted risk probability problems appear more complicated because the deterministic flow function and the state-dependent discount factor should be considered, while the states evolve according to the deterministic flow function are unchanged and the state-dependent discount factor is fixed constant in CTMDPs [20].

The current work aims to solve the infinite discounted risk probability optimality problem for PDMDPs, investigate the condition for the existence of a risk probability optimal policy, and build an approach to derive the value function.

3. MAIN RESULTS

Let $\mathcal{U}_m$ be the class of entire Borel-measurable mappings $U : E \times R \to [0,1]$, which satisfies $U(x,\lambda) = 0$ for each $(x,\lambda) \in E \times (-\infty, 0)$. For any $U \in \mathcal{U}_m, x \in E, a \in A(x)$ and $f \in \Pi_s$, the operators M^f, M on $\mathcal{U}_m$ are defined as the following: if $\lambda \geq 0$,

$$\begin{aligned} M^a U(x,\lambda) &:= I_{[0,\lambda]}\left(\int_0^{+\infty} e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s\right) e^{-\int_0^{+\infty} q_{\phi(x,s)}(a)\mathrm{d}s} \\ &\quad + \int_0^{+\infty}\int_{E\setminus\{\phi(x,u)\}} U\Big(y, e^{\int_0^u \alpha(\phi(x,t))\mathrm{d}t} \\ &\quad \times(\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s)\Big) \\ &\quad \times e^{-\int_0^u q_{\phi(x,s)}(a)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u),a)\mathrm{d}u, \\ M^f U(x,\lambda) &:= M^f U(x,\lambda), \qquad (12) \\ MU(x,\lambda) &:= \inf_{a\in A(x)} M^a U(x,\lambda). \qquad (13) \end{aligned}$$

If $\lambda < 0$,

$$M^f U(x,\lambda) = M^a U(x,\lambda) = MU(x,\lambda) := 0, \qquad (14)$$

where $q_{\phi(x,s)}(a) := -q(\phi(x,s)|\phi(x,s),a)$.

Furthermore, the operators $(M^f)^n, M^n, n \geq 2$ on $\mathcal{U}_m$ are also defined as follows:

$$(M^f)^n U(x,\lambda) = M^f((M^f)^{n-1}U(x,\lambda)) \text{ and } M^n U(x,\lambda) = M(M^{n-1}U(x,\lambda)).$$

Let $\widetilde{\mathcal{U}}_m$ be the set of all Borel measurable functions $\widetilde{V} : E \times R \to [-1,1]$, where $\widetilde{V}(x,\lambda) = 0$ if $\lambda < 0$. For any $(x,\lambda) \in E\times R, \widetilde{V} \in \widetilde{\mathcal{U}}_m, f \in \Pi_s$, the operators $(\widetilde{M^f})^n\widetilde{V}, n \geq 1$ are defined as the following:

$$\begin{aligned} \widetilde{M^f}\widetilde{V}(x,\lambda) &:= \int_0^{+\infty}\int_{E\setminus\{\phi(x,u)\}} \widetilde{V}\Big(y, e^{\int_0^u \alpha(\phi(x,t))\mathrm{d}t} \\ &\quad \times(\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),f(x,\lambda))\mathrm{d}s)\Big) \\ &\quad \times e^{-\int_0^u q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u),f(x,\lambda))\mathrm{d}u, \\ (\widetilde{M^f})^n\widetilde{V}(x,\lambda) &:= M^f((M^f)^{n-1}\widetilde{V}(x,\lambda)), n \geq 2. \qquad (15) \end{aligned}$$

Remark 3.1. Compared with the operators (12) and (14) in CTMDPs [20], we know that the infinite discounted risk probability case in PDMDPs is more complex and difficult to deal with than the case in CTMDPs [20].

Based on [5, 11, 20], the following condition is established to ensure the existence of optimal policies.

Assumption 3.1. For any $(x,\lambda) \in E \times R$,

(a) $A(x)$ is compact.

(b) For all $x, y \in E$, the function $c(x,a)$ and $q(y|x,a)$ are continuous in $a \in A(x)$.

(c) For each fixed $U \in \mathcal{U}_m$, $\int_0^{+\infty} \int_{E\setminus\{\phi(x,u)\}} U\Big(y, e^{\alpha(x)u}(\lambda - \int_0^u e^{-\alpha(x)s} r(\phi(x,s),a)\mathrm{d}s)\Big)$ $\times e^{-\int_0^u q_{\phi(x,s)}(a)\mathrm{d}s}\, q(\mathrm{d}y|\phi(x,u),a)\mathrm{d}u$ is lower semi-continuous in $a \in A(x)$.

Remark 3.2. Assumption 3.1 is also referred to as the continuity-compactness condition. The feature of the flow function makes Assumption 3.1 more extensive than the standard continuity-compactness condition in [11, 20] for CTMDPs.

The operators have the following characteristics.

Lemma 3.3. Under Assumption 3.1, the subsequent assertions are fulfilled.

(a) If $U, V \in \mathcal{U}_m$, and $U \geq V$, then $M^a U(x,\lambda) \geq M^a V(x,\lambda)$ for each $a \in A(x)$, and $MU(x,\lambda) \geq MV(x,\lambda)$ for each $(x,\lambda) \in E \times R$.

(b) If $U \in \mathcal{U}_m$, then an $f \in \Pi_s$ exists such that $MU(x,\lambda) = M^f U(x,\lambda)$ for each $(x,\lambda) \in E \times R$.

Proof. (a) According to the definition of the operator M, it can be known that part (a) is satisfied.

(b) For any $(x,\lambda) \in E \times R$, under the continuity-compactness condition in Assumption 3.1, it can be seen from the measurable selection theorem (proposition D.5 in [14]) that the existence of $f \in \Pi_s$ is proved. □

The state process can be explosive due to the unbounded transition rates, which means that the state process jumps infinitely in a limited time. To avoid this situation, the following condition needs to be established.

Assumption 3.2. For each $\pi \in \Pi, (x,\lambda) \in E \times R$, $P^{\pi}_{(x,\lambda)}(S_\infty = \infty) = 1$.

Assumption 3.2 is also referred to as the non-explosive condition of the state process $\{\xi_s, s \geq 0\}$. To verify Assumption 3.2, a generalized "drift condition" is established based on [11, 13, 19] through the determined flow and transition rates.

Lemma 3.4. If $W \geq 1$ on E with the parameters $c_0 > 0$, $b_0 \geq 0$ represents a measurable mapping satisfying

(a) $\int_E W(\phi(y,s)) q(\mathrm{d}y \mid x,a) \leq c_0 W(\phi(x,s)) + b_0$, for any $(x,a) \in K, s \geq 0$;

(b) There is a sequence $\{E_n, n \geq 1, E_n \subseteq E\}$ which satisfies $E_n \uparrow E$, $\lim_{n\to\infty} \inf_{x\notin E_n} W(\phi(x,s)) = \infty, \sup_{x\in E_n} q^*(x) < \infty$ for all $n \geq 1$ with $s \geq 0, q^*(x) = \sup_{a\in A(x)} q_x(a)$.

Then, Assumption 3.2 holds.

P r o o f . Lemma 3.4 can be proved from [12, 16]. The histories include reward levels and the structure of the probability measure P^{π}_{γ} in (7) is slightly different from that in [12, 16] for the expected criterion, there are some slight differences between our proof and those in [12, 16], e. g., the equalities (A3) in [12, 16]. □

Remark 3.5. (a) The conditions of Lemma 3.4 are the extension of the drift condition for CTMDPs [19, 20]. The main difference is that our model imposes the drift condition on the transition rates and the determined flow function, while the model of CTMDPs in [19, 20] only imposes some conditions on the transition rates. When the determined flow function $\phi(x, s) = x$ for $s \geq 0$, our conditions degenerate to the case of CTMDPs in [19, 20].

(b) Under the bounded transition rates (i. e., $\sup_{x\in E} q^*(x) < \infty$), Assumption 3.2 is fulfilled by considering $W \equiv 1, E_n \equiv E$.

For any $(x, \lambda) \in E \times R$ and $\pi \in \Pi$, under Assumption 3.2, the state process $\{\xi_s, s \geq 0\}$ is non-explosive. By the continuity of a probability measure, $U^{\pi}(x, \lambda)$ can be represented as follows:

$$\begin{aligned}
U^{\pi}(x,\lambda) &= P^{\pi}_{(x,\lambda)}\Big(\int_0^{+\infty} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\
&= P^{\pi}_{(x,\lambda)}\Big(\sum_{m=0}^{\infty}\int_{S_m}^{S_{m+1}} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\
&= P^{\pi}_{(x,\lambda)}\Big(\bigcap_{n=1}^{\infty}\sum_{m=0}^{n}\int_{S_m}^{S_{m+1}} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\
&= \lim_{n\to\infty} P^{\pi}_{(x,\lambda)}\Big(\sum_{m=0}^{n}\int_{S_m}^{S_{m+1}} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\
&:= \lim_{n\to\infty} U^{\pi}_n(x,\lambda).
\end{aligned}$$

Then, a monotone non-increasing sequence $\{U^{\pi}_n(x, \lambda), n = -1, 0, 1, \ldots\}$ with $U^{\pi}_{-1}(x, \lambda) := I_{[0,\infty)}(\lambda)$ can be obtained.

To prove that U^* is a solution to the corresponding optimality equation, several lemmas need to be given.

Lemma 3.6. Assume that Assumptions 3.1 and 3.2 hold. Then, for any $(x, \lambda) \in E \times R, n \geq -1, \pi = \{f_0, f_1, \ldots\} \in \Pi$.

(a) $U^{\pi}_n \in \mathcal{U}_m$ and $U^{\pi} \in \mathcal{U}_m$;

(b) $U^{\pi}_{n+1}(x, \lambda) = M^{f_0} U^{{}^1\pi}_n(x, \lambda)$ and $U^{\pi}(x, \lambda) = M^{f_0} U^{{}^1\pi}(x, \lambda)$, where ${}^1\pi := \{\tilde{f}_0, \tilde{f}_1, ...\}$ denotes the 1-shift policy of π, $\tilde{f}_k(x_1, \lambda_1, \ldots, \theta_{k+1}, x_{k+1}, \lambda_{k+1}) := f_{k+1}(x, \lambda, \theta_1, x_1, \lambda_1, \ldots, \theta_{k+1}, x_{k+1}, \lambda_{k+1}), k = 0, 1, \ldots$.

When for any $\pi = f \in \Pi_s$, $U^f(x, \lambda) = M^f U^f(x, \lambda)$.

P r o o f. (a) For any $(x,\lambda)\in E\times(-\infty,0), \pi\in\Pi$, by Remark 2.3, we know that part (a) is true. For any $(x,\lambda)\in E\times R^+, \pi\in\Pi$, we prove parts (a) and (b) together by induction for $n\geq -1$. Clearly, $U^\pi_{-1}=1\in\mathcal{U}_m$. Assume that the results hold for $n=-1,0,\ldots,k$. Then, by (7), the following is attained,

$$
\begin{aligned}
& U^\pi_{k+1}(x,\lambda)\\
= \;& P^\pi_{(x,\lambda)}\Big(\sum_{m=0}^{k+1}\int_{S_m}^{S_{m+1}} e^{-\int_0^s\alpha(\xi_t)\mathrm{d}t}r(\xi_s,\pi_s)\mathrm{d}s\leq\lambda\Big)\\
= \;& E^\pi_{(x,\lambda)}[I_{\{\int_0^{S_1} e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),\pi_s)\mathrm{d}s\leq\lambda, S_1=\infty\}}]\\
& +E^\pi_{(x,\lambda)}[I_{\{\sum_{m=0}^{k+1}\int_{S_m}^{S_{m+1}} e^{-\int_0^s\alpha(\xi_t)\mathrm{d}t}r(\xi_s,\pi_s)\mathrm{d}s\leq\lambda, S_1<\infty\}}]\\
= \;& I_{[0,\lambda]}(\int_0^{+\infty} e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)e^{-\int_0^{+\infty}q_{\phi(x,s)}(f_0)\mathrm{d}s}\\
& +E^\pi_{(x,\lambda)}[E^\pi_{(x,\lambda)}[I_{\{\sum_{m=0}^{k+1}\int_{S_m}^{S_{m+1}} e^{-\int_0^s\alpha(\xi_t)\mathrm{d}t}r(\xi_s,\pi_s)\mathrm{d}s\leq\lambda, S_1<\infty\}}|\xi_{S_0},\Lambda_0,S_1,\xi_{S_1},\Lambda_1]]\\
= \;& I_{[0,\lambda]}(\int_0^{+\infty} e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)e^{-\int_0^{+\infty}q_{\phi(x,s)}(f_0)\mathrm{d}s}\\
& +\int_0^{+\infty}\int_{E\setminus\{\phi(x,u)\}} P^\pi_{(x,\lambda)}\Big(\sum_{m=1}^{k+1}\int_{S_m}^{S_{m+1}} e^{-\int_0^s\alpha(\xi_t)\mathrm{d}t}r(\xi_s,\pi_s)\mathrm{d}s\\
& \leq\lambda-\int_0^u e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s|\xi_{S_0}=x,\Lambda_0=\lambda,S_1=u,\\
& \xi_{S_1}=y,\Lambda_1=e^{\int_0^u\alpha(\phi(x,t))\mathrm{d}t}(\lambda-\int_0^u e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)\Big)\\
& \times e^{-\int_0^u q_{\phi(x,s)}(f_0)\mathrm{d}s}q(\mathrm{d}y|\phi(x,u),f_0)\mathrm{d}u\\
= \;& I_{[0,\lambda]}(\int_0^{+\infty} e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)e^{-\int_0^{+\infty}q_{\phi(x,s)}(f_0)\mathrm{d}s}\\
& +\int_0^{+\infty}\int_{E\setminus\{\phi(x,u)\}} P^\pi_{(x,\lambda)}\Big(\sum_{m=1}^{k+1}\int_{S_m}^{S_{m+1}} e^{-\int_0^{l+u}\alpha(\xi_t)\mathrm{d}t}r(\xi_{l+u},\pi_{l+u})\mathrm{d}l\\
& \leq\lambda-\int_0^u e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s|\xi_{S_0}=x,\Lambda_0=\lambda,S_1=u,\xi_{S_1}=y,\\
& \Lambda_1=e^{\int_0^u\alpha(\phi(x,t))\mathrm{d}t}(\lambda-\int_0^u e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)\Big)\\
& \times e^{-\int_0^u q_{\phi(x,s)}(f_0)\mathrm{d}s}q(\mathrm{d}y|\phi(x,u),f_0)\mathrm{d}u\\
= \;& I_{[0,\lambda]}(\int_0^{+\infty} e^{-\int_0^s\alpha(\phi(x,t))\mathrm{d}t}r(\phi(x,s),f_0)\mathrm{d}s)e^{-\int_0^{+\infty}q_{\phi(x,s)}(f_0)\mathrm{d}s}\\
& +\int_0^{+\infty}\int_{E\setminus\{\phi(x,u)\}} P^{1\,\pi}_{(y,e^{\int_0^u\alpha(\phi(x,t))\mathrm{d}t}(\lambda-\int_0^u e^{-\alpha(x)s}r(\phi(x,s),f_0)\mathrm{d}s))}\\
& \times\Big(\sum_{m=0}^{k}\int_{S_m}^{S_{m+1}} e^{-\int_0^s\alpha(\xi_t)\mathrm{d}t}r(\xi_s,\pi_s)\mathrm{d}s\leq e^{\int_0^u\alpha(\phi(x,s))\mathrm{d}t}
\end{aligned}
$$

$$\times(\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s), f_0)\mathrm{d}s)\Big) e^{-\int_0^u q_{\phi(x,s)}(f_0)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u), f_0)\mathrm{d}u$$

$$\begin{aligned}
= \quad & I_{[0,\lambda]}(\int_0^{+\infty} e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s), f_0)\mathrm{d}s) e^{-\int_0^{+\infty} q_{\phi(x,s)}(f_0)\mathrm{d}s} \\
& + \int_0^{+\infty} \int_{E\setminus\{\phi(x,u)\}} U_k^1{}^{\pi}\Big(y, e^{\int_0^u \alpha(\phi(x,t))\mathrm{d}t}(\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s), f_0)\mathrm{d}s)\Big) \\
& \times e^{-\int_0^u q_{\phi(x,s)}(f_0)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u), f_0)\mathrm{d}u \\
:= \quad & M^{f_0} U_k^1{}^{\pi}(x,\lambda).
\end{aligned}$$

Thus, through the induction, it can be obtained that $U_n^\pi \in \mathcal{U}_m$ and $\lim_{n\to\infty} U_n^\pi = U^\pi \in \mathcal{U}_m$.

(b) For each $(x,\lambda) \in E \times R, n \geq -1$, according to the proof of part (a), we have

$$U_{n+1}^\pi(x,\lambda) = M^{f_0} U_n^1{}^{\pi}(x,\lambda).$$

Letting $n \to \infty$, based on the dominated convergence theorem, $U^\pi(x,\lambda) = M^{f_0} U^1{}^{\pi}(x,\lambda)$ is attained. In particular, if $\pi = f \in \Pi_s$, $U^f(x,\lambda) = M^f U^f(x,\lambda)$. □

Theorem 3.7. Under Assumptions 3.1 and 3.2, for any $(x,\lambda) \in E \times R, f \in \Pi_s$, the following assertions hold.

(a) If $u, v \in \mathcal{U}_m, n \geq 1$, then $(\widetilde{M}^f)^n(u-v)(x,\lambda) \leq P^f_{(x,\lambda)}(S_n < \infty)$ on $E \times R$.

(b) If $u, v \in \mathcal{U}_m$, $u(x,\lambda) - v(x,\lambda) \leq \widetilde{M}^f(u(x,\lambda) - v(x,\lambda))$, then $u(x,\lambda) \leq v(x,\lambda)$ on $E \times R$.

(c) $U^f \in \mathcal{U}_m$ is the unique solution to the equation $U^f(x,\lambda) = M^f U^f(x,\lambda)$ on $E \times R$.

P r o o f. (a) The following expression is proved by induction

$$(\widetilde{M}^f)^n(u-v)(x,\lambda) \leq P^f_{(x,\lambda)}(S_n < \infty) \quad \forall\ (x,\lambda) \in E \times R, n \geq 1. \tag{16}$$

When $n = 1$, for any $(x,\lambda) \in E \times R, f \in \Pi_s$, since $u, v \in \mathcal{U}_m$, by (12), we know that

$$\begin{aligned}
& \widetilde{M}^f(u-v)(x,\lambda) \\
= \quad & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} (u-v)\Big(y, e^{\int_0^t \alpha(\phi(x,l))\mathrm{d}l}(\lambda - \int_0^t e^{-\int_0^s \alpha(\phi(x,l))\mathrm{d}l} r(\phi(x,s), f)\mathrm{d}s)\Big) \\
& \times e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t, \\
\leq \quad & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t.
\end{aligned} \tag{17}$$

Moreover, by (7), we have

$$\begin{aligned} & P^f_{(x,\lambda)}(S_1 < \infty) \\ = \; & E^f_{(x,\lambda)}[E^f_{(x,\lambda)}[I_{\{S_1<\infty\}}|\xi_{S_0}, \Lambda_0, S_1, \xi_{S_1}, \Lambda_1]] \\ = \; & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t. \end{aligned} \tag{18}$$

Comparing (17) with (18), we know that (16) holds for $n = 1$.

Suppose that (16) is valid for $n = k$. For $(x, \lambda) \in E \times R$, the following relation can be derived from the induction

$$\begin{aligned} & (\widetilde{M}^f)^{k+1}(u - v)(x, \lambda) \\ = \; & \widetilde{M}^f(\widetilde{M}^f)^k(u - v)(x, \lambda) \\ = \; & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} (\widetilde{M}^f)^k(u - v)\Big(y, e^{\int_0^t \alpha(\phi(x,l))\mathrm{d}l} \\ & (\lambda - \int_0^t e^{-\int_0^s \alpha(\phi(x,l))\mathrm{d}l} r(\phi(x, s), f)\mathrm{d}s)\Big) \\ & \times e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t, \\ \leq \; & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} P^f_{(y, e^{\int_0^t \alpha(\phi(x,l))\mathrm{d}l}(\lambda - \int_0^t e^{-\int_0^s \alpha(\phi(x,l))\mathrm{d}l} r(\phi(x,s),f)\mathrm{d}s))}(S_k < \infty) \\ & \times e^{-\int_0^t q_{\phi(x,s)}(f)ds} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t. \end{aligned} \tag{19}$$

Alternatively, by the characteristic of conditional expectation, we obtain

$$\begin{aligned} & P^f_{(x,\lambda)}(S_{k+1} < \infty) \\ = \; & E^f_{(x,\lambda)}[E^f_{(x,\lambda)}[I_{\{S_{k+1}<\infty\}}|\xi_{S_0}, \Lambda_0, S_1, \xi_{S_1}, \Lambda_1]] \\ = \; & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} P^f_{(x,\lambda)}\Big(S_{k+1} < \infty|\xi_{S_0} = x, \Lambda_0 = \lambda, S_1 = t, \\ & \xi_{S_1} = y, \Lambda_1 = e^{\int_0^t \alpha(\phi(x,l))\mathrm{d}l}(\lambda - \int_0^t e^{-\int_0^s \alpha(\phi(x,l))\mathrm{d}l} r(\phi(x, s), f)\mathrm{d}s)\Big) \\ & \times e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}u \\ = \; & \int_0^{+\infty} \int_{E\setminus\{\phi(x,t)\}} P^f_{(y, e^{\int_0^t \alpha(\phi(x,l))\mathrm{d}l}(\lambda - \int_0^t e^{-\int_0^s \alpha(\phi(x,l))\mathrm{d}l} r(\phi(x,s),f)\mathrm{d}s))}(S_k < \infty) \\ & \times e^{-\int_0^t q_{\phi(x,s)}(f)\mathrm{d}s} q(\mathrm{d}y|\phi(x,t), f)\mathrm{d}t. \end{aligned} \tag{20}$$

Comparing (19) with (20), the induction hypothesis holds. Thus, part (a) is proved.

(b) For each $(x, \lambda) \in E \times R, f \in \Pi_s$, since $u(x, \lambda) - v(x, \lambda) \leq \widetilde{M}^f(u(x, \lambda) - v(x, \lambda))$, by induction and part (a), we have

$$u(x, \lambda) - v(x, \lambda) \leq (\widetilde{M}^f)^n(u - v)(x, \lambda) \leq P^f_{(x,\lambda)}(S_n < \infty) \;\; \forall n \geq 1. \tag{21}$$

Letting $n \to \infty$ in (21), and utilizing Assumption 3.2, the following expression holds

$$u(x,\lambda) - v(x,\lambda) \leq \lim_{n\to\infty} P^f_{(x,\lambda)}(S_n < \infty) = 0,$$

which implies $u(x,\lambda) \leq v(x,\lambda)$.

(c) According to Lemma 3.6 (b), it can be seen that $U^f \in \mathcal{U}_m$ is a solution to the equation $U^f = M^f U^f$ on $E \times R$ for any $f \in \Pi_s$. Suppose that there exists $V^{'} \in \mathcal{U}_m$ such that $V^{'} = M^f V^{'}$. Now, the definition of the operator yields $V^{'} - U^f = \widetilde{M}^f(V^{'} - U^f)$ on $E \times R$. Then, based on part (b), it can be seen that $V^{'} = U^f$, and the uniqueness of U^f is proved. □

Theorem 3.8. Assume that Assumptions 3.1 and 3.2 hold, for each $(x,\lambda) \in E \times R$, let $U^*_{-1} := I_{[0,+\infty)}(\lambda), U^*_{n+1} := MU^*_n, n \geq -1$. Then, $\lim_{n\to\infty} U^*_n = U^*$.

Proof. For any $(x,\lambda) \in E \times R$, since $U^*_{n+1} := MU^*_n, n \geq -1$, by the monotonicity of the operator M and $U^*_{-1} := I_{[0,+\infty)}(\lambda)$, we obtain $0 \leq U^*_{n+1} \leq U^*_n \leq 1$. Then, $\lim_{n\to\infty} U^*_n := \widetilde{U}$ exists. We will prove that $\tilde{U} = U^*$.

To prove $\tilde{U} \leq U^*$ by induction

$$U^*_n(x,\lambda) \leq U^\pi_n(x,\lambda), \tag{22}$$

for any $(x,\lambda) \in E \times R, \pi \in \Pi, n \geq -1$. When $n = -1$, for any $\pi \in \Pi$, since $U^*_{-1}(x,\lambda) = U^\pi_{-1}(x,\lambda) := I_{[0,+\infty)}(\lambda)$, this fact holds. Assume that $U^*_k(x,\lambda) \leq U^\pi_k(x,\lambda)$ for $(x,\lambda) \in E \times R, \pi \in \Pi$. Then, the following expression can be obtained from the induction hypothesis and Lemma 3.6(b):

$$U^*_{k+1}(x,\lambda) = MU^*_k(x,\lambda) \leq MU^{{}^1\pi}_k(x,\lambda) \leq M^{\varphi^0}U^{{}^1\pi}_k(x,\lambda) = U^\pi_{k+1}(x,\lambda).$$

Thus, the induction hypothesis holds and $U^*_n(x,\lambda) \leq U^\pi_n(x,\lambda)$ for any $(x,\lambda) \in E \times R, \pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi$. Hence, letting $n \to \infty$, we obtain $\tilde{U}(x,\lambda) \leq U^\pi(x,\lambda)$, indicating that $\tilde{U} \leq U^*$ as π is arbitrary.

To show the opposite, it is first shown that for each $(x,\lambda) \in E \times R$ and $n \geq -1$, there exists a policy $\eta \in \Pi_{RM}$ such that $U^*_n(x,\lambda) = U^\eta_n(x,\lambda)$. This fact trivially holds for $n = -1$, that is for any $\pi \in \Pi_{RM}$, $U^*_{-1}(x,\lambda) = U^\pi_{-1}(x,\lambda) = I_{[0,+\infty)}(\lambda)$. Assume that there exists a policy $\eta \in \Pi_{RM}$ such that $U^*_k(x,\lambda) = U^\eta_k(x,\lambda)$ for $k \geq -1$. Moreover, the existence of $f \in \Pi_s$ satisfying $M^f U^*_k(x,\lambda) = MU^*_k(x,\lambda)$ is guaranteed from Lemma 3.3(b). Then, Letting $\theta = \{f, \eta\} \in \Pi_{RM}$. Now, from the induction hypothesis and Lemma 3.6(b), we obtain

$$U^*_{k+1}(x,\lambda) = MU^*_k(x,\lambda) = M^f U^*_k(x,\lambda) = M^f U^\eta_k(x,\lambda) = U^\theta_{k+1}(x,\lambda),$$

and the fact holds. Then, we further prove that there is a policy $\xi \in \Pi_{RM}$ such that

$$U^*_n(x,\lambda) = U^\xi_n(x,\lambda) \geq U^\xi(x,\lambda) \geq U^*(x,\lambda),$$

which implies that letting $n \to \infty$, $\tilde{U} \geq U^*$. This proves $\tilde{U} = U^*$. □

Remark 3.9. For each $(x, \lambda) \in E \times R$, based on Theorem 3.8, we develop a so-called the value iteration algorithm to derive the value U^* as follows: Let $U^*_{-1} := I_{(-\infty,0)}(\lambda)$ and $U^*_{n+1} = MU^*_n, n \geq -1$. Then, $\lim_{n\to\infty} U^*_n = U^*$.

Theorem 3.10. Under Assumptions 3.1 and 3.2, for any $(x, \lambda) \in E \times R$, the subsequent conclusions can be derived.

(a) U^* is the unique solution of the corresponding optimality equation $U^* = MU^*$.

(b) There exists a policy $f^* \in \Pi_s$ such that $U^* = M^{f^*}U^*$ and $U^* = U^{f^*}$. Then, the optimal policy of risk probability $\pi^* := \{\tilde{f}^*_0, \tilde{f}^*_1, \ldots, \tilde{f}^*_k, \ldots\}$ is optimal, where $\tilde{f}^*_0(x, \lambda) := f^*(x, \lambda)$, $\tilde{f}^*_1(x, \lambda, \theta_1, x_1, \lambda_1) := f^*(x_1, \hat{\lambda}_1)$, $\ldots$, $\tilde{f}^*_k(x, \lambda, \theta_1, x_1, \lambda_1, \ldots, \theta_k, x_k, \lambda_k) := f^*(x_k, \hat{\lambda}_k)$ for any $(x, \lambda, \theta_1, x_1, \lambda_1, \ldots, \theta_k, x_k, \lambda_k) \in H_k, k \geq 0$.

P r o o f. (a) For each $(x, \lambda) \in E \times R, \pi = \{f_0, f_1, \ldots\} \in \Pi$, based on Lemma 3.6(b) and (13), the following expression holds

$$U^\pi(x, \lambda) = M^{f_0}U^{1\pi}(x, \lambda) \geq M^{f_0}U^*(x, \lambda) \geq MU^*(x, \lambda),$$

which implies $U^*(x, \lambda) \geq MU^*(x, \lambda)$, as π is arbitrary.

To prove the converse, for any $(x, \lambda) \in E \times R, a \in A(x)$, by using Theorem 3.8 and (13), we have

$$U^*_{n+1}(x, \lambda) = MU^*_n(x, \lambda) \leq M^a U^*_n(x, \lambda).$$

Hence, by applying the dominated convergence theorem and part (a), letting $n \to \infty$, we obtain

$$U^*(x, \lambda) \leq M^a U^*(x, \lambda),$$

which implies $U^*(x, \lambda) \leq MU^*(x, \lambda)$, as $a \in A(x)$ is arbitrary. Then, $U^* = MU^*$ is proved.

For any $(x, \lambda) \in E \times R$, since $U^* = MU^*$, according to Lemma 3.3(b), we know that there exists a policy $f^* \in \Pi_s$ such that

$$U^* = M^{f^*}U^*. \tag{23}$$

Besides, suppose that $U' \in \mathcal{U}_m$ is another solution to the risk probability optimality equation $U' = MU'$. Correspondingly, the existence of $f' \in \Pi_s$ fulfilling

$$U' = M^{f'}U'. \tag{24}$$

is guaranteed from Lemma 3.3(b). Combining (23) and (24), we have $U^* - U' \leq \widetilde{M}^{f'}(U^* - U')$ and $U' - U^* \leq \widetilde{M}^{f^*}(U' - U^*)$, which together with Theorem 3.7(b) indicates that $U^* = U'$. The uniqueness of the solution U^* has been proved. Since $U^* = MU^* = M^{f^*}U^*$ by Theorem 3.7(c), we obtain that $U^* = U^{f^*}$.

(b) For any $(x, \lambda) \in E \times R, k \geq 0$, let $\tilde{f}^*_0(x, \lambda) := f^*(x, \lambda), \tilde{f}^*_k(x, \lambda, \theta_1, x_1, \lambda_1, \ldots, \theta_k, x_k, \lambda_k) := f^*(x_k, \hat{\lambda}_k)$, $\pi^* := \{\tilde{f}^*_0, \tilde{f}^*_1, \ldots, \tilde{f}^*_k, \ldots\}$. By part (b), (3), (6) and (7), for all $k \geq 0$, it can be obtained that $P^{f^*}_{\gamma,k} = P^{\pi^*}_{\gamma,k}$, which indicates that $P^{f^*}_\gamma = P^{\pi^*}_\gamma, P^{f^*}_\gamma(\int_0^{+\infty} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s, \pi^*_s)\,\mathrm{d}s \neq \int_0^{+\infty} e^{-\int_0^s \alpha(\xi_t)\mathrm{d}t} r(\xi_s, f^*)\,\mathrm{d}s) = 0$ and $U^{\pi^*}(x, \lambda) = U^{f^*}(x, \lambda) = U^*(x, \lambda)$. Therefore, π^* is optimal. □

Remark 3.11. The new fact (16) in Theorem 3.7 is established using the condition of Assumption 3.2 to guarantee the existence of the optimal policy of risk probability, that is the controlled state process is non-explosive. However, the authors in [20] need to establish the additional first passage condition, indicating that the system will gradually reach the target set within a finite time for all the initial states. Then, it can be concluded that the established condition is weaker than that in [20].

According to Theorem 3.10,the value iteration method could be employed to derive the value U^* as follows:

The value iteration algorithm:

Step 1: Let $U^*_{-1}(x,\lambda) := I_{[0,\infty)}(\lambda)$, for $(x,\lambda) \in E \times R$.

Step 2: For each fixed $(x,\lambda) \in E \times R, a \in A(x), n \geq 0$,

$$\begin{aligned} M^a U^*_n(x,\lambda) &= I_{[0,\lambda]}\Big(\int_0^{+\infty} e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s) e^{-\int_0^{+\infty} q_{\phi(x,s)}(a)\mathrm{d}s} \\ &\quad + \int_0^{+\infty} \int_{E\setminus\{\phi(x,u)\}} U\Big(y, e^{\int_0^u \alpha(\phi(x,t))\mathrm{d}t}(\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} \\ &\quad \times r(\phi(x,s),a)\mathrm{d}s)\Big) e^{-\int_0^u q_{\phi(x,s)}(a)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u),a)\mathrm{d}u, \\ U^*_{n+1}(x,\lambda) &= \min_{a\in A(x)}\{M^a U^*_n(x,\lambda)\}. \end{aligned}$$

Step 3: If $|U^*_{n+1} - U^*_n| < 10^{-12}$, the value U^*_{n+1} is approximately received as the value function U^*. Otherwise, the program continues to run step 2 for $n+1$.

4. EXAMPLES

In this section, two examples are used to illustrate the main results of the discounted optimality PDMDPs. The first example verifies the existence condition of the optimal policy and the optimality equation. The second example uses the value iteration algorithm to show the numerical calculations of the value function and the optimal policy of risk probability.

Example 4.1. (Optimal control of an investment system) Consider a controlled investment system with the state x describing the enterprise's capital value. When the status of the system is $x \in E := [0,+\infty)$, the reward level is $\lambda \in R^+$, the decision-maker can borrow a loan a from a finite set $A(x) \subset [0,x]$ to make a new investment plant. After this action is selected, the system state evolves according to the determined flow function $\phi(x,s)$ until the new state jump time. At this time, the system will reach a new state Now, the system state jumps into a new state $x_1 \in E$, which is governed by transition rate $q(\mathrm{d}x_1|x,a)$, and get some rewards at the rate $r(x,a) \geq 0$. We formulate this system as a piecewise deterministic MDPs with the state space $E = [0,\infty)$, the action space $A = [0,\infty)$, the set of admissible actions $A(x)$, the deterministic flow $\phi(x,s) = xe^{-s}$. The transition rates can be described as

$$q(C|x,a) = (x-a+1)\Big[\int_{C\setminus\{x\}} \frac{1}{x-a+1} e^{-y/(x-a+1)}\mathrm{d}y - \delta_{\{x\}}(C)\Big], \tag{25}$$

for $(x,a) \in K, C \in \mathcal{B}(E)$. For the system, the purpose of the decision maker is to derive an optimal policy of risk probability.

The existence of the optimal policy of risk probability for the proposed model is investigated. First, Assumption 3.2 is verified by taking $W(x) = x + 1$. For $(x,a) \in K$, we obtain

$$\begin{aligned} & \int_E W(\phi(y,s))q(\mathrm{d}y \mid x,a) \\ = & \int_0^\infty (ye^{-s}+1)(x-a+1)\Big[\frac{1}{x-a+1}e^{-y/(x-a+1)}\mathrm{d}y - \delta_{\{x\}}(\mathrm{d}y)\Big] \\ = & e^{-s}(x-a+1)(1-a) \\ \leq & e^{-s}(x+1) \\ = & W(\phi(x,s)). \end{aligned}$$

Then, condition (a) in Lemma 3.4 holds, where $c_0 = 1, b_0 = 0$.

Taking $E_n = [0,n], n = 1, \ldots,$ then $E_n \uparrow E, \sup_{x\in E_n} q^*(x) = n+1 < \infty$, and $\lim_{n\to\infty} \inf_{x\notin E_n} W(\phi(x,s)) = \lim_{n\to\infty}(ne^{-s}+1) = \infty$ for $s > 0$. Thus, condition (b) in Lemma 3.4 is also verified. It follows from Lemma 3.4, we know that Assumption 3.2 holds. Since $A(x)$ is a bounded closed set for $x \in E$, and it is confirmed that Assumption 3.1 and 3.2 hold. Also, the existence of an optimal policy of the risk probability is guaranteed by using Theorem 3.10.

Example 4.2. (Optimal production management) Consider a production management system of a industry corporation where the state $x \in E := [0, +\infty)$ denotes the number of products, the constant $k > 0$ represents the product quantity threshold. At the initial moment $s = 0$, when the company has a small number of products $x \in (0,k)$, the decision maker can use the production plan $a \in \{a_{11}, a_{12}\}$ to expand the production scale.When the company has a lot of products $x \in [k, +\infty)$, based on the initial reward level λ_0, the decision maker can choose the production plan $a \in \{a_{11}, a_{12}, a_{21}, a_{22}\}$. Consequently, the change of the system state is based on the flow $\phi(x,s)(s \in [0, s_1))$ up to the time s_1. At a new decision-making moment s_1, the system state jumps into a new state $x_1 \in E$ according to the transition rate $q(\mathrm{d}x_1|x,a_0)$. Considering the influence of the varying discount factor $\alpha(\cdot)$, the decision maker obtains the rewards $\int_0^{s_1} e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s$. The system state undergoes repeated evolution. When the company doesn't have any products $x = 0$, the decision-maker cannot choose any production plan (which is denoted by a_{01}) and will not receive any reward $r(0, a_{01}) = 0$.

We formulate this idle fund management system as a PDMDP, where the system states between two jumps change according to the determined drift function $\phi(x,s) = xe^s$. The model parameters are provided as follows: The state threshold $k = 2$; the state space $E = [0, +\infty)$; the action sets $A(0) = \{a_{01}\}$, $A(x) = \{a_{11}, a_{12}\}$ for $x \in (0,2)$; $A(x) = \{a_{11}, a_{12}, a_{21}, a_{22}\}$ for $x \in [2, +\infty)$; the discount factor is given by

$$\alpha(x) = \begin{cases} 1, & x = 0; \\ x, & 0 < x < 1; \\ \frac{1}{2x}, & x \geq 1. \end{cases} \tag{26}$$

The transition rates are assumed to be as follows: for each $D \in \mathcal{B}(E)$, $q(D|0, a_{01}) = 0$. For $x \in (0,2) \cup (2,+\infty)$,

$$q(D|x,a_{11}) = \begin{cases} 0.056, & D = \{0\}; \\ -0.28, & D = \{x\}; \\ 0.224, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \qquad q(D|x,a_{12}) = \begin{cases} 0.056, & D = \{0\}; \\ -0.08, & D = \{x\}; \\ 0.024, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \tag{27}$$

For $x = 2$,

$$q(D|2,a_{11}) = \begin{cases} 0.28, & D = \{0\}; \\ -0.28, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \qquad q(D|2,a_{12}) = \begin{cases} 0.08, & D = \{0\}; \\ -0.08, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \tag{28}$$

For $x \in [2,+\infty)$,

$$q(D|x,a_{21}) = \begin{cases} 0.084, & D = \{0\}; \\ 0.056, & D = \{1\}; \\ -0.14, & D = \{x\}; \\ 0, & \text{others.} \end{cases} \qquad q(D|x,a_{22}) = \begin{cases} 0.13, & D = \{0\}; \\ 0.13, & D = \{1\}; \\ -0.26, & D = \{x\}; \\ 0, & \text{others.} \end{cases} \tag{29}$$

For each $x \in E$, the reward rates is provided as

$$r(0,a_{01}) = 0, \quad r(x,a_{11}) = x, \quad r(x,a_{12}) = 2x, \quad r(x,a_{21}) = \sqrt{x}, \quad r(x,a_{22}) = 2\sqrt{x}.$$

For this system, the decision-maker mainly focuses on the existence of the optimal risk probability policy and the calculation of the value function.

First, we need to verify Assumption 3.2 to assure the existence of the optimal policy of the risk probability. Since the set $A(x)$ is finite for any $x \in E$, by Remark 3.2 we know that Assumption 3.1 is satisfied. Based on the uniform boundedness of the transition rates in (27)-(29), we know that Assumption 3.2 holds. Then, from Theorem 3.10, the existence of the discounted optimal policy of the risk probability.

To further show the feasibility and effectiveness of the algorithm, we choose the calculation result of the system in state $x \in \{0,1,2\}$ as an example. When the system appears in other states, it can be calculated similarly. Since $r(0,a_{01}) = 0$, we know that $U^*(0,\lambda) = I_{[0,+\infty)}(\lambda)$. The value iteration method is exploited to derive the value $U^*(1,\lambda)$ and $U^*(2,\lambda)$ as follows :

Step 1: For $\lambda \in R^+$ and $x \in [0,+\infty)$, let $U^*_{-1}(x,\lambda) := 1$.

Step 2: For $x = 1, n \geq 0$,

$$\begin{aligned} M^{a_{11}}U^*_n(1,\lambda) &= 0.2 \times (1 - e^{-0.56\ln(\lambda/2+1)}) \\ &+ 0.8 \times 0.28 \times \int_0^{+\infty} U^*_n(2, e^{0.5u}(\lambda - \int_0^u e^{0.5s}ds))e^{-0.28u}\,\mathrm{d}u, \end{aligned} \tag{30}$$

$$\begin{aligned} M^{a_{12}}U_n^*(1,\lambda) &= 0.3\times(1-e^{-0.16\ln(\lambda/4+1)}) \\ &+ 0.7\times 0.08\times\int_0^{+\infty} U_n^*(2,e^{0.5u}(\lambda-2\int_0^u e^{0.5s}ds))e^{-0.08u}\,\mathrm{d}u, \\ U_{n+1}^*(1,\lambda) &= \min\{M^{a_{11}}U_n^*(1,\lambda), M^{a_{12}}U_n^*(1,\lambda)\}. \end{aligned}$$

For $x=2, n\geq 0$,

$$\begin{aligned} M^{a_{11}}U_n^*(2,\lambda) &= 1-e^{-0.6\ln(1+\frac{3}{8}\lambda)}, \\ M^{a_{12}}U_n^*(2,\lambda) &= 1-e^{-\frac{8}{75}\ln(1+\frac{3}{16}\lambda)}, \end{aligned}$$

$$\begin{aligned} M^{a_{21}}U_n^*(2,\lambda) &= 0.6\times(1-e^{-0.56\ln(\frac{\sqrt{2}\lambda}{8}+1)}) \\ &+0.4\times 0.14\times\int_0^{+\infty} U_n^*(1,e^{0.25u}(\lambda-\sqrt{2}\int_0^u e^{0.25s}d\mathrm{d}s))e^{-0.14u}\,\mathrm{d}u, \\ M^{a_{22}}U_n^*(2,\lambda) &= 0.5\times(1-e^{-1.04\ln(\frac{\sqrt{2}\lambda}{16}+1)}) \\ &+0.5\times 0.26\times\int_0^{+\infty} U_n^*(1,e^{0.25u}(\lambda-2\sqrt{2}\int_0^u e^{0.25s}\mathrm{d}s))e^{-0.26u}\,\mathrm{d}u, \\ U_{n+1}^*(2,\lambda) &= \min\{M^{a_{11}}U_n^*(2,\lambda), M^{a_{12}}U_n^*(2,\lambda), M^{a_{21}}U_n^*(2,\lambda), M^{a_{22}}U_n^*(2,\lambda)\}. \end{aligned}$$

Step 3: If $|U_{n+1}^*-U_n^*|<10^{-12}$, the program goes to step 4. Then, U_{n+1}^* is approximately received as the value function U^*; otherwise, the program continues to run step 2 for $n+1$.

Step 4: Drawing the figures of $M^aU_n^*(x,\lambda), U^*(x,\lambda), x\in\{1,2\}$ by using Matlab software, see Figure 1 and Figure 2.

Remark 4.1. It is worth pointing out that the integral in (30) is calculated by the trapezoidal integration method in [21] with $t=e^{0.5u}(\lambda+2)-2e^u$, which is shown as follows:

$$\begin{aligned} &\int_0^{+\infty} U_n^*(2,e^{0.5u}(\lambda-\int_0^u e^{0.5s}ds))e^{-0.28u}\,\mathrm{d}u \\ =\ &\int_0^{+\infty} U_n^*(2,(e^{0.5u}(\lambda+2)-2e^u))e^{-0.28u}\,\mathrm{d}u \\ =\ &2\int_0^{\lambda} U_n^*(2,t)(\frac{(\lambda+2)+\sqrt{(\lambda+2)^2-8t}}{4})^{-1.56}\frac{1}{\sqrt{(\lambda+2)^2-8t}}\,\mathrm{d}t \\ \approx\ &\sum_{k=0}^{l-1}[U_n^*(2,kh)(\frac{(\lambda+2)+\sqrt{(\lambda+2)^2-8kh}}{4})^{-1.56}\frac{1}{\sqrt{(\lambda+2)^2-8kh}} \\ &+U_n^*(2,(k+1)h)(\frac{(\lambda+2)+\sqrt{(\lambda+2)^2-8(k+1)h}}{4})^{-1.56}\frac{1}{\sqrt{(\lambda+2)^2-8(k+1)h}}]\frac{h}{2}, \end{aligned}$$

where $k\leq l, k, l\in\mathbb{N}$, $lh=\lambda$, h is the step length, $\mathbb{N}$ is the set of all positive integers.

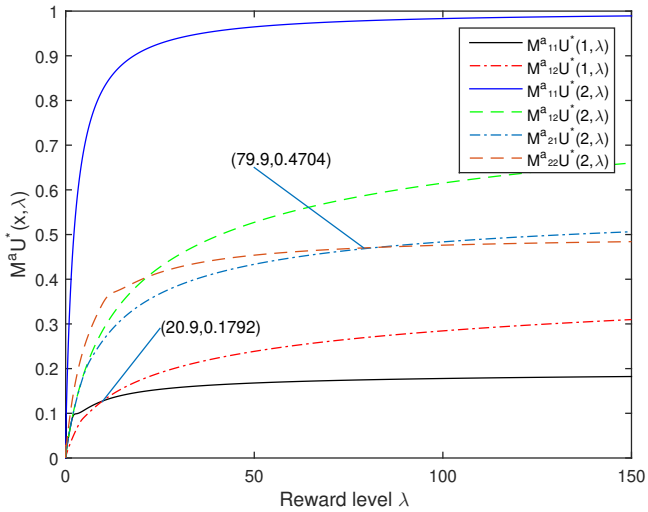

Fig. 1. The function $M^aU^*(x,\lambda)$.

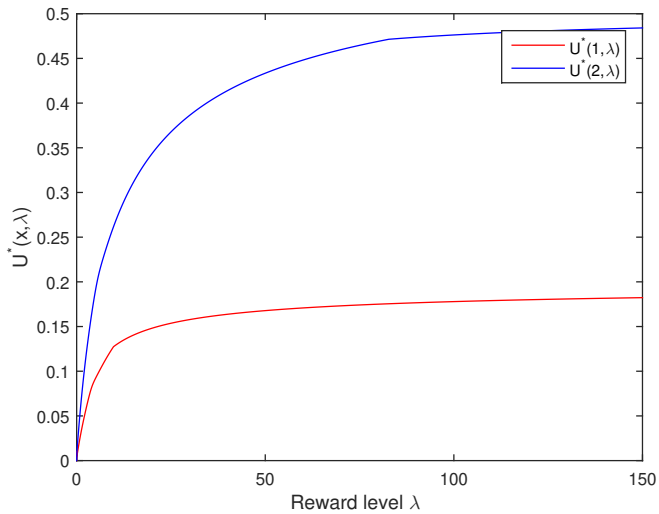

Fig. 2. The value function $U^*(x,\lambda)$.

According to the value function calculation and Figures 1 and 2, we obtain the following conclusions.

(a) In Figure 1, we know that when the system state $x = 1$, $\lambda \in (0, 20.9)$, the value $M^{a_{12}}U^*(1,\lambda)$ is lower than that $M^{a_{11}}U^*(1,\lambda)$; if $\lambda \in [20.9, +\infty)$, the value $M^{a_{11}}U^*(1,\lambda)$ is lower than that $M^{a_{12}}U^*(1,\lambda)$. When the system state $x = 2$, $\lambda \in (0, 79.9)$, the value $M^{a_{22}}U^*(2,\lambda)$ is the minimum value. $\lambda \in [79.9, +\infty)$, the value $M^{a_{21}}U^*(2,\lambda)$ is the minimum value. This implies that in state $x = 1$, if $\lambda \in (0, 20.9)$, the decision maker should choose low-risk action a_{12} instead of the action a_{11}. Conversely, if $\lambda \in [20.9, +\infty)$, the decision maker should choose low-risk action a_{11} instead of the action a_{12}. In state $x = 2$, if $\lambda \in (0, 79.9)$, the decision maker should choose low-risk action a_{22} instead of other actions. if $\lambda \in [79.9, +\infty)$, the decision maker should choose low-risk action a_{21} instead of other actions.

(b) As seen in Figures 1 and 2, when the system state $x \in \{1, 2\}$, the choice of optimal action is based on the following expression:

$$f^*(1,\lambda) = \begin{cases} a_{12}, & 0 \le \lambda < 20.9; \\ a_{11}, & \lambda \ge 20.9. \end{cases}, \quad f^*(2,\lambda) = \begin{cases} a_{21}, & 0 \le \lambda < 79.9; \\ a_{22}, & \lambda \ge 79.9. \end{cases} \tag{31}$$

At the initial time $s_0 = 0$, if the system state (x_0, λ_0) and (31) the control action $\tilde{f}_0^*(x_0,\lambda_0) := f^*(x_0,\lambda_0) \in A(x_0)$ is chosen by the decision maker. Consequently, according to $\phi(x_0, s) = x_0 e^s (s \in [s_0, s_1))$ the system states evolve up to the time s_1. Now, the system enters a new state x_1. During the period $[s_0, s_1)$, the decision maker gets the reward $\int_0^{s_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0,s), \tilde{f}_0^*)ds$. At time s_1, according to the historical information $(x_0, \lambda_0, \theta_1, x_1, \hat{\lambda}_1)$ and (31), the decision-maker selects a control action $\tilde{f}_1^*(x_0,\lambda_0,\theta_1,x_1,\lambda_1) := f^*(x_1,\hat{\lambda}_1) \in A(x_1)$, where the remaining reward goal becomes $\hat{\lambda}_1 = e^{\int_0^{\theta_1} \alpha(\phi(x_0,t))\mathrm{d}t}(\lambda_0 - \int_0^{\theta_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0,s), \tilde{f}_0^*)ds)$. The development of the system is repeated. Now, the optimal policy of the risk probability $\pi^* = \{\tilde{f}_0^*, \tilde{f}_1^*, \ldots\}$ is determined from Theorem 3.10 and (31).

ACKNOWLEDGEMENT

This work was supported by Guangxi science and technology base and talent project(Grant No.AD21159005); National Natural Science Foundation of China (Grant No.12361091, 11961005); Guangxi Natural Science Foundation Program(Grant No.2020GXNSFAA297196); ;Foundation of Guangxi Educational Committee(Grant No.KY2022KY0342); The Doctoral Foundation of Guangxi University of Science and Technology(Grant No.18Z06).

(Received January 26, 2023)

REFERENCES

[1] A. Almudevar: A dynamic programming algorithm for the optimal control of piecdewise deterministic Markov processes. SIAM J. Control Optim. *40* (2001), 525–539. DOI:10.1137/S0363012999364474

[2] D. Bertsekas and S. Shreve: Stochastic Optimal Control: The Discrete-Time Case. Academic Press Inc, New York 1978.

[3] O. L. V. Costa, F. Dufour: The vanishing discount approach for the average continuous of piecewise deterministic Markov processes. J. Appl. Probab. *46* (2009), 1157–1183. DOI:10.1017/S0021900200006203

[4] O. L. V. Costa, F. Dufour: Continuous Average Control of Piecewise Deterministic Markov Processes. Springer-Vrelag, New York 2013.

[5] N. Bauerle and U. Rieder: Markov Decision Processes with Applications to Finance. Springer, Heidelberg 2011.

[6] D. Bertsekas, S. Shreve: Stochastic Optimal Control: The Discrete-Time Case. Academic Press Inc, New York 1978.

[7] K. Boda, J. A. Filar, and Y. L. Lin: Stochastic target hitting time and the problem of early retirement. IEEE Trans. Automat. Control.*49* (2004), 409–419. DOI:10.1109/TAC.2004.824469

[8] M. H. A. Davis: Piecewise deterministic Markov processes: a general class of nondiffusion stochastic models. J. Roy. Statist. Soc. *46* (1984), 353–388. DOI:10.1111/j.2517-6161.1984.tb01308.x

[9] M. H. A. Davis: Markov Models and Optimization. Chapman and Hall 1993. DOI:10.1007/978-1-4899-4483-2

[10] F. Dufou, M. Horiguchi, and A. Piunovskiy: Optimal impulsive control of piecewise deterministic Markov processes. Stochastics *88* (2016), 1073–1098. DOI:10.1080/17442508.2016.1197925

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

[12] X. P. Guo and A. Piunovskiy: Discounted continuous-time Markov decision processes with constraints: unbounded transition and loss rates. Math. Oper. Res. *36* (2011), 105–132. DOI:10.1287/moor.1100.0477

[13] 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

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

[15] J. P. Hespanha: A model for stochastic hybrid systems with applications to communication networks. Nonlinear Anal. *62* (2005), 1353–1383. DOI:10.1016/j.na.2005.01.112

[16] Y. H. Huang and X. P. Guo: Finite-horizon piecewise deterministic Markov decision processes with unbounded transition rates. Stochastics *91* (2019), 67–95. DOI:10.1080/17442508.2018.1518450

[17] 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

[18] 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

[19] H. F. Huo, X. Wen: First passage risk probability optimality for continuous time Markov decision processes. Kybernetika *55* (2019), 114–133. DOI:10.14736/kyb-2019-1-0114

[20] 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

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

[22] 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

[23] Y. Ohtsubo and K. Toyonaga: Optimal policy for minimizing risk models in Markov decision processes. J. Math. Anal. Appl. *271* (2002), 66–81. DOI:10.1016/s0022-247x(02)00097-5

[24] A. Piunovskiy and Y. Zhang: Continuous-Time Markov Decision Processes: Borel Space Models and General Control Strategies. Springer, 2020.

[25] X. Wen, H. F. Huo, X. P. Guo: First passage risk probability minimization for piecewise deterministic Markov decision processes. Acta Math. Appl. Sinica *38* (2022), 549–567. DOI:10.1007/s10255-022-1098-0

[26] 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

[27] 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

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

Jinhua Cui, School of Science, Guangxi University of Science and Technology, Liuzhou, 545006. P. R. China.
e-mail: liveinsidecoco@163.com

Xian Wen, Corresponding author. School of Science, Guangxi University of Science and Technology, Liuzhou, 545006. P. R. China.
e-mail: wenxian879@163.com