Haifeng Huo; Xian Wen

The exponential cost optimality for finite horizon semi-Markov decision processes

Kybernetika, Vol. 58 (2022), No. 3, 301–319

Persistent URL: http://dml.cz/dmlcz/151031

KYBERNETIKA — VOLUME *58* (2022), NUMBER 3, PAGES 301–319

THE EXPONENTIAL COST OPTIMALITY FOR FINITE HORIZON SEMI-MARKOV DECISION PROCESSES

Haifeng Huo and Xian Wen

This paper considers an exponential cost optimality problem for finite horizon semi-Markov decision processes (SMDPs). The objective is to calculate an optimal policy with minimal exponential costs over the full set of policies in a finite horizon. First, under the standard regular and compact-continuity conditions, we establish the optimality equation, prove that the value function is the unique solution of the optimality equation and the existence of an optimal policy by using the minimum nonnegative solution approach. Second, we establish a new value iteration algorithm to calculate both the value function and the ϵ-optimal policy. Finally, we give a computable machine maintenance system to illustrate the convergence of the algorithm.

Keywords: semi-Markov decision processes, exponential cost, finite horizon, optimality equation, optimal policy

Classification: 90C40, 60E20

1. INTRODUCTION

As is well known, semi-Markov decision processes (SMDPs) form a more general stochastic optimal model, in which the sojourn time periods of the system state in SMDPs are allowed to follow any arbitrary probability distribution. This feature makes it widely applicable in many fields such as the queuing systems, reliability engineering, risk analysis and finance, and so on [2, 10, 11, 18, 24]. In recent years, a lot of research has been conducted on the classical expected criteria of SMDPs; see e. g., [14, 22, 24] for the finite horizon expected criteria; [2, 4, 10, 21, 23] for the expected discounted criteria; [10, 19, 27] for the expected average criteria.

These traditional criteria do not reflect the attitude of the decision maker towards the risk. Based on these situations, SMDPs with the expected exponential utility criteria (also known as risk-sensitive SMDPs) have attracted attention of scholars; see e. g.,[7, 13]. More precisely, the authors in [7] discussed risk sensitive SMDPs with a long-run average cost, established the optimality equation for the risk-sensitive average cost, and proved the existence of an optimal stationary policy under the continuity-compactness conditions. Recently, Huang, Lian and Guo [13] adopted the convex analytic approach to solve the unconstrained and constrained risk sensitive problems for

DOI: 10.14736/kyb-2022-3-0301

SMDPs, and proved the existence of the Bellman equation and the optimal policies under some continuity-compactness conditions. Risk sensitive MDPs is an important dynamic programming model and has been widely studied in discrete time Markov decision processes (DTMDPs) [3, 5, 6, 8, 17, 20], and continuous time Markov decision processes (CTMDPs) [9, 25, 28]. More specifically, Ghosh and Saha [9] considered finite and infinite horizon risk-sensitive problems for CTMDPs with bounded transition rates, proved that the value function is a unique solution to the HJB equation, showed the existence of an optimal Markov control. Wei [25] further considered the finite horizon risk sensitive cost criterion for CTMDPs with unbounded transition rates under the so-called drift condition, proved the existence of an optimal deterministic Markov policy by using the Feynman-Kac formula. Different from the mentioned works [9, 25], Zhang [28] reduced the infinite horizon risk-sensitive CTMDPs to an equivalent risk-sensitive DTMDP. Under the compactness-continuity condition, the author proved the existence of a deterministic stationary optimal policy. In this paper, we are going to discuss further the exponential cost optimality for finite horizon SMDPs.

Compared with the existing research work, we pay more attention to the calculation of the optimal policies and the feasibility and effectiveness of the algorithm, whereas the authors in [13] are more focused on the existing conditions of the optimal policies. More specifically, we use a so-called minimum nonnegative solution technique to establish the corresponding optimality equation and the existence of an optimal policy, which is different from the application of the convex analytic method in Huang [13]. Secondly, from a practical standpoint, we establish NEW facts (see Theorem 2 and 3) to ensure the existence of optimal policies, and develop a new value iteration algorithm to calculate the value function and the optimal policy. Moreover, we prove the convergence of the value iteration algorithm. However, there is little discussion on the calculation of the optimal optimal policy and the feasibility and convergence of the algorithm in [13].

The main contributions of this paper are reflected as follow: Since the criterion is the optimization of the expected exponential cost over a fixed finite horizon, it is normal to consider planning horizons in the standard histories, see Definition 1. Thus, we define a class of policies depending on histories with the additional planning horizons, construct the corresponding probability space, and provide the so-called regularity condition to ensure that the state process is non-explosive; see Lemma 1. Under some suitable conditions, we establish the corresponding optimality equation, prove the existence and uniqueness of the solution, and show the existence of an optimal policy using a minimum nonnegative solution approach, which is slightly different from the Feyman-Kac formula method in [25], the reduction method in [28] and the occupation measure method in [13]. Furthermore, we develop a value iteration algorithm for computing both the value function and the ε-exponential cost optimal policy; see Theorem 3. Moreover, we provided an example of machine maintenance system to analyse the convergence of the value iteration algorithm, and compute both the value function and the ε-exponential cost optimal policy.

The rest of this paper is organized as follows. In Section 2, we describe the optimal control model for semi-Markov decision processes. The main results are presented and proved in Section 3. In Section 4, an example illustrating our main results is given.

2. THE CONTROL MODEL

The control model is given by

$$\{E, A, (A(i) \subseteq A, i \in E), Q(u, j|i, a), c(i, a)\} \tag{1}$$

where E is a Borel state space with a Borel σ-algebras $\mathcal{B}(E)$; A is a Borel action space; $A(i)$ denotes the set of admissible actions at state $i \in E$; The set of pairs of states and actions is denoted by $K := \{(i, a)|i \in E, a \in A(i)\}$. $Q(\cdot, \cdot|x, a)$ denotes the semi-Markov kernel on $R_+ \times E$ given K with $R_+ := [0, \infty)$. For any $u \in R_+, D \in \mathcal{B}(S)$, $Q(u, D|i, a)$ represents the joint probability that the sojourn time in state i does not exceed $u \in [0, \infty)$ and the state i transitions into the set D if action a is taken.

The semi-Markov kernel has the following properties:

(a) $Q(\cdot, j|i, a)$ is a non-decreasing, right continuous function from $[0, \infty)$ to $[0, 1]$ with $Q(0, j|i, a) = 0$ for any fixed $(i, a) \in K, j \in E$.

(b) $Q(u, \cdot|i, a)$ is a sub-stochastic kernel on E given K for any fixed $u \in [0, \infty)$.

(c) $p(j|i, a) := \lim_{u\to\infty} Q(u, j|i, a)$ is a stochastic kernel on E given K.

The cost function $c(i, a)$ is a nonnegative real-valued function on K.

The evolution of this system is described as follows: At an initial decision epoch $s_0 = 0$, based on the initial state i_0 and planning horizon t_0, the decision maker chooses an action $a_0 \in A(i_0)$. As a result of this chosen action, the system stays at the state i_0 until time s_1, at which point the system jumps into a new state $i_1 \in D$ with the transition law $Q(s_1, D|i_0, a_0)$, and pays the cost $c(i_0, a_0)(s_1 - s_0)$. Then, the next decision epoch s_1 arrives, and based on the current state i_1, the planning horizon $t_1 = [t_0 - (s_1 - s_0)]_+$, the previous state i_0, and the planning horizon t_0, the decision marker chooses an action $a_1 \in A(i_1)$, where $[x]_+ = \max(0, x)$. The system continues to evolve in the same way, and produces along a vector

$$h_k := (i_0, t_0, a_0, s_1, i_1, t_1, a_1, \ldots, s_k, i_k, t_k), \tag{2}$$

which is called an admissible history up to the kth decision epoch. In (2), s_k $(k \geq 1)$ represents the kth decision epoch, i_{k-1} denotes the state of the system on $[s_{k-1}, s_k)$, a_{k-1} is the action taken by the decision marker at time s_{k-1}. $\theta_k := s_k - s_{k-1}$ represents the sojourn time at state i_{k-1}, which can be made to follow any probability distribution. t_k is the planning horizon at time s_k and is defined by

$$t_k := [t_{k-1} - \theta_k]_+. \tag{3}$$

Moreover, the state process is assumed to be absorbed in an isolated point $\Delta \notin E$ after $s_\infty := \lim_{k\to\infty} s_k$, $c(\Delta, a_\infty) :\equiv 0$, $A(\Delta) := \{a_\infty\}$, and $A_\infty := A \cup \{a_\infty\}$, where a_∞ is an isolated point.

Let H_k denote the sets of all admissible histories h_k defined by

$$H_0 := E \times R_+, \ H_1 := E \times R_+ \times A \times (0, +\infty] \times E \times R_+, \text{ and}$$
$$H_k := E \times R_+ \times A \times ((0, +\infty] \times E \times R_+ \times A)^{k-1} \times (0, +\infty] \times E \times R_+ \text{ for } k \geq 2.$$

To state our optimal control problem rigorously, we define some admissible policies.

Definition 1. A *stochastic history-dependent policy* π is defined by a sequence $\{\pi_k, k \geq 0\}$, where π_k is a stochastic kernel on A given H_k, and such that

$$\pi_k(A(i_k)|h_k) = 1 \quad \forall h_k \in H_k, k = 0, 1, 2 \ldots.$$

The class of all stochastic history-dependent policies is denoted by Π.

Let ϕ be the set of stochastic kernels φ on A given $E \times R_+$ satisfying $\varphi(A(i)|i,t) = 1$ for all $(i,t) \in E \times R_+$. Denote by F the set of measurable functions $f : E \times R_+ \to A$ such that $f(i,t) \in A(i)$ for all $(i,t) \in E \times R_+$.

Definition 2. A policy $\pi = \{\pi_k\} \in \Pi$ is called *stochastic Markov* if $\pi_k(\cdot|h_k) = \varphi_k(\cdot|i_k, t_k)$ for some stochastic kernels $\varphi_k \in \phi$ with $k \geq 0$ and $h_k \in H_k$. Such a stochastic Markov policy is denoted by $\pi = \{\varphi_k\}$ for simplicity.

A stochastic Markov policy $\pi = \{\varphi_k\}$ is said to be *stochastic stationary* if each φ_k is independent of k. Such a stationary policy is denoted by $\pi = \{\varphi\}$ for simplicity.

A stochastic Markov policy $\pi = \{\varphi_k\}$ is said to be *deterministic Markov* if $\varphi_k(\cdot|i_k, t_k)$ is concentrated at $f_k(i_k, t_k) \in A(i_k)$ for some measurable function f_k from $E \times R_+$ to $A(i_k)$ with $k \geq 0$ and $(i_k, t_k) \in E \times R_+$.

A deterministic Markov policy $\pi = \{f_k\}$ is said to be *deterministic stationary* if each f_k is independent of k. Such a stationary policy is denoted by f for simplicity.

Let $\Pi_{RM}, \Pi_{RS}, \Pi_{DM}$ and Π_{DS} denote the class of all stochastic Markov, the class of stochastic stationary, deterministic Markov, and deterministic stationary policies, respectively. It is clear that $\phi = \Pi_{RS} \subset \Pi_{RM} \subset \Pi$ and $F = \Pi_{DS} \subset \Pi_{DM} \subset \Pi$.

To ensure the rationality of the optimal control problem, we need to construct the probability space as follows: The sample space Ω is defined by $\Omega := \{(i_0, t_0, a_0, s_1, i_1, t_1, a_1, \ldots, s_k, i_k, t_k, a_k, \ldots,)|\ i_0 \in E, t_0 \in R_+, a_0 \in A, s_l \in (0, \infty], i_l \in E, t_l \in R_+, a_l \in A$ for each $1 \leq l \leq k, k \geq 1\}$. The sample space Ω endowed with its Borel σ-algebra $\mathcal{F}$. For each $\omega := (i_0, t_0, a_0, s_1, i_1, t_1, a_1, \ldots, s_k, i_k, t_k, a_k, \ldots) \in \Omega, k \geq 1$, we define some random variables on $(\Omega, \mathcal{F})$ as follows: $S_0(\omega) := 0$, $X_0(\omega) := i_0, T_0(\omega) := t_0, A_0(\omega) = a_0, S_k(\omega) := s_k, \Theta_k(\omega) := \theta_k, X_k(\omega) := i_k, T_k(\omega) := t_k, A_k(\omega) := a_k, S_\infty(\omega) := \lim_{k\to\infty} S_k(\omega)$. For simplicity, the argument ω will be omitted from now on. Hence, the state process $\{x_s, s \geq 0\}$ and the action process $\{A_s, s \geq 0\}$ are defined by

$$x_s := \sum_{k\geq 0} I_{\{S_k \leq s < S_{k+1}\}} i_k + \Delta I_{\{s \geq S_\infty\}}, \tag{4}$$

$$A_s := \sum_{k\geq 0} I_{\{S_k \leq s < S_{k+1}\}} a_k + a_\Delta I_{\{s \geq S_\infty\}}, \tag{5}$$

where I_D represents the indicator function on a set D.

For any $(i,t) \in E \times R_+$ and $\pi \in \Pi$, by the Ionescu Tulcea theorem (e. g., Proposition 7.45 in [1]), there exist a unique probability space $(\Omega, \mathcal{F}, P^\pi_{(i,t)})$ and a stochastic process $\{x_s, A_s, s \geq 0\}$ such that for each $k \geq 0$,

$$P^\pi_{(i,t)}(A_k \in \Gamma|h_k) = \pi_k(\Gamma|h_k), \tag{6}$$

$$P^\pi_{(i,t)}(\theta_{k+1} \leq u, X_{k+1} \in D|S_0, X_0, T_0, A_0, \ldots, S_k, X_k, T_k, A_k) = Q(u, D|X_k, A_k). \tag{7}$$

The expectation operator is denoted by $\mathbb{E}^{\pi}_{(i,t)}$ associated with $P^{\pi}_{(i,t)}$.

To ensure that the explosion of state process, we need to impose the following basic assumption, which has been used previously in MDPs, see, for instance, [10, 16] for CTMDPs and [13–15] for SMDPs.

Assumption 1. For any $\pi \in \Pi, (i,t) \in E \times [0,T]$, $P^{\pi}_{(i,t)}(S_{\infty} = \infty) = 1$.

Assumption 1 means that the state process $\{x_s, s \geq 0\}$ is non-explosive, that is the state process $\{x_s, s \geq 0\}$ cannot jump infinitely many times during any finite horizon. Moreover, the following lemma gives an easily verifiable sufficient condition on the semi-Markov kernel ensuring the validity of Assumption 1.

Lemma 1. If there exist some constants δ and $\varepsilon_0 > 0$ such that

$$Q(\delta, E|i,a) \leq 1 - \varepsilon_0, \tag{8}$$

for all $(i,a) \in K$, then Assumption 1 holds.

P r o o f. It follows from Proposition 2.1 in [14]. □

Remark 1. The condition in Lemma 1 is usually called the "standard regular condition", and which is widely used in SMDPs [13–15].

For any fixed $T \in R_+$, $i \in E$ and $\pi \in \Pi$, the finite horizon exponential cost criterion $V^{\pi}(i,T)$ of SMDPs is defined by

$$V^{\pi}(i,T) := E^{\pi}_{(i,T)}\Big(e^{-\gamma \int_0^T c(x_s, A_s)\mathrm{d}s}\Big), \tag{9}$$

where $\gamma > 0$ denotes the risk aversion coefficient, which shows the degree of risk aversion of the decision maker.

Definition 3. A policy $\pi^* \in \Pi$ is said to be an *optimal policy* if

$$V^{\pi^*}(i,T) = \sup_{\pi \in \Pi} V^{\pi}(i,T), i \in E. \tag{10}$$

The value function is given by

$$V^*(i,T) := \sup_{\pi \in \Pi} V^{\pi}(i,T), i \in E. \tag{11}$$

3. MAIN RESULTS

The aim of the present section is to exhibit the main results on the problem of minimizing exponential cost for SMDPs on finite horizon.

Notation: For any policy $\pi \in \Pi$ and initial state $i \in E$, we define the expected exponential cost during a planning horizon$[0,t]$ as follows: $V^{\pi}(i,t) := E^{\pi}_{(i,t)}(e^{-\gamma \int_0^t c(x_s,A_s)\,\mathrm{d}s})$, where $0 \le t \le T$.

Let

$$V^*(i,t) := \sup_{\pi\in\Pi} V^{\pi}(i,t) \quad \forall (i,t) \in E \times [0,T]. \tag{12}$$

Let $\mathcal{V}_m$ be the set of all Borel-measurable functions $V : E \times [0,T] \to [0,1]$.

For any $(i,t) \in E \times [0,T], V \in \mathcal{V}_m$, $\varphi \in \phi$, and $a \in A(i)$, we define the operators $L^{\varphi}V$ and LV by:

$$\begin{aligned} L^aV(i,t) &:= e^{-\gamma c(i,a)t}(1 - D(t|i,a)) \\ &\quad + \int_E \int_0^t e^{-\gamma c(i,a)u} V(j,t-u) Q(\mathrm{d}u, \mathrm{d}j|i,a) \qquad (13) \\ L^{\varphi}V(i,t) &:= \int_A \varphi(\mathrm{d}a|i,t) L^aV(i,t) \qquad (14) \\ LV(i,t) &:= \sup_{A(i)} L^aV(i,t), \qquad (15) \end{aligned}$$

where $D(t|i,a) := \int_E Q(t,\mathrm{d}j|i,a)$.

Moreover, we iteratively define the operators $(L^nV, n \ge 1)$, $((L^{\varphi})^nV, n \ge 1)$ by

$$L^1V = LV, L^{n+1}V = L(L^nV),\ (L^{\varphi})^1V = L^{\varphi}V,\ (L^{\varphi})^{n+1}V = L^{\varphi}((L^{\varphi})^nV),\ n \ge 1.$$

Let $\mathcal{U}_m$ be the set of all Borel-measurable functions $U : E \times [0,T] \to [-1,1]$. For any $(i,t) \in E \times [0,T]$, $U \in \mathcal{U}_m$, $\varphi \in \phi$, and $a \in A(i)$, we define the operators $\widetilde{L}^{\varphi}V$, $((\widetilde{L}^{\varphi})^nV, n \ge 1)$ by:

$$\begin{aligned} \widetilde{L}^aU(i,t) &:= \int_E \int_0^t e^{-\gamma c(i,a)u} U(j,t-u) Q(\mathrm{d}u, \mathrm{d}j|i,a) \qquad (16) \\ \widetilde{L}^{\varphi}U(i,t) &:= \int_A \varphi(\mathrm{d}a|i,t) \widetilde{L}^aU(i,t) \qquad (17) \\ (\widetilde{L}^{\varphi})^{n+1}U &= \widetilde{L}^{\varphi}((\widetilde{L}^{\varphi})^nU). \qquad (18) \end{aligned}$$

In order to ensue the existence of the optimal policy, we need to establish the following compact-continuity condition, which is satisfied for the finite set $A(i)$ with $i \in E$. See, for instance, [10, 11, 13].

Assumption 2. **(a)** For any $i \in E$, $A(i)$ is compact.

(b) For each fixed $V \in \mathcal{V}_m$, $\int_E \int_0^t e^{-\gamma c(i,a)u} V(j,t-u) Q(\mathrm{d}u, \mathrm{d}j|i,a)$ is is upper semicontinuous and inf-compact on K.

In the following, we give some properties of the operator L.

Lemma 2. Under Assumptions 1 and 2, the following assertions hold:

(a) For $V, G \in \mathcal{V}_m$ such that $V \geq G$, we have $L^a V(i,t) \geq L^a G(i,t)$, $LV(i,t) \geq LG(i,t)$ for any $a \in A(i), (i,t) \in E \times [0,T]$.

(b) For $V \in \mathcal{V}_m$, there exists a policy $f \in F$ satisfying $LV(i,t) = L^f V(i,t)$ for any $(i,t) \in E \times [0,T]$.

P r o o f. (a) It is a straightforward consequence of the definition of the operators L^a and L.

(b) Under Assumptions 1 and 2, the measurable selection theorem (proposition D.5 in [11]) provides the existence of a policy $f \in F$ satisfying $L^f V(i,t) = LV(i,t) = \sup_{a \in A(i)} L^a V(i,t)$ for $V \in \mathcal{V}_m, (i,t) \in E \times [0,T]$. □

For any given $(i,t) \in E \times [0,T]$, $\pi \in \Pi$, based on the non-explosion of state process $\{x_s, s \geq 0\}$, the nonnegativity of the cost rate, the continuity of probability measures and the monotone convergence theorem, we can rewrite $V^\pi(i,t)$ as follows:

$$\begin{aligned} V^\pi(i,t) &= E^\pi_{(i,t)}\Big(e^{-\gamma \int_0^t c(x_s,A_s)\mathrm{d}s}\Big) \\ &= E^\pi_{(i,t)}\Big(e^{-\gamma \sum_{m=0}^{\infty} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s,A_s)\mathrm{d}s}\Big) \\ &= \lim_{n\to\infty} E^\pi_{(i,t)}\Big(e^{-\gamma \sum_{m=0}^{n} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s,A_s)\mathrm{d}s}\Big). \end{aligned}$$

Then, the sequence $\{V^\pi_n(i,t), n = -1, 0, 1, \ldots\}$ is defined as follows:

$$\begin{aligned} V^\pi_{-1}(i,t) &:= 1, \\ V^\pi_n(i,t) &:= E^\pi_{(i,t)}\Big(e^{-\gamma \sum_{m=0}^{n} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s,A_s)\mathrm{d}s}\Big) \end{aligned} \tag{19}$$

for all $(i,t) \in E \times [0,T]$. It is clear that $V^\pi_n(i,t) \geq V^\pi_{n+1}(i,t), n \geq -1$ and $\lim_{n\to\infty} V^\pi_n(i,t) = V^\pi(i,t)$.

Proposition 1. For each $(i,t) \in E \times [0,T]$ and $\pi = \{\pi_0, \pi_1, \ldots\} \in \Pi$, there is a policy $\pi' = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$ which satisfies $V^\pi(i,t) = V^{\pi'}(i,t)$.

P r o o f. The proof of Proposition 1 follows from the same arguments as in the proof of Proposition 2.2 in [14]. □

This Proposition states that it suffices to find optimal policies for our optimality problem 10 in the family randomized Markov policies. Then, we will restrict our attention to the case of randomized Markov policies.

The following lemma is required to establish the optimality equation.

Lemma 3. Suppose that Assumptions 1 and 2 hold. For any $(i,t) \in E \times [0,T]$, $n \geq -1$, and $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$, the following assertions hold.

(a) $V_n^\pi \in \mathcal{V}_m$ and $V^\pi \in \mathcal{V}_m$.

(b) $V_{n+1}^\pi(i,t) = L^{\varphi_0} V_n^{{}^1\pi}(i,t)$ and $V^\pi(i,t) = L^{\varphi_0} V^{{}^1\pi}(i,t)$, where ${}^1\pi := \{\varphi_1, \varphi_2, \ldots\}$ being the 1-shift policy of π.

(c) In particular, for any $f \in F$, $V_{n+1}^f(i,t) = L^f V_n^f(i,t)$ and $V^f(i,t) = L^f V^f(i,t)$.

P r o o f. (a) For any $(i,t) \in E \times [0,T], \pi \in \Pi_{RM}$, we prove (a) by induction. Obviously, $V_{-1}^\pi(i,t) = 1 \in \mathcal{V}_m$, and part (a) is valid for $n = -1$. Assume that part (a) holds for $-1 < k \leq n$. it follows from (7) and the property of conditional expectation, we have

$$
\begin{aligned}
V_{k+1}^\pi(i,t) &= E_{(i,t)}^\pi\Big(e^{-\gamma \sum_{m=0}^{k+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s, A_s)\mathrm{d}s}\Big) \\
&= E_{(i,t)}^\pi[E_{(i,t)}^\pi[e^{-\gamma \sum_{m=0}^{k+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s, A_s)\mathrm{d}s} | S_0, x_{S_0}, T_0, a_0, S_1, x_{S_1}, T_1]] \\
&= \int_A \varphi_0(\mathrm{d}a|i,t) \int_E \int_0^{+\infty} E_{(i,t)}^\pi\Big(e^{-\gamma \sum_{m=0}^{k+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s, A_s)\mathrm{d}s} | S_0 = 0, \\
&\quad x_{S_0} = i, T_0 = t, a_0 = a, S_1 = u, x_{S_1} = j, T_1 = [t-u]^+\Big) Q(\mathrm{d}u, \mathrm{d}j|i,a) \\
&= \int_A \varphi_0(\mathrm{d}a|i,t) \int_E \int_t^{+\infty} E_{(i,t)}^\pi\Big(e^{-\gamma \int_0^t c(x_s, A_s)\mathrm{d}s} | S_0 = 0, x_{S_0} = i, \\
&\quad T_0 = t, a_0 = a, S_1 = u, x_{S_1} = j, T_1 = [t-u]^+\Big) Q(\mathrm{d}u, \mathrm{d}j|i,a) \\
&\quad + \int_A \varphi_0(\mathrm{d}a|i,t) \int_E \int_0^t E_{(i,t)}^\pi\Big(e^{-\gamma \sum_{m=0}^{k+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} c(x_s, A_s)\mathrm{d}s} | S_0 = 0, \\
&\quad x_{S_0} = i, T_0 = t, a_0 = a, S_1 = u, x_{S_1} = j, T_1 = [t-u]^+\Big) Q(\mathrm{d}u, \mathrm{d}j|i,a) \\
&= \int_A \varphi_0(\mathrm{d}a|i,t) \Big[e^{-\gamma c(i,a)t}(1 - D(t|i,a)) \\
&\quad + \int_E \int_0^t e^{-\gamma c(i,a)u} E_{(j,t-u)}^{{}^1\pi}\Big(e^{-\gamma \sum_{m=0}^{k} \int_{S_m \wedge (t-u)}^{S_{m+1} \wedge (t-u)} c(x_s, A_s)\mathrm{d}s}\Big) Q(\mathrm{d}u, \mathrm{d}j|i,a)\Big] \\
&= \int_A \varphi_0(\mathrm{d}a|i,t) \Big[e^{-\gamma c(i,a)t}(1 - D(t|i,a)) \\
&\quad + \int_E \int_0^t e^{-\gamma c(i,a)u} V_k^{{}^1\pi}(j, t-u) Q(\mathrm{d}u, \mathrm{d}j|i,a)\Big] \\
&:= L^{\varphi_0} V_k^{{}^1\pi}(j, t-u).
\end{aligned}
$$

Thus, by the induction hypothesis and (20), we deduce that $V_n^\pi(i,t)$ is measurable and that $V_n^\pi \in \mathcal{V}_m$ for all $n \geq -1$. Since the limit of a sequence of measurable functions is still measurable, we obtain $\lim_{n\to\infty} V_n^\pi = V^\pi \in \mathcal{V}_m$.

(b) For any $(i,t) \in E \times [0,T], n \geq -1$, from part (a), we know that $V_{n+1}^\pi(i,t) = L^{\varphi_0} V_n^{{}^1\pi}(i,t)$. Letting $n \to \infty$, and invoking the monotone convergence theorem, we

obtain

$$V^{\pi}(i,t) = L^{\varphi_0} V^{^1\pi}(i,t).$$

(c) This part follows from part (a) and (b). □

Remark 2. The Lemma 3 provides an algorithm to calculate the function $V^f(i,t)$ as follows: letting $V^f_{-1}(i,t) := 1$, Then $V^f_{n+1}(i,t) = L^f V^f_n(i,t)$ and $V^f(i,t) = \lim_{n\to\infty} V^f_n(i,t)$ for any $(i,t) \in E \times [0,T]$, $f \in F$ and $n \geq -1$.

In what follows, we establish the optimality equation, and prove the existence of the optimal policies.

Theorem 1. Under Assumptions 1 and 2, for any $(i,t) \in E \times [0,T]$, let $V^*_{-1}(i,t) := 1$, $V^*_{n+1}(i,t) := LV^*_n(i,t),\ n \geq -1$. Then, $\lim_{n\to\infty} V^*_n(i,t) = V^*(i,t) \in \mathcal{V}_m$.

P r o o f. For any $(i,t) \in E \times [0,T], n \geq -1$, since $V^*_{-1}(i,t) := 1$, by Lemma 2(a) and the definition of V^*_n, we have $0 \leq V^*_n(i,t) \leq V^*_{n+1}(i,t) \leq 1, V^*_n \in \mathcal{V}_m$ and $\tilde{V} := \lim_{n\to\infty} V^*_n \in \mathcal{V}_m$.

To prove $\tilde{V}(i,t) \geq V^*(i,t)$, we first need to prove $V^*_n(i,t) \geq V^{\pi}_n(i,t)$ by induction for any $\pi \in \Pi_{RM}, (i,t) \in E \times [0,T]$ and $n \geq -1$. Since $V^*_{-1}(i,t) = V^{\pi}_{-1}(i,t) = 1$ for any $\pi \in \Pi_{RM}$, the property holds for $n = -1$. Suppose that $V^*_k(i,t) \geq V^{\pi}_k(i,t)$ for all $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$, $-1 \leq k \leq n$. Then, it follows from the induction hypothesis and Lemma 3(b) that

$$V^*_{k+1}(i,t) = LV^*_k(i,t) \geq LV^{^1\pi}_k(i,t) \geq L^{\varphi_0} V^{^1\pi}_k(i,t) = V^{\pi}_{k+1}(i,t).$$

Thus, by induction, we obtain

$$V^*_n(i,t) \geq V^{\pi}_n(i,t), \tag{20}$$

for all $\pi \in \Pi_{RM}, (i,t) \in E \times [0,T]$ and $n \geq -1$. Letting $n \to \infty$ in (20), we obtain $\tilde{V}(i,t) = \lim_{n\to\infty} V^*_n(i,t) \geq V^{\pi}(i,t)$ for all $\pi \in \Pi$. The arbitrariness of π shows that $\tilde{V}(i,t) \geq V^*(i,t)$.

To prove $\tilde{V}(i,t) \leq V^*(i,t)$ for any $(i,t) \in E \times [0,T], n \geq -1$. Letting $A_n := \{a \in A(i) | L^a V^*_n(i,t) \geq L\tilde{V}(i,t)\}$ and $A^* := \{a \in A(i) | L^a \tilde{V}(i,t) = L\tilde{V}(i,t)\}$. Under Assumption 2, since $V^*_n \downarrow \tilde{V}$, we know that A_n and A^* are nonempty and compact, and $A_n \downarrow A^*$. Then, by using the measurable selection theorem (Theorem B.6 in [24]), we know that is an action $a_n \in A_n$ satisfying $L^{a_n} V^*_n(i,t) = LV^*_n(i,t)$. Hence, the existence of an action $a^* \in A^*$ and a subsequence $\{a_{n_k}\}$ of $\{a_n\}$ satisfying $a_{n_k} \to a^*$ are ensured by the compactness of A_n and $A_n \downarrow A^*$. It follows from Lemma 3(a) that for any given $n \geq 1$,

$$L^{a_{n_k}} V^*_{n_k}(i,t) \leq L^{a_{n_k}} V^*_n(i,t) \quad \forall n_k \geq n.$$

Letting $k \to \infty$ and using the upper semicontinuity condition in Assumption 2 give

$$\tilde{V}(i,t) \leq L^{a^*} V^*_n(i,t).$$

Letting $n \to \infty$, we obtain

$$\tilde{V}(i,t) \leq L^{a^*}\tilde{V}(i,t) \leq L\tilde{V}(i,t).$$

Hence, using Lemma 2(b), we see that there exists a stationary policy $f \in F$ such that

$$\tilde{V}(i,t) \leq L\tilde{V}(i,t) = L^f\tilde{V}(i,t) \leq (L^f)^n\tilde{V}(i,t) \leq (L^f)^n V_{-1}^f(i,t) = V_{n-1}^f(i,t).$$

Letting $n \to \infty$, and using (19), we have $\tilde{V}(i,t) \leq V^f(i,t) \leq V^*(i,t)$. Then, $\tilde{V}(i,t) = V^*(i,t)$.

□

Theorem 2. Suppose that Assumptions 1 and 2 hold. Then, for any $(i,t) \in E \times [0,T]$, $f \in \Pi_s, u, v \in \mathcal{V}_m$.

(a) If $u(i,t) - v(i,t) \leq \widetilde{L}^f(u-v)(i,t)$, then $u(i,t) \leq v(i,t)$.

(b) The function $V^f(i,t)$ is the unique solution to the equation $V = L^f V(i,t)$.

Proof. (a) For any $(i,t) \in E \times [0,T], f \in \Pi_s, u, v \in \mathcal{V}_m, u - v \in \mathcal{U}_m$, based on the mathematical induction method, we first prove the following fact:

$$(\widetilde{L}^f)^n(u-v)(i,t) \leq P_{(i,t)}^f(S_n < t), n \geq 1. \tag{21}$$

When $n = 1$, since $u, v \in \mathcal{V}_m, c(i,f) \geq 0$, by the definition of the operator $\widetilde{L}$, we have

$$\begin{aligned}
\widetilde{L}^f(u-v)(i,t) &= \int_E \int_0^t e^{-\gamma c(i,f)s}(u-v)(j,t-s)Q(\mathrm{d}s,\mathrm{d}j|i,f) \\
&\leq \int_0^t D(\mathrm{d}s|i,f) \\
&= P_{(i,t)}^f(S_1 \leq t).
\end{aligned}$$

Assume the fact (21) is satisfied for $n = k$. On the basis of the induction hypothesis, we obtain

$$\begin{aligned}
(\widetilde{L}^f)^{k+1}(u-v)(i,t) &= \widetilde{L}^f(\widetilde{L}^f)^k(u-v)(i,t) \\
&= \int_E \int_0^t e^{-\gamma c(i,f)s}(\widetilde{L}^f)^k(u-v)(j,t-s)Q(\mathrm{d}s,\mathrm{d}j|i,f) \\
&\leq \int_E \int_0^t e^{-\gamma c(i,f)s} P_{(j,t-s)}^f(S_k \leq t-s)Q(\mathrm{d}s,\mathrm{d}j|i,f) \\
&\leq \int_E \int_0^t P_{(j,t-s)}^f(S_k \leq t-s)Q(\mathrm{d}s,\mathrm{d}j|i,f).
\end{aligned} \tag{22}$$

Further, according to the properties of conditional expectation, we have

$$\begin{aligned}
& P^f_{(i,t)}(S_{k+1} \le t) \\
= \; & E^f_{(i,t)}[I_{\{S_{k+1}\le t\}}] \\
= \; & E^f_{(i,t)}[E^f_{(i,t)}[I_{\{S_{k+1}\le t\}}|S_0, x_{S_0}, T_0, S_1, x_{S_1}, T_1] \\
= \; & \int_E \int_0^t P^f_{(i,t)}\Big(S_{k+1} \le t|S_0 = 0, \\
& x_{S_0} = i, T_0 = t, S_1 = s, x_{S_1} = j, T_1 = [t-s]^+\Big) Q(\mathrm{d}s, \mathrm{d}j|i, f) \\
= \; & \int_E \int_0^t P^f_{(j,t-s)}(S_k \le t - s) Q(\mathrm{d}s, \mathrm{d}j|x, f),
\end{aligned}$$

which together with (22) and the mathematical induction gives

$$u(i,t) - v(i,t) \le (\widetilde{L}^f)^n (u-v)(i,t) \le P^f_{(i,t)}(S_n \le t) \quad \forall n \ge 1. \tag{23}$$

On the basis of Assumption 1, as $n \to \infty$, we have

$$u(i,t) - v(i,t) \le \lim_{n\to\infty} P^f_{(i,t)}(S_n \le t) = P^f_{(i,t)}(S_\infty \le t) = 0.$$

Thus, $u(i,t) \le v(i,t)$ for $(i,t) \in E \times [0,T]$.

(*b*) For any $(i,t) \in E \times [0,T]$, $f \in F$, $V^f(i,t) \in \mathcal{V}_m$ satisfying the equation $V = L^f V$ is proved in Lemma 2(b). If the equation $V = L^f V$ has another solution $U(i,t)$ on $E \times [0,T]$, then $U(i,t) - V^f(i,t) = \widetilde{L}^f(U - V^f)(i,t)$. This implies $U(i,t) = V^f(i,t)$ by the conclusion in part (a). □

Theorem 3. Suppose that Assumptions 1 and 2 hold, for any $(i,t) \in E \times [0,T], n \ge -1$.

(a) V^* is the unique solution in $\mathcal{V}_m$ to the optimality equation $V = LV$.

(b) There exists a policy $f^* \in F$ satisfying $V^* = L^{f^*}V^*$, and such a policy f^* is optimal.

P r o o f. (*a*) For any $(i,t) \in E \times [0,T], \pi \in \Pi_{RM}$, from Lemma 3(b), we have

$$V^\pi(i,t) = L^{\varphi_0} V^{1\pi}(i,t) \ge L^{\varphi_0} V^*(i,t) \ge LV^*(i,t),$$

which together with the arbitrariness of π gives that $V^*(i,t) \ge LV^*(i,t)$.

On the other hand, for each $(i,t) \in E \times [0,T]$ and $a \in A(i)$, by the definition of V^*_n, we have

$$V^*_{n+1}(i,t) = LV^*_n(i,t) \le L^a V^*_n(i,t),$$

which together with the monotone convergence theorem implies $V^*(i,t) \le L^a V^*(i,t)$. Hence, the arbitrariness of $a \in A(i)$ gives $V^*(i,t) \le LV^*(i,t)$. Thus, $V^* = LV^*$.

Since $V^* = LV^*$ for any $(i,t) \in E \times [0,T]$, using Lemma 2, we know that there exists an $f^* \in F$ such that $V^*(i,t) = L^{f^*}V^*(i,t)$. Moreover, suppose that G is a

another solution in $\mathcal{V}_m$ to the equation $Lu = u$. Similarly, we know that there is a policy $f \in F$ such that $G^*(i,t) = L^f G^*(i,t)$ for any $(i,t) \in E \times [0,T]$. Then, we have $V^* - G \le L^{f^*}(V^* - G)$ and $G - V^* \le L^f(G - V^*)$, which together with Theorem 2 give $G = V^*$. This concludes the proof of part (a).

(*b*) For any $(i,t) \in E \times [0,T]$, it follows from Lemma 2 that there exists an $f^* \in F$ such that $V^*(i,t) = L^{f^*}V^*(i,t)$. Moreover, since $V^* \in \mathcal{V}_m$, using Lemma 3, we obtain

$$V^* = \lim_{n\to\infty} (L^{f^*})^n V^* \le \lim_{n\to\infty} (L^{f^*})^n V_{-1}^{f^*} = \lim_{n\to\infty} V_{n-1}^{f^*} = V^{f^*},$$

which implies $V^* = V^{f^*}$. Thus, f^* is optimal. □

4. THE VALUE ITERATION ALGORITHM

Based on Theorem 3, for any given error, we establish the following value iterative algorithm to calculate the value function and the ε-optimal policy with finite iteration.

Definition 4. For any $\varepsilon > 0$, a policy $f_\varepsilon^* \in \Pi_s$ is called exponential cost ε-optimal if

$$-\varepsilon \le V^*(i,t) - V^{f_\varepsilon^*}(i,t) \le \varepsilon \quad \forall (i,t) \in E \times [0,T].$$

Theorem 4. Suppose that Assumptions 1 and 2 hold. The sequence $\{V_n^*\}$ is defined in Theorem 3, δ, ε_0 are defined in (8), $0 < \alpha = (1 - \varepsilon_0^k)^{1/k} < 1$, where k is defined as a non-negative integer and satisfies $k > T/\delta$. For every $(i,t) \in E \times [0,T]$, the following assertions hold.

(a) For any sufficiently small $\epsilon > 0$, letting $n_0 = [\log_\alpha \frac{\epsilon(1-\alpha)}{2}]^+ + k$, where $[x]^+$ denotes the largest integer not bigger than $x \in R_+$. Then we get

$$|V^* - V_{n_0}^*| \le \frac{\epsilon}{2}.$$

(b) There exists $f_\epsilon^* \in \Pi_s$ such that $V_{n_0+1}^* = L^{f_\epsilon^*} V_{n_0}^*$, and such policy $f_\epsilon^* \in \Pi_s$ is an ϵ-optimal policy.

P r o o f. (a) For any $(i,t) \in E \times [0,T]$, from Lemma 2, we know that: there is a policy f^* such that $V_{n+1}^* = LV_n^* = L^{f^*}V_n^*$. Then, under Assumption 1, by the proof of Theorem 2 and (21), we have for $n \ge 1$

$$\begin{aligned} |V_n^* - V_{n+1}^*| &= V_n^* - V_{n+1}^* \\ &= (L)^{n+1} V_{-1}^* - L^{n+1} V_0^* \\ &\le (\widetilde{L}^{f^*})^{n+1}(V_{-1}^* - V_0^*) \\ &\le P_{(i,t)}^{f^*}(S_{n+1} < t). \end{aligned} \tag{24}$$

Letting

$$F_\delta(t) = \begin{cases} 0, & t < 0, \\ 1 - \varepsilon_0, & 0 \le t \le \delta, \\ 1, & t > \delta, \end{cases}$$

where δ, ε_0 are defined in (8).

Moreover, it follows from the proof of Theorem 1 in [22], we know that

$$F_\delta^{(n)}(t) \leq (1-\varepsilon_0^k)^{[n/k]^+}, \ \forall n > k, \tag{25}$$

Then, under Assumption 1, since $Q(\delta, E|i,a) \leq F_\delta(t)$, by (7),(24) and (25), we obtain

$$|V_n^* - V_{n+1}^*| \leq F_\delta^{(n+1)}(t) \leq (1-\varepsilon_0^k)^{[n+1/k]^+}. \tag{26}$$

For any given $\epsilon > 0$, letting $0 < \alpha := (1-\varepsilon_0^k)^{1/k} < 1, n_0 = [\log_\alpha \frac{\epsilon(1-\alpha)}{2}]^+ + k$, from (26), we have

$$\begin{aligned} & |V_{n_0}^*(i,t) - V^*(i,t)| \\ = & \sum_{l=1}^{+\infty} [V_{n_0+l-1}^*(i,t) - V_{n_0+l}^*(i,t)] \\ < & \sum_{l=1}^{+\infty} \alpha^{n_0+l-k} \\ = & \frac{\alpha^{n_0+1-k}}{1-\alpha} \\ < & \frac{\epsilon}{2}. \end{aligned} \tag{27}$$

(b) For any $(i,t) \in E \times [0,T]$, it follows from the measurable selection theorem (proposition D.5 in [11]) and Theorem 3 that, there exists an $f_\epsilon^* \in \Pi_s$ satisfying

$$V_{n_0+1}^*(i,t) = LV_{n_0}^*(i,t) = L^{f_\epsilon^*} V_{n_0}^*(i,t).$$

According to the similar argument in (26) and Theorem 2, we have

$$\begin{aligned} & |V_{n_0}^* - V^{f_\epsilon^*}| \\ = & |(L)^{n_0+1} V_{-1}^* - (L^{f_\epsilon^*})^{n_0+1} V^{f_\epsilon^*}| \\ \leq & |(\widetilde{L}^{f_\epsilon^*})^{n_0+1}(V_{-1}^* - V^{f_\epsilon^*})| \\ \leq & F_\delta^{(n_0+1)}(t) \\ < & \alpha^{n_0+1-k} \\ < & \frac{1-\alpha}{2}\epsilon, \end{aligned}$$

which gives that $|V^* - V^{f_\epsilon^*}| \leq |V^* - V_{n_0}^*| + |V_{n_0}^* - V^{f_\epsilon^*}| \leq \frac{2-\alpha}{2}\epsilon < \epsilon$, and so (b) follows. □

The value iteration algorithm procedure:

Step 1: For any $(i,t) \in E \times [0,T]$, set $V_{-1}^*(i,t) = 1$.

Step 2: For all $n \geq 0, a \in A(i)$, by Theorem 3, the functions $L^a V_n^*(i,t)$ and $V_{n+1}^*(i,t)$ are computed as follows:

$$\begin{aligned} L^a V_n^*(i,t) &= e^{-\gamma c(i,a)t}(1 - D(t|i,a)) \\ &\quad + \int_E \int_0^t e^{-\gamma c(i,a)u} V_n^*(j, t-u) Q(\mathrm{d}u, \mathrm{d}j|i,a) \\ &\approx e^{-\gamma c(i,a)t}(1 - D(t|i,a)) \\ &\quad + \int_E \sum_{l=1}^{m-1} \frac{1}{2}\Big[V_n^*\Big(j, t-lh\Big) e^{-\gamma c(i,a)lh} Q(lh, \mathrm{d}j|i,a) \\ &\quad + V_n^*\Big(j, t-(l+1)h\Big) e^{-\gamma c(i,a)(l+1)h} Q((l+1)h, \mathrm{d}j|i,a)\Big] h. \quad (28) \\ V_{n+1}^*(i,t) &\approx \sup_{a \in A(i)} \{L^a V_n^*(i,t)\}, \quad (29) \end{aligned}$$

where the step length h satisfies $mh = t$ and $l \leq m$ with $l, m \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of natural numbers.

Step 3: For any sufficiently small $\epsilon > 0$, when $n = n_0 = [\log_\alpha \frac{\epsilon(1-\alpha)}{2}]^+ + k$, the iteration stops, and $V_{n_0}^*$ is accepted as a good approximation of the value function V^*. Hence, the existence of the ϵ-optimal policy f_ϵ^* is determined by Theorem 4.

Remark 3. The formula (28) is due to the trapezoidal integration rule [18] given as follows:

$$\int_a^b g(x)\mathrm{d}x \approx \sum_{l=0}^{m-1} \frac{g(a+lh) + g(a+(l+1)h)}{2} h, \quad (30)$$

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

5. EXAMPLE

In this section, we apply our main results obtained to a machine maintenance problem, in which we exhibit the usefulness of the value iteration algorithm in computing the value function and optimal policies.

Example 1. Consider a machine maintenance system with three states: the bad state, medium and good state, which are denoted by 0, 1 and 2. When the system is in a state $i \in \{0,1,2\}$, the decision-maker can choose a rapid maintenance action a_{i1} or a general maintenance action a_{i2}, according to the actual maintenance cost at rate $c(i,a_{i1})$ or $c(i,a_{i2})$. Assuming that the transition mechanism of this maintenance system is primarily interested in the evolution of the control model of SMDPs (1). Additionally, the model parameter are given as follows:

The state space $E = \{0, 1, 2\}$ and admissible action space $A(0) = \{a_{01}\}, A(1) = \{a_{11}, a_{12}\}, A(2) = \{a_{21}, a_{22}\}$, the risk-sensitivity coefficient $\gamma = 1$. The transition probabilities are given as follows

$$\begin{array}{lll} p(0|1, a_{11}) = 0.5, & p(2|1, a_{11}) = 0.5, & p(0|1, a_{12}) = 0.3, \\ p(2|1, a_{12}) = 0.7, & p(0|2, a_{21}) = 0.9, & p(1|2, a_{21}) = 0.1, \\ p(0|2, a_{22}) = 0.3, & p(1|2, a_{22}) = 0.7, & p(0|0, a_{01}) = 1. \end{array} \tag{31}$$

Correspondingly, for any $u \in [0, +\infty)$, the semi-Markov decision kernels are given by

$$\begin{array}{ll} Q(u, 0|1, a_{11}) = p(0|1, a_{11})(1 - e^{-0.15u}), & Q(u, 2|1, a_{11}) = p(2|1, a_{11})(1 - e^{-0.15u}), \\ Q(u, 0|1, a_{12}) = p(0|1, a_{12})(1 - e^{-0.08u}), & Q(u, 2|1, a_{12}) = p(2|1, a_{12})(1 - e^{-0.08u}), \\ Q(u, 0|2, a_{21}) = p(0|2, a_{21})(1 - e^{-0.11u}), & Q(u, 1|2, a_{21}) = p(1|2, a_{21})(1 - e^{-0.11u}), \\ Q(u, 0|2, a_{22}) = p(0|2, a_{22})(1 - e^{-0.05u}), & Q(u, 1|2, a_{22}) = p(1|2, a_{22})(1 - e^{-0.05u}), \end{array} \tag{32}$$

and the cost rates are given as follows:

$$c(1, a_{11}) = 0.05,\ c(1, a_{12}) = 0.03,\ c(2, a_{21}) = 0.04,\ c(2, a_{22}) = 0.02,\ c(0, a_{01}) = 0.$$

Our goal is to compute the exponential cost optimal policy by using a value iteration algorithm over the finite horizon $[0, 90]$.

From (32), one can easily verify the sufficient condition of Lemma 1 and thus the validity of Assumption 1 for $\delta = 10, \varepsilon_0 = 0.95$. Moreover, based on the denumerable state space and the finite action space, we know that Assumption 2 is trivially satisfied. Thus, the existence of the value function and the optimal policy is ensured by Theorem 3. Therefore, it follows from (31), $c(0, a_{01}) = 0$ and Theorem 3 that $V^*(0, t) = 1$ for $t \in [0, 90]$. Letting $k = 15, \epsilon = 10^{-4}$ Then, the value iteration algorithm in Theorem 4 can be used to compute the value function $V^*(1, t), V^*(2, t)$ and the ϵ-optimal policy as follows.

Step 1: For $i = 1, 2,\ n = -1,\ t \in [0, 90]$, set $V^*_{-1}(i, t) := 1$.

Step 2: For $i = 1, 2,\ n \geq 0$ and $a \in A(i)$, by Theorem 3, we have

$$\begin{aligned} L^{a_{11}} V^*_n(1, t) &= e^{-0.2t} + 0.5 \times 0.15 \times \int_0^t V^*_n(2, t - u) e^{-0.2u} \mathrm{d}u \\ &\quad + 0.5 \times 0.15 \times \int_0^t e^{-0.2u} \mathrm{d}u, \\ L^{a_{12}} V^*_n(1, t) &= e^{-0.11t} + 0.7 \times 0.08 \times \int_0^t V^*_n(2, t - u) e^{-0.11u} \mathrm{d}u \\ &\quad + 0.3 \times 0.08 \times \int_0^t e^{-0.11u} \mathrm{d}u, \end{aligned}$$

$$
\begin{aligned}
V_{n+1}^*(1,t) &= \max\{L^{a_{11}}V_n^*(1,t), L^{a_{12}}V_n^*(1,t)\},\\
L^{a_{21}}V_n^*(2,t) &= e^{-0.15t} + 0.1\times 0.11\times \int_0^t V_n^*(1,t-u)e^{-0.15u}\mathrm{d}u\\
&\quad +0.9\times 0.11\times \int_0^t e^{-0.15u}\mathrm{d}u,\\
L^{a_{22}}V_n^*(2,t) &= e^{-0.07t} + 0.7\times 0.05\times \int_0^t V_n^*(1,t-u)e^{-0.07u}\mathrm{d}u\\
&\quad +0.3\times 0.05\times \int_0^t e^{-0.07u}\mathrm{d}u,\\
V_{n+1}^*(2,t) &= \max\{L^{a_{21}}V_n^*(2,t), L^{a_{22}}V_n^*(2,t)\}.
\end{aligned}
$$

Step 3: For $i = 1, 2$, when $n = n_0 = 334$, the iteration stops. Then, go to step 4, the value $V_{n_0+1}^*$ is accepted as an approximate value of the value V^*.

Step 4: For $i = 1, 2$, plot out the graphs of the functions $L^aV^*(i,t)$, $V^*(i,t)$ (see Figures 1 – 2).

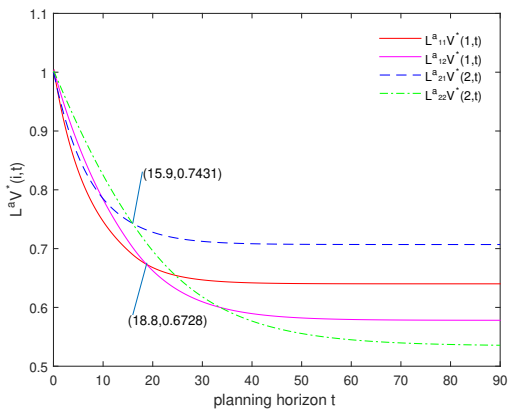

Fig. 1. The function $L^aV^*(i,t)$.

By analyzing the computing procedure and Figures 1 – 2, it is found the following conclusions:

(*a*) In state 1, $L^{a_{11}}V^*(1,t)$ is smaller than $L^{a_{12}}V^*(1,t)$ with the planning horizon $t \in [0, 18.8)$, and $L^{a_{12}}V^*(1,t)$ is smaller than $L^{a_{11}}V^*(1,t)$ with the planning horizon $t \in [18.8, 90]$. This means that the action a_{11} has lower exponential expected costs than the action a_{12} when the planning horizon is $t \in [0, 18.8)$, but the action a_{12} has lower exponential expected costs than the action a_{11} when the planning horizon is $t \in [18.8, 90]$. This implies the decision maker should select the action a_{11} rather than the action a_{12} when the planning horizon $t \in [0, 18.8)$, or select the action a_{12} rather than the action a_{11} if the planning horizon $t \in [18.8, 90]$.

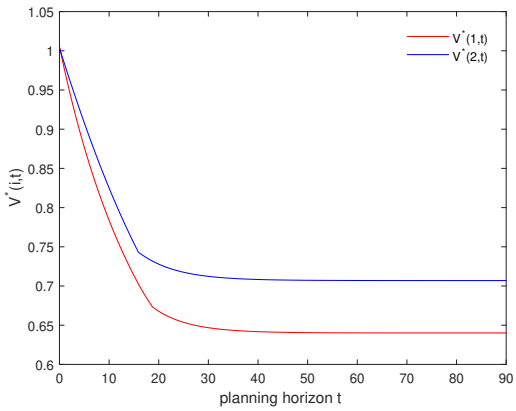

Fig. 2. The value function $V^*(i,t)$.

Similarly, in state 2, the action a_{21} has lower exponential expected costs than the action a_{22} when the planning horizon is $t \in [0, 15.9)$, but the action a_{22} has lower exponential expected costs than the action a_{21} when the planning horizon is $t \in [15.9, 90]$.

(b) From Figures 1 – 2 and part (a), we obtain the following optimal policy f_ϵ^* as follows:

$$f_\epsilon^*(1,t) = \begin{cases} a_{11}, & 0 \le t < 18.8, \\ a_{12}, & 18.8 \le t \le 90, \end{cases} \quad f_\epsilon^*(2,t) = \begin{cases} a_{21}, & 0 \le t < 15.9, \\ a_{22}, & 15.9 \le t \le 90, \end{cases} \tag{33}$$

which satisfy $V^*(i,t) = L^{f_\epsilon^*} V^*(i,t)$ for $i = 1, 2,\ t \in [0, 90]$.

This means that at the initial decision epoch $s_0 = 0$, and according to the initial system state $i_0 \in \{1, 2\}$, the planning horizon $t_0 = 90$ and (33), the decision marker chooses the optimal action $f_\epsilon^*(i_0, t_0) \in A(i_0)$. As a consequence of this action, the system remains in state i_0 until time s_1, at which point the system jumps to another state i_1 with probability $p(i_1|i_0, f_\epsilon^*(i_0, t_0))$. Then, the next decision epoch s_1 arrives, and again based on the current state i_1, the planning horizon $t_1 = [t_0 - s_1]_+$, and (33), the decision marker chooses the optimal action $f_\epsilon^*(i_1, t_1) \in A(i_1)$. Thus, at the decision epoch $s_k, k = 2, 3, \ldots$, the decision maker chooses actions repeatedly in the same way. By Theorem 4, we know that $\pi^* = \{f_\epsilon^*(i_0, t_0), f_\epsilon^*(i_1, t_1), \ldots\}$ is the ϵ-optimal policy.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (Grant No.11961005; 12161007); Guangxi Natural Science Foundation Program (Grant No.2020GXNSFAA297196); Guangxi science and technology base and talent project(Grant No.AD21159005); Foundation of Guangxi Educational Committee (Grant No.KY2022KY0342); PhD research startup foundation of Guangxi University of Science and Technology (Grant No.18Z06).

(Received December 26, 2021)

REFERENCES

[1] D. P. Bertsekas and S. E. Shreve: Stochastic Optimal Control: The Discrete-Time Case. Academic Press, Inc. 1978.

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

[3] N. Baüerle and U. Rieder: More risk-sensitive Markov decision processes. Math. Oper. Res. *39* (2014), 105–120. DOI:10.1287/moor.2013.0601

[4] X. R. Cao: Semi-Markov decision problems and performance sensitivity analysis. IEEE Trans. Automat. Control *48* (2003), 758–769. DOI:10.1109/TAC.2003.811252

[5] R. Cavazos-Cadena and R. Montes-De-Oca: Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards. Appl. Math. *27* (2000), 167–185. DOI:10.4064/am-27-2-167-185

[6] R. Cavazos-Cadena and R. Montes-De-Oca: Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. Math. Methl Oper. Res. *52* (2000), 133–167. DOI:10.1155/S107379280000009X

[7] S. Chávez-Rodríguez, R. Cavazos-Cadena, and H. Cruz-Suárez: Controlled Semi-Markov chains with risk-sensitive average cost criterion. J. Optim. Theory Appl. *170* (2016), 670–686. DOI:10.1007/s10957-016-0916-z

[8] K. J. Chung, M. J. Sobel: Discounted MDP's: distribution functions and exponential utility maximization. SIAM J. Control Optim. *25* (1987), 49–62. DOI:10.1137/0325004

[9] M. K. Ghosh and S. Saha: Risk-sensitive control of continuous time Markov chains. Stoch. Int. J. Probab. Stoch. Process. *86* (2014), 655–675. DOI:10.1080/17442508.2013.872644

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

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

[12] R. A. Howard and J. E. Matheson: Risk-sensitive Markov decision processes. Management Sci. *18* (1972), 356–369. DOI:10.1287/mnsc.18.7.356

[13] Y. H. Huang, Z. T. Lian, and X. P. Guo: Risk-sensitive semi-Markov decision processes with general utilities and multiple criteria. Adv. Appl. Probab. *50* (2018), 783–804. DOI:10.1017/apr.2018.36

[14] Y. H. Huang and X. P. Guo: Finite horizon semi-Markov decision processes with application to maintenance systems. Europ. J. Oper. Res. *212* (2011), 131–140. DOI:10.1016/j.ejor.2011.01.027

[15] 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

[16] H. F. Huo and 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

[17] A. Jaśkiewicz: A note on negative dynamic programming for risk-sensitive control. Oper. Res. Lett. *36* (2008), 531–534. DOI:10.1016/j.mpmed.2008.07.009

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

[19] A. Jaśkiewicz: On the equivalence of two expected average cost criteria for semi Markov control processes. Math. Oper. Res. *29* (2013), 326–338. DOI:10.1002/dmrr.2420

[20] S. C. Jaquette: A utility criterion for Markov decision processes. Manag Sci. *23* (1976), 43–49. DOI:10.5951/AT.23.1.0043

[21] F. Luque-Vasquez and J. A. Minjarez-Sosa: Semi-Markov control processes with unknown holding times distribution under a discounted criterion. Math. Methods Oper. Res. *61* (2005), 455–468. DOI:10.1007/s001860400406

[22] J. W. Mamer: Successive approximations for finite horizon semi-Markov decision processes with application to asset liquidation. Oper. Res. *34* (1986), 638–644. DOI:10.1287/opre.34.4.638

[23] V. Nollau: Solution of a discounted semi-markovian descision problem by successiveoe-varrelaxation. Optimization. *39*, (1997), 85–97. DOI:10.1080/02331939708844273

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

[25] Q. Wei: Continuous-time Markov decision processes with risk-sensitive finite-horizon cost criterion. Math. Oper. Res. *84* (2016), 461–487. DOI:10.1007/s00186-016-0550-4

[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] A. A. Yushkevich: On semi-Markov controlled models with average reward criterion. Theory Probab. Appl. *26* (1982), 808–815. DOI:10.1137/1126089

[28] Y. Zhang: Continuous-time Markov decision processes with exponential utility. SIAM J. Control Optim. *55* (2017), 2636–2666. DOI:10.1137/16m1086261

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