First Passage Exponential Optimality Problem for Semi-Markov Decision Processes

Haifeng Huo(✉) and Xian Wen

Department of School of Science, Guangxi University of Science and Technology, Liuzhou 5451006, China
xiaohuo08ok@163.com, wenxian879@163.com

Abstract. This paper deals with the exponential utility maximization problem for semi-Markov decision process with Borel state and action spaces, and nonnegative reward rates. The criterion to be optimized is the expected exponential utility of the total rewards before the system state enters the target set. Under the regular and compactness-continuity conditions, we establish the corresponding optimality equation, and prove the existence of an exponential utility optimal stationary policy by an invariant embedding technique. Moreover, we provide an iterative algorithm for calculating the value function as well as the optimal policies. Finally, we illustrate the computational aspects of an optimal policy with an example.

Keywords: Semi-Markov decision processes · Exponential utility · First passage time · Value iterative approach · Optimality equation · Optimal policy

AMS(2020) subject classification: Primary 90C40 · Secondary 90C39

1 Introduction

Semi-Markov decision processes (SMDPs), as an important class of stochastic control problems, have been widely studied [1,10,11,15,20,28,31]. The commonly used criteria for SMDPs are the finite horizon expected criterion [8,14,26,28], the expected discounted criterion [1,3,10,13,25,27], and the average criterion [10,23,31–33]. These criteria are linear utility functions of the total rewards (i.e. are risk-neutral), which only focus on the expected total rewards of a system during a fixed or a random horizon, and therefore cannot reflect the decision maker's attitude toward risk.

To exhibit the attitude of a decision maker in the face of risk (i.e. risk-seeking or risk-averse), the risk sensitive criteria, which include the exponential utility criterion, have been considered for discrete-time MDPs (DTMDPs) [2,4–6,21,22], and continuous-time MDPs (CTMDPs) [7,9,30,34]. Specifically, Jaquette [21] first introduced the exponential utility to DTMDPs. For the resulting

A. Piunovskiy and Y. Zhang (Eds.): *Modern Trends in Controlled Stochastic Processes*, ECC 41, pp. 19–37, 2021. https://doi.org/10.1007/978-3-030-76928-4_2

optimization problem, Chung and Sobel [6] established the corresponding optimality equation by means of the Banach fixed point theorem. Cavazos-Cadena and Montes-De-Oca [4,5] gave conditions ensuring the existence of optimal policies for the positive dynamic programming, where the state space is considered to be finite in [4], and denumerable in [5]. Jaśkiewicz [22] considered the Borel state and action spaces, and establish the convergence of the n-stage optimal expected total reward and the existence of an optimal stationary policy. Baüerle and Rieder [2] considered a more general problem than the classic risk sensitive optimization problem, which is called minimizing a certainty equivalent. They solved the optimization problem by an ordinary MDP with extended state space, and proved the existence of an optimal policy under some suitable conditions. For the case of CTMDPs, Ghosh and Saha [7] studied the risk sensitive control in discrete state space. They obtain the value function as a solution to the Hamilton Jacobi Bellman equation, and proved the existence of an optimal Markov control for finite horizon problem, and the existence of an optimal stationary control for infinite horizon problem. Wei [30] dealt with risk sensitive cost criterion for finite horizon CTMDPs with denumerable state space and Borel action space. Under suitable conditions, he proved the existence of the Feynman-Kac formula and an optimal deterministic Markov policy. For the same problem as in [30], Guo, Liu and Zhang [9] investigated the case when the transition and cost rates may be unbounded. They proved that the value function is the unique solution to the optimality equation, and showed the existence of an optimal policy via the Feynman-Kac formula. Few literature [34] applied the uniformization technique to reducing the CTMDPs problem with exponential utility to an equivalent DTMDPs. Recently, Huang, Lian and Guo [17] considered the risk sensitive unconstrained and constrained problems for SMDPs with Borel state space, unbounded cost rates and general utility functions, and proved the existence of the Bellman equation and the optimal policies under some continuity-compactness conditions by using the occupation measure approach.

All of this existing literature shows that all the aforementioned MDPs for the risk-sensitive criterion have two common features: the horizon is finite or infinite, the control model is DTMDPs or CTMDPs. However, such as those encountered in many real world situations, many models in ruin problems [20,29], reliability [20,24], and maintenance [20] are considered with a random horizon, and described as SMDPs. Moreover, compared to DTMDPs and CTMDPs (under stationary policies), SMDPs are more general stochastic optimal models, in which the holding time of the system state can be allowed to follow any arbitrary probability distribution. This is the main reason for considering a random horizon for SMDPs in this paper.

Compared with the existing research work for risk-sensitive SMDPs in [17], this paper has some new features as follows: First, in order to make the conclusion more closely fit the actual situation, we pay more attention to the time horizon is the random first passage time, which is more general than those in [17]. Second, since the random first passage time is considered in our control model, by Remark

4.2 in [17], we know that the occupation measure approach is not suitable for our model, because the definition of the occupation measure is based on the discount factor. Instead, we use a so-called minimum nonnegative solution approach to establish the optimality equation and prove the existence of optimal policies. Third, we are mainly concerned with the calculation and existence of the optimal policies, while the purpose of the works in [17] is to establish the existence condition of the optimal policies. Due to these, we develop a value iteration algorithm to calculate the value function and the optimal policy, which is new and the key feature in our paper.

To the best of our knowledge, the risk-sensitive optimality problem for SMDPs in first passage has not been studied yet.

Motivated by the above discussion, we investigate in this paper the first passage risk-sensitive optimality problems for SMDPs. We focus on both the existence conditions and the computational algorithms of an optimal policy, thus we limit the choice of risk-sensitive criteria to the exponential utility criterion (e.g. [2,6,21,34]), which maximizes the expected exponential utility of the total rewards before the state of system enters the target set. More precisely, in order to ensure the existence of an optimal stationary policy, we impose the standard regular condition to ensure that the state process is non-explosive, which is similar to those given in [13–15,18] for SMDPs (see Lemma 1). Second, compared with [13–15,18], which are mainly limited to denumerable state space and finite action set, we consider more general Borel state and action spaces. Then, we need to introduce a new continuity-compactness condition (see Assumption 2). Under the regular and continuity-compactness conditions, we establish the corresponding optimality equation, and prove that the value function is a solution to this optimality equation. Moreover, we show the existence of an exponential utility optimal stationary policy by using an invariant embedding technique (see Assumption 1). Furthermore, a value iteration algorithm for computing the value function as well as the optimal policies, in a finite number of iterations, is provided. Finally, an example illustrating the computational methodology of an optimal stationary policy and the value function is given.

The rest of this paper is organized as follows. In Sect. 2, we introduce the semi-Markov decision model and state the first passage exponential utility optimality problem. The main optimality results are stated and proved in Sect. 3. In Sect. 4, an example is provided to illustrate the computational aspects of an optimal policy.

2 Model Description

Models of first passage exponential utility SMDPs are defined by

$$\{S, A, (A(x), x \in S), Q(u, y|x, a), B, r(x, a)\} \tag{1}$$

with the following components:

(a) S denotes a Borel state space, endowed with the Borel σ-algebras $\mathcal{B}(S)$.

(b) A denotes a Borel action space, endowed with the Borel σ-algebras $\mathcal{B}(A)$.
(c) $A(x) \in \mathcal{B}(A)$ represents the set of allowable actions when the system is at state $x \in S$. $K := \{(x,a)|x \in S, a \in A(x)\}$ represents the set of all feasible pairs of states and actions.
(d) $Q(\cdot,\cdot|x,a)$ is a semi-Markov kernel on $R^+ \times S$ given K, where $R^+ := [0,\infty)$. For any $u \in R^+, D \in \mathcal{B}(S)$, when the action $a \in A(x)$ is taken in state x, $Q(u,D|x,a)$ denotes the joint probability that the holding time of the system is no more than $u \in R^+$ and the state x changes into the set D. The semi-Markov kernel $Q(\cdot,\cdot|x,a), (x,a) \in K$ has the following features:
  (i) For any $D \in \mathcal{B}(S)$, $Q(\cdot,D|x,a)$ is a non-decreasing, right continuous function from R^+ to $[0,1]$ with $Q(0,D|x,a) = 0$.
  (ii) For any $u \in R^+$, $Q(u,\cdot|x,a)$ is a sub-stochastic kernel on the state space S.
(e) B is target set, which is a measurable subset of S, and usually represents the set of failure (or ruin) states of a system.
(f) $r(x,a)$ denotes the reward rate, which is assumed to be nonnegative measurable function on K such that $r(x,\cdot) \equiv 0$ for all $x \in B$.

The first passage SMDP with exponential utility evolves as follows: When the system state is $x_0 \in B^c$ at time $t_0 = 0$, the decision maker selects an admissible action a_0 from the action set $A(x_0)$, where B^c denotes the complement of B. Consequently, the system stays in the state x_0 up to time t_1. At this point the system jumps to state x_1 with probability $p(x_1|x_0,a_0)$, and earns a reward $r(x_0,a_0)(t_1-t_0)$. If the state $x_1 \in B$, the system will stay at the target set B forever. If the state $x_1 \in B^c$, a new decision epoch t_1 comes along. Then, based on the present state x_1 and the previous state x_0, the decision maker chooses an action $a_1 \in A(x_1)$ and the process is repeated. Thus, during its evolution, the system receives a series of rewards. The decision maker aims at maximizing the exponential utility of the total rewards before the state of the system first reaches the target set B.

Let

$$h_k := (x_0, a_0, t_1, x_1, a_1, \ldots, t_k, x_k), \tag{2}$$

be an admissible history up to the k-th decision epoch, where $t_{m+1} \geq t_m \geq 0$, $x_m \in S, a_m \in A(x_m)$ for $m = 0,1,\ldots,k-1, x_k \in S$. From the evolution of SMDPs, we know that t_{k+1} $(k \geq 0)$ denotes the $(k+1)$-th decision epoch, x_k denotes the state of the system on $[t_k, t_{k+1})$, a_k denotes an action, which is chosen by the decision maker at time t_k. $\theta_{k+1} := t_{k+1} - t_k$ denotes the sojourn time at state x_k, which may follow any given probability distribution.

The set of all admissible histories h_k is denoted by H_k, that is $H_0 := S$ and $H_k := (S \times A \times (0,+\infty])^k \times S$.

For the sake of the optimality problem, we shall pay close attention to some classes of policies that we introduce below.

Definition 1. *A sequence* $\pi = \{\pi_k, k \geq 0\}$ *is called* stochastic history-dependent policy *if, for any* $k = 0,1,2\ldots$, *the stochastic kernel* π_k *on* $A(x_k)$ *given* H_k *satisfies*

$$\pi_k(A(x_k)|h_k) = 1 \quad \textit{for any} \quad h_k \in H_k.$$

Denote by Π the set of all stochastic history-dependent policies, ϕ the set of all stochastic kernels φ on $A(x)$ given S such that $\varphi(A(x)|x) = 1$, and F the family of all Borel measurable functions f from S to $A(x)$ for all $x \in S$.

Definition 2. *A policy $\pi = \{\pi_k\} \in \Pi$ is called* stochastic Markov *if there exists a sequence of stochastic kernels $\{\varphi_k\}$ such that $\pi_k(\cdot|h_k) = \varphi_k(\cdot|x_k)$ for $k \geq 0, h_k \in H_k$, and $\varphi_k \in \phi$. For simplicity, we denote such a policy by $\pi = \{\varphi_k\}$.*

A stochastic Markov policy $\pi = \{\varphi_k\}$ is called stochastic stationary *if all the φ_k are independent of k. Such a policy is denoted by φ, for simplicity.*

A stochastic Markov policy $\pi = \{\varphi_k\}$ is called deterministic Markov *if each $\varphi_k(\cdot|x_k)$ is concentrated at $f_k(x_k) \in A(x_k)$ for some measurable functions $\{f_k\}$ with $k \geq 0, x_k \in S$, and $f_k \in F$.*

A deterministic Markov policy $\pi = \{f_k\}$ is called deterministic stationary *if all the measurable functions f_k are independent of k. For simplicity, such a policy is denoted by f.*

The class of all stochastic Markov, stochastic stationary, deterministic Markov, and deterministic stationary policies are, respectively, denoted by Π_{RM}, Π_{RS}, Π_{DM} and Π_{DS}. Clearly, $\phi = \Pi_{RS} \subset \Pi_{RM} \subset \Pi$ and $F = \Pi_{DS} \subset \Pi_{DM} \subset \Pi$.

For the sake of mathematical rigor, we need to construct a well-suited probability space. Define a sample space $\Omega := \{(x_0, a_0, t_1, x_1, a_1, \ldots, t_k, x_k, a_k, \ldots)|$ $x_0 \in S$, $a_0 \in A(x_0), t_l \in (0, \infty], x_l \in S, a_l \in A(x_l)$ for each $1 \leq l \leq k, k \geq 1\}$. Let F be the Borel σ-algebra of the sample space Ω. For any $\omega := (x_0, a_0, t_1, x_1, a_1,$ $\ldots, t_k, x_k, a_k, \ldots) \in \Omega$, we define the random variables T_k, X_k, A_k on $(\Omega, \mathcal{F})$ as follows:

$$T_k(\omega) := t_k, X_k(\omega) := x_k, A_k(\omega) := a_k, T_\infty(\omega) := \lim_{k\to\infty} T_k(\omega). \tag{3}$$

In what follows, for the purpose of simplicity, we omit the argument ω.

Moreover, we define the state process $\{x_t, t \geq 0\}$ and the action process $\{A_t, t \geq 0\}$ on $(\Omega, \mathcal{F})$ by

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

$$A_t := \sum_{k\geq 0} I_{\{T_k \leq t < T_{k+1}\}} A_k + a_\Delta I_{\{t \geq T_\infty\}},$$

where $I_D(\cdot)$ denotes the indicator function on the set D, $\Delta \notin E$ is a cemetery state, and a_Δ is an isolated point.

For any policy $\pi \in \Pi$ and initial state $x \in S$, in the light of the Ionescu Tulcea theorem (e.g., Proposition C.10 in [11]), there exist a unique probability measure P_x^π on the measurable space $(\Omega, \mathcal{F})$ such that,

$$P_x^\pi(A_k \in \Gamma|T_0, X_0, A_0, \ldots, T_k, X_k) = \pi_k(\Gamma|T_0, X_0, A_0, \ldots, T_k, X_k), \tag{4}$$

$$P_x^{\pi}(T_{k+1} - T_k \leq u, X_{k+1} \in D | T_0, X_0, A_0, \ldots, T_k, X_k, A_k) = Q(u, D | X_k, A_k),$$

for each $u \in R^+, \Gamma \in \mathcal{B}(A), D \in \mathcal{B}(S), k \geq 0$. We shall use $\mathbb{E}_x^{\pi}$ to represent the expectation operator with respect to P_x^{π}.

To avoid the possibility that the system generates an infinite number of jumps within a fixed finite horizon, we need to impose the following condition.

Assumption 1 *For any $\pi \in \Pi, x \in S$, $P_x^{\pi}(T_{\infty} = \infty) = 1$.*

To ease the verification of Assumption 1, we state the following sufficient condition for its validity.

Lemma 1. *If $Q(\delta, S|x, a) \leq 1 - \varepsilon$ with some constants δ, $\varepsilon > 0$ and $(x, a) \in K$, then Assumption 1 holds.*

Proof. The proof follows directly from Proposition 2.1 in [14]. □

Remark 1.(a) A key feature of Lemma 1 is that the condition is imposed on the semi-Markov kernel, and can be directly verified.

(b) Lemma 1 is the standard regular condition, which is similar to the classic expected criteria for SMDPs, see, for instance [13–15,18].

The random variable τ_B is given by

$$\tau_B = \begin{cases} \inf\{t \geq 0 : x_t \in B\}, & if \quad \{t \geq 0 : x_t \in B\} \neq \emptyset; \\ +\infty, & \text{otherwise.} \end{cases} \tag{5}$$

represents the first passage time for which the state process $\{x_t, t \geq 0\}$ first enters the target set B.

For any $x \in S$ and $\pi \in \Pi$, we define the first passage exponential utility criterion by

$$V^{\pi}(x) := E_x^{\pi}\left(e^{-\gamma \int_0^{\tau_B} r(x_t, A_t)dt}\right), \tag{6}$$

where $\gamma > 0$ represents the risk aversion coefficient, which expresses the degree of risk aversion that the decision makers face to the level of the total rewards before the state of the system first enters the target set.

Definition 3. *A policy $\pi^* \in \Pi$ is called an* optimal policy, *if*

$$V^{\pi^*}(x) = \sup_{\pi \in \Pi} V^{\pi}(x), x \in S. \tag{7}$$

The corresponding value function is given by

$$V^*(x) := \sup_{\pi \in \Pi} V^{\pi}(x), x \in S. \tag{8}$$

Remark 2. Note that for any $\pi \in \Pi$ and initial state $x \in B$, in view of (5), (6) and (8), we have $\tau_B = 0$ and $V^*(x) = V^{\pi}(x) = 1$. In order to avoid this trivial case, our arguments consider only the case $x \in B^c$.

3 Main Results

In this section, we will state the main results concerning the first passage exponential utility optimality problem for SMDPs.

Notation: Let $\mathcal{V}_m$ denotes the set of all Borel measurable functions from S to $[0,1]$. For any $x \in B^c, V \in \mathcal{V}_m, \varphi \in \phi, a \in A(x)$, we define the operators $M^aV, M^\varphi V$ and MV as follows:

$$\begin{aligned} M^aV(x) &:= \int_B \int_0^{+\infty} e^{-\gamma r(x,a)u} Q(du, dy|x, a) \\ &\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,a)u} V(y) Q(du, dy|x, a), \\ M^\varphi V(x) &:= \int_{A(x)} \varphi(da|x) M^a V(x), \\ MV(x) &:= \sup_{a \in A(x)} M^a V(x). \end{aligned}$$

For any $\varphi \in \phi$, we also define the operators $(M^nV, n \geq 1), ((M^\varphi)^n V, n \geq 1)$ as follows:

$$M^{n+1}V = M(M^nV), (M^\varphi)^{n+1}V = M^\varphi((M^\varphi)^n V), n \geq 1.$$

Since the state and action space are Borel space, in order to ensure the existence of optimal policies, it follows from [28,31,32], we need establish the following continuity-compactness condition, and which is trivially satisfied for the case of denumerable state space and finite action set $A(x)$ with $x \in S$.

Assumption 2. *(a) For any $x \in B^c$, $A(x)$ is compact;*
(b) For each fixed $V \in \mathcal{V}_m$, $\int_{y \in S} \int_0^{+\infty} e^{-\gamma r(x,a)u} V(y) Q(du, dy|x, a)$ is upper semicontinuous and inf-compact on K.

Lemma 2. *Suppose that Assumptions 1 and 2 hold. Then the operators M^a and M have the following properties:*

(a) For any $U, V \in \mathcal{V}_m$, if $U \geq V$, then $M^aU(x) \geq M^aV(x)$ and $MU(x) \geq MV(x)$ for any $x \in S$ and $a \in A(x)$.
(b) For any $V \in \mathcal{V}_m$, there exists a policy $f \in \Pi_{DS}$ such that $MV(x) = M^fV(x)$ for any $x \in S$.

Proof. (a) This statement follows from the definitions of operators M^a and M.
(b) Assuming the validity of Assumption 1 and 2, and invoking the measurable selection theorem (Theorem B.6 in [28]), we conclude that, for each $x \in S$, there is a stationary policy $f \in F$ with $M^fV(x) = MV(x) = \sup_{a \in A(x)} M^a V(x)$.

□

Since state process $\{x_t, t \geq 0\}$ is non-explosive and the reward rate is non-negative, in view of the monotone convergence theorem, we can rewrite $V^{\pi}(x)$ as follows:

$$\begin{aligned} V^{\pi}(x) &= E_x^{\pi}\left(e^{-\gamma \int_0^{\tau_B} r(x_t, A_t)dt}\right) \\ &= E_x^{\pi}\left(e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\tau_B > t\}} r(x_t, A_t)dt}\right) \\ &= E_x^{\pi}\left(e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k} \in B^c\}\}} r(x_t, A_t)dt}\right) \\ &= \lim_{n\to\infty} E_x^{\pi}\left(e^{-\gamma \sum_{m=0}^{n} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k} \in B^c\}\}} r(x_t, A_t)dt}\right). \end{aligned} \tag{9}$$

We shall find it essential to define the sequence $\{V_n^{\pi}(x), n = -1, 0, 1, \ldots\}$ by

$$\begin{aligned} V_{-1}^{\pi}(x) &:= 1, \\ V_n^{\pi}(x) &:= E_x^{\pi}\left(e^{-\gamma \sum_{m=0}^{n} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k} \in B^c)\}\}} r(x_t, A_t)dt}\right). \end{aligned}$$

Obviously, $V_n^{\pi}(x) \geq V_{n+1}^{\pi}(x)$ for any $n \geq -1$ and $\lim_{n\to\infty} V_n^{\pi}(x) = V^{\pi}(x)$ for all $x \in B^c$.

Proposition 1. *For each $\pi = \{\pi_0, \pi_1, \ldots\} \in \Pi$ and $x \in S$. Then, there exists a policy $\pi' = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$, satisfying $V^{\pi}(x) = V^{\pi'}(x)$.*

Proof. Since $V^{\pi}(x) = E_x^{\pi}\left(e^{-\gamma \sum_{m=0}^{\infty} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^{m}\{x_{T_k} \in B^c\}\}} r(x_t, A_t)dt}\right)$ in (9), to prove this proposition we need to prove that, for each $x \in S$, there exists a randomized Markov policy $\pi' = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$ such that

$$\begin{aligned} &P_x^{\pi'}(X_k \in D, T_{n+1} - T_n > u, A_k \in \Gamma) \\ &= P_x^{\pi}(X_k \in D, T_{n+1} - T_n > u, A_k \in \Gamma) \end{aligned}$$

with $k = 0, 1, \ldots, u \in R^+, D \in \mathcal{B}(S), \Gamma \in \mathcal{B}(A)$.

Thus, in view of property (4), it suffices to show that

$$P_x^{\pi'}(X_k \in D, A_k \in \Gamma) = P_x^{\pi}(X_k \in D, A_k \in \Gamma). \tag{10}$$

Along the same arguments as in the proof of Theorem 5.5.1 in [28], one can prove (10) by induction on the integer k. □

Proposition 1 states, in particular, that in seeking optimal policies for (7), it is sufficient to limit the search to the set of randomized Markov policies. Thus, from now on, we will limit our attention to Π_{RM}.

The following lemma is required to establish the optimality equation.

Lemma 3. *Under Assumption 1 and 2, for any $x \in S$, $n \geq -1$, and $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$, the following statements hold.*

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

(b) $V_{n+1}^\pi(x) = M^{\varphi_0} V_n^{{}^1\pi}(x)$ *and* $V^\pi(x) = M^{\varphi_0} V^{{}^1\pi}(x)$, *with* ${}^1\pi := \{\varphi_1, \varphi_2, \ldots\}$ *being the 1-shift policy of* π.
In particular, for any $f \in F$, $V_{n+1}^f(x) = M^f V_n^f(x)$ *and* $V^f(x) = M^f V^f(x)$.

Proof. (a) We shall prove the first statement of (a) by induction on the integer $n \geq -1$. The statement is trivial for $n = -1$ since $V_{-1}^\pi(x) = 1 \in \mathcal{V}_m$ for any $x \in S$ and $\pi \in \Pi_{RM}$. Assume the statement holds for any $n < k$. Then, by (4) and the property of conditional expectation, we have

$$
\begin{aligned}
& V_{k+1}^\pi(x) \\
&= E_x^\pi\Big(e^{-\gamma \sum_{m=0}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt}\Big) \\
&= E_x^\pi[E_x^\pi[e^{-\gamma \sum_{m=0}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt} | T_0, x_{T_0}, A_0, T_1, x_{T_1}]] \\
&= \int_{A(x)} \varphi_0(da|x) \\
&\quad \times \int_S \int_0^{+\infty} E_x^\pi\Big(e^{-\gamma(\int_0^{T_1} r(x_t, A_t) dt + \sum_{m=1}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=1}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt)} \\
&\quad |T_0 = 0, x_{T_0} = x, A_0 = a, T_1 = u, x_{T_1} = y\Big) Q(du, dy|x, a) \\
&= \int_{A(x)} \varphi_0(da|x) \int_B \int_0^{+\infty} e^{-\gamma r(x,a) u} Q(du, dy|x, a) + \int_{A(x)} \varphi_0(da|x) \\
&\quad \times \int_{B^c} \int_0^{+\infty} E_x^\pi\Big(e^{-\gamma(\int_0^{T_1} r(x_t, A_t) dt + \sum_{m=1}^{k+1} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=1}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt)} \\
&\quad |T_0 = 0, x_{T_0} = x, A_0 = a, T_1 = u, x_{T_1} = y\Big) Q(du, dy|x, a) \\
&= \int_{A(x)} \varphi_0(da|x) [\int_B \int_0^{+\infty} e^{-\gamma r(x,a) u} Q(du, j|x, a) \\
&\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,a) u} E_y^{{}^1\pi}\Big(e^{-\gamma \sum_{m=0}^{k} \int_{T_m}^{T_{m+1}} I_{\{\bigcap_{k=0}^m \{x_{T_k} \in B^c\}\}} r(x_t, A_t) dt}\Big) \\
&\quad \times Q(du, dy|x, a)] \\
&= \int_{A(x)} \varphi_0(da|x) [\int_B \int_0^{+\infty} e^{-\gamma r(x,a) u} Q(du, dy|x, a) \\
&\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,a) u} V_k^{{}^1\pi}(y) Q(du, dy|x, a)] \\
&:= M^{\varphi_0} V_k^{{}^1\pi}(x)
\end{aligned}
$$

which together with induction hypothesis implies that $V_{k+1}^\pi(x)$ is a measurable function and $V_{k+1}^\pi(x) \leq 1$. Thus, $V_n^\pi \in \mathcal{V}_m$ for all $n \geq -1$. Since the limit of a convergent sequence of measurable functions is itself a measurable function, we obtain $\lim_{n\to\infty} V_n^\pi = V^\pi \in \mathcal{V}_m$. This concludes the proof of (a).

(b) From the proof of part (a), we can deduce that, for any $x \in B^c$ and $n \geq -1$,

$$V_{n+1}^{\pi}(x) = M^{\varphi_0} V_n^{^1\pi}(x). \tag{11}$$

Letting $n \to \infty$ in (11) and using the monotone convergence theorem, we obtain

$$V^{\pi}(x) = M^{\varphi_0} V^{^1\pi}(x).$$

In particular, for $\pi = f \in F$, we have $V^f(x) = M^f V^f(x)$. □

Remark 3. For any $x \in B^c$ and $f \in F$, one can use Lemma 3 to develop an efficient iteration algorithm for the computation of the function $V^f(x)$ based on the following: $V^f(x) = \lim_{n\to\infty} V_n^f(x)$ where $V_{-1}^f(x) := 1$ and $V_{n+1}^f(x) = M^f V_n^f(x)$ for $n \geq 0$.

The following theorem states the existence of an optimality equation.

Theorem 1. *Under Assumption 1 and 2, the following hold.*

(a) For each $n \geq -1$, let $V_{n+1}^* := MV_n^*$ with $V_{-1}^* := 1$. Then, $\lim_{n\to\infty} V_n^* = V^* \in \mathcal{V}_m$.
(b) The value function V^* is a solution to the optimality equation $V^* = MV^*$.
(c) There is a policy $f^* \in F$ such that $V^*(x) = M^f V^*(x), x \in B^c$.

Proof. (a) Using Lemma 2(a) and the definition of the operator M, we obtain $0 \leq V_{n+1}^*(x) \leq V_n^*(x) \leq 1$ and $V_n^* \in \mathcal{V}_m, n \geq -1$, for any $x \in B^c$. Thus, $\tilde{V} := \lim_{n\to\infty} V_n^* \in \mathcal{V}_m$, since the limit of a convergent sequence of measurable function is also measurable. To complete the proof of part (a), we need to prove that $\tilde{V} = V^*$.

We first show by induction on $n \geq -1$ that for any $x \in B^c$ and $\pi = \{\varphi_0, \varphi_1, \ldots\} \in \Pi_{RM}$

$$V_n^*(x) \geq V_n^{\pi}(x). \tag{12}$$

It is clear that $V_{-1}^* = V_{-1}^{\pi} = 1$ for any $\pi \in \Pi_{RM}$. Suppose that (12) holds for any $n \leq k$. By the induction hypothesis, the definition of the operator M and Lemma 3(b), we have

$$V_{k+1}^*(x) = MV_k^*(x) \geq MV_k^{^1\pi}(x) \geq M^{\varphi_0} V_k^{^1\pi}(x) = V_{k+1}^{\pi}(x).$$

Letting $n \to \infty$ in (12), we obtain $\tilde{V}(x) = \lim_{n\to\infty} V_n^*(x) \geq V^{\pi}(x)$ with $\pi \in \Pi_{RM}$. Since π is arbitrary, we conclude that $\tilde{V}(x) \geq V^*(x)$.

We need, now, to prove the reverse inequality $\tilde{V}(x) \leq V^*(x)$. For any $x \in B^c, n \geq -1$, let $A_n := \{a \in A(x) | M^a V_n^*(x) \geq M\tilde{V}(x)\}$ and $A^* := \{a \in A(x) | M^a \tilde{V}(x) = M\tilde{V}(x)\}$. By the compactness-continuity condition in Assumption 2 and the convergence $V_n^* \downarrow \tilde{V}$, we conclude that A_n and A^* are nonempty and compact, and that $A_n \downarrow A^*$. It follows from the measurable selection theorem (Theorem B.6 in [28]) that, for each $n \geq 1$, there exist $a_n \in A_n$

such that $M^{a_n}V_n^*(x) = MV_n^*(x)$. Hence, using compactness and the convergence $A_n \downarrow A^*$, we deduce that there exist an $a^* \in A^*$ and a subsequence $\{a_{n_k}\}$ of $\{a_n\}$ such that $a_{n_k} \to a^*$. Since $V_n^* \downarrow \tilde{V}$, by Lemma 3(a), for any given $n \geq 1$, we have

$$M^{a_{n_k}}V_{n_k}^*(x) \leq M^{a_{n_k}}V_n^*(x) \quad \forall n_k \geq n.$$

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

$$\tilde{V}^*(x) \leq M^{a^*}V_n^*(x),$$

which together with the convergence $V_n^* \downarrow \tilde{V}$ imply

$$\tilde{V}^*(x) \leq M^{a^*}\tilde{V}(x) \leq M\tilde{V}(x),$$

By Lemma 2(b), there exists a stationary policy $f \in F$ such that

$$\tilde{V}(x) \leq M\tilde{V}(x) = M^f\tilde{V}(x).$$

Moreover, using Lemma 2(a), Lemma 3(b) and Remark 3, we obtain

$$\tilde{V}(x) \leq (M^f)^n\tilde{V}(x) \leq (M^f)^nV_{-1}^f(x) = V_{n-1}^f(x).$$

Letting $n \to \infty$, and invoking Remark 3, we obtain $\tilde{V}(x) \leq V^f(x) \leq V^*(x)$, which proves the part (a) of the theorem.

(b) By virtue of Lemma 3(b), we know that for any $x \in B^c$ and $\pi \in \Pi_{RM}$, we have

$$V^\pi(x) = M^{\varphi_0}V^{1\pi}(x) \leq M^{\varphi_0}V^*(x) \leq MV^*(x).$$

Taking the supremum over all policies $\pi \in \Pi_{RM}$ implies $V^*(x) \leq MV^*(x)$.

The reverse inequality is proved as follows: From the definition of V_n^*, for any $x \in B^c$ and $a \in A(x)$,

$$V_{n+1}^*(x) = MV_n^*(x) \geq M^aV_n^*(x).$$

Letting $n \to \infty$ and using the monotone convergence theorem, we obtain

$$V^*(x) \geq M^aV^*(x),$$

which implies that $V^*(x) \geq MV^*(x)$ since $a \in A(x)$ is arbitrary. This proves $V^* = MV^*$.

(c) The statement in (c) follows from Lemma 2. □

To guarantee the uniqueness of solution of the optimality equation and the existence of the optimal policies, we require the following additional condition (i.e., Assumption 3).

Assumption 3 *For any* $x \in B^c, f \in \Pi_s, P_x^f(\tau_B < +\infty) = 1$.

Remark 4.(a) Assumption 3 means that, when the initial state of such system is $X_0 = x \in S$, the controlled state process $\{x_t, t \geq 0\}$ will eventually enter the target set B under the policy $f \in F$.

(b) Letting $X_n := x_{T_n}, n = 0, 1, \ldots$, T_n denotes the jump epoch. Then, we obtain a discrete-time embedded chain $\{X_n, n \geq 0\}$. For every $x \in B^c$, using Theorem 3.3 in [16], we know that Assumption 3 can be rewritten as follows:.

$$P_x^f(\tau_B < +\infty) = P_x^f(\bigcup_{n=1}^{\infty} \{X_n \in B\}) = 1,$$

which is equivalent to

$$P_x^f(\bigcap_{n=1}^{\infty} \{X_n \in B^c\}) = 0. \tag{13}$$

(c) Using Proposition 3.3 in [19], we also obtain a sufficient condition to verify Assumption 3. There exist a constant $\alpha > 0$ such that $\int_B P(dy|x, a) \geq \alpha$ for $(x, a) \in B^c \times A(x)$, then Assumption 3 holds.

Lemma 4. *Suppose that Assumptions 1 and 3 hold.*

(a) If $U, V \in \mathcal{V}_m$ are such that $U(x) - V(x) \leq M^f(U - V)(x)$ with $x \in B^c, f \in \Pi_s$, then $U(x) \leq V(x)$.
(b) For any $f \in \Pi_s$, $V^f \in \mathcal{V}_m$ is the unique solution to the equation $V = M^f V$.

Proof. (a) For any $U, V \in \mathcal{V}_m$, $x \in B^c, f \in \Pi_s$, we will show the following conclusion by induction,

$$(M^f)^n(U - V)(x) \leq P_x^f(\bigcap_{k=1}^{n} \{X_k \in B^c\}), n \geq 1. \tag{14}$$

For $n = 1$, it follows from $U, V \in \mathcal{V}_m$ that

$$\begin{aligned} M^f(U - V)(x) &= M^f U(x) - M^f V(x) \\ &= \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u}(U - V)(y) Q(du, dy|x, a) \\ &\leq \int_{B^c} \int_0^{+\infty} Q(du, dy|x, a) \\ &= P_x^f(X_1 \in B^c). \end{aligned}$$

Suppose that (14) holds for $n = k$. Then, by using the induction hypothesis and the nonnegativity of the reward rate, we have

$$
\begin{aligned}
(M^f)^{k+1}(U-V)(x) &= M^f(M^f)^k(U-V)(x)\\
&= \int_{B^c}\int_0^{+\infty} e^{-\gamma r(x,f)u}(M^f)^k(U-V)(y)\\
&\quad \times Q(du,dy|x,a)\\
&= \int_{B^c}\int_0^{+\infty} e^{-\gamma r(x,f)u}P_y^f(\bigcap_{l=1}^{k}\{X_l\in B^c\})\\
&\quad \times Q(du,dy|x,a)\\
&\le \int_{B^c}\int_0^{+\infty} P_y^f(\bigcap_{l=1}^{k}\{X_l\in B^c\})Q(du,dy|x,a). \qquad (15)
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&P_x^f(\bigcap_{l=1}^{k+1}\{X_l\in B^c\})\\
&= E_x^f[I_{\{\bigcap_{l=1}^{k+1}\{X_l\in B^c\}\}}]\\
&= E_x^f[E_x^f[I_{\bigcap_{l=1}^{k+1}\{X_l\in B^c\}}|X_0,X_1]\\
&= \int_{B^c}\int_0^{+\infty} P_x^f\Big(\bigcap_{l=1}^{k+1}\{X_l\in B^c\}|X_0=x,X_1=y\Big)Q(du,dy|x,a)\\
&= \int_{B^c}\int_0^{+\infty} P_y^f\Big(\bigcap_{l=1}^{k}\{X_l\in B^c\}\Big)Q(du,dy|x,a),
\end{aligned}
$$

from which together with (15) and the induction, we have for all $n\ge 1$,

$$
U(x)-V(x)\le (M^f)^n(U(x)-V(x))\le P_x^f(\bigcap_{k=1}^{n}\{X_k\in B^c\}). \qquad (16)
$$

Letting $n\to\infty$, using (13), we obtain

$$
U(x)-V(x)\le P_x^f(\bigcap_{k=1}^{\infty}\{X_k\in B^c\})=0.
$$

Then, $U(x)\le V(x)$, for $x\in S$.

(*b*) For any $x\in S, f\in F$, it follows from Lemma 2(b) that $V^f(x)\in\mathcal{V}_m$ satisfies the equation $V(x)=M^fV(x)$. If $U(x)$ is another solution to the equation $U(x)=M^fU(x)$ on S, and thus $U(x)-V^f(x)=M^f(U(x)-V^f(x))$, which together with the statement in part (a), we know $U(x)=V^f(x)$ and the uniqueness of solution to the equation is proved. □

Theorem 2. *Suppose that Assumption 1,2 and 3 hold. Then, the following statements hold.*

(a) The value function V^* is the unique solution to the optimality equation $V^* = MV^*$.
(b) There is a policy $f^* \in F$ which satisfies $V^* = M^{f^*}V^*$, $V^* = V^{f^*}$ and such a policy $f^* \in F$ is optimal.

Proof. (a) It follows from Lemma 3 (b) that V^* satisfies the equation $V^* = MV^*$. Then, by Lemma 2(b), there exists a stationary policy $f^* \in F$ such that $V^* = M^{f^*}V^*$. Moreover, U is another solution of the equation $U = MU$. Similarly, the existence of a policy $f' \in F$ satisfying $U = M^{f'}U$ is ensured by Lemma 2(b). Then, we have $V^* - U \leq M^{f^*}(V^* - U)$. Combining this inequality and Lemma 4 yields that $V^* \leq U$. Similarly, we obtain $U - V^* \leq M^{f'}(U - V^*)$ and $U \leq V^*$, which implies $U = V^*$ and the uniqueness of V^* is achieved.

(b) Since $V^* \in \mathcal{V}_m$, for any $x \in B^c$, Lemma 2 guarantees the existence of a stationary policy $f^* \in F$ such that

$$V^*(x) = M^{f^*}V^*(x),$$

which together with Lemma 3 and Remark 11 yield

$$V^* = \lim_{n\to\infty}(M^{f^*})^n V^* \leq \lim_{n\to\infty}(M^{f^*})^n V_{-1}^{f^*} = \lim_{n\to\infty} V_{n-1}^{f^*} = V^{f^*}.$$

This implies the optimality of f^*. □

Theorem 1 leads to the following iterative algorithm for computing the value function and the corresponding optimal policies.

The value iteration algorithm procedure:

Step 1: For any $x \in B^c$, set $V_{-1}^*(x) := 1$.

Step 2: According to Theorem 1, the value $V_{n+1}^*(x), n \geq 1$, is iteratively computed as:

$$\begin{aligned} M^a V_n^*(x) &= \int_B \int_0^{+\infty} e^{-\gamma r(x,f)u} Q(du, dy|x, a) \\ &\quad + \int_{B^c} \int_0^{+\infty} e^{-\gamma r(x,f)u} V_n^*(y) Q(du, dy|x, a), \\ V_{n+1}^*(x) &= \sup_{a\in A(x)} \{M^a V_n^*(x)\}. \end{aligned}$$

Step 3: When $|V_{n+1}^* - V_n^*| < 10^{-12}$, the iteration stops. Since V_n^* is very close to V_{n+1}^*, one can view V_{n+1}^* as a good approximation of the value function V^*. In addition, Lemma 2 and Theorem 2 ensure the existence of a policy $f^* \in F$ such that $MV^* = M^{f^*}V^*$, and this policy f^* is optimal. Or else, go back to step 2 and replace n with $n + 1$.

4 Example

In this section, an example is given to illustrate our main results, and to demonstrate the computation of an optimal stationary policy and the corresponding value function using the above described iterative algorithm.

Example 1. Consider a company using idle funds for financial management. When the company has some idle funds (which is denoted by state 1), the decision maker gets the reward at the rate of return $r(1, a_{11}) \geq 0$ through deposit method a_{11} or the reward at the rate of return $r(1, a_{12}) \geq 0$ through another deposit method a_{12}. When the company has plenty of idle funds (which is denoted by state 2), the decision maker can choose a financial management a_{21} earning in a reward rate $r(2, a_{21}) \geq 0$ or another financing way a_{22} earning in a reward rate $r(2, a_{22}) \geq 0$. When the company goes bankrupt (which is denoted by state 0), the decision-maker does not need to choose any way of financing a_{01} and cannot get any reward $r(0, a_{01}) = 0$.

Suppose that the evolution mechanism of this system is described as a SMDP. When the system state is 1, the decision maker selects an admissible action $a_{1n}, n = 1, 2$. Then, the system stays at the state 1 with a random time satisfying the uniform distribution in the region $[0, u(1, a_{1n})], n = 1, 2$. After the system state lingers for a period of time, it will move to a new state $j \in \{0, 2\}$ with the probability $p(j|1, a_{1n}), n = 1, 2$. When the action a_{2n} is selected $n = 1, 2$, the system stays at 2 with a random time satisfying the exponential distribution with the parameter $\lambda(2, a_{2n})$. Consequently, the system jumps to state $j \in \{0, 1\}$ with the probability $p(j|2, a_{2n}), n = 1, 2$.

The corresponding parameters of this SMDPs are given as follows: The state space $S = \{0, 1, 2\}$, the target set $B = \{0\}$ and the admissible action sets $A(0) = \{a_{01}\}, A(1) = \{a_{11}, a_{12}\}, A(2) = \{a_{21}, a_{22}\}$, the risk-sensitivity coefficient $\gamma = 1$. The transition probabilities are assumed to be given

$$\begin{aligned} &p(0|0, a_{01}) = 1, && p(0|1, a_{11}) = \frac{1}{2}, && p(2|1, a_{11}) = \frac{1}{2}, \\ &p(0|1, a_{12}) = \frac{2}{3}, && p(2|1, a_{12}) = \frac{1}{3}, && p(0|2, a_{21}) = \frac{3}{10}, \\ &p(1|2, a_{21}) = \frac{7}{10}, && p(0|2, a_{22}) = \frac{2}{5}, && p(1|2, a_{22}) = \frac{3}{5}. \end{aligned} \tag{17}$$

In addition, the corresponding distribution parameters are given by

$$\begin{aligned} &u(1, a_{11}) = 30, && u(1, a_{12}) = 40, \\ &\lambda(2, a_{21}) = 0.11, && \lambda(2, a_{22}) = 0.13. \end{aligned} \tag{18}$$

and the reward rates are given by

$$\begin{aligned} &r(1, a_{11}) = 0.0035, && r(1, a_{12}) = 0.011, \\ &r(2, a_{21}) = 0.013, && r(2, a_{22}) = 0.015. \end{aligned}$$

In this model, we mainly focus on the existence and calculation parts of an optimal policy and the value function for first passage exponential utility criterion. As can be seen from the discussion in Sect. 3 above, we first need to verify Assumption 1, 2 and 3. Indeed, by (17) and (18), we know that Assumption 1 and 3 are satisfied. Moreover, since the state space is denumerable and the action space A is finite, Assumption 2 is trivially satisfied. Thus, by Theorem 1 and 2,

the value iteration technique can be used for evaluating the value function and the exponential optimal policies as follows:

Step 1: Let $V_{-1}^*(x) := 1, x = 1, 2$.

Step 2: For $x = 1, 2, n \geq 1$, using Theorem 1 (a), we obtain

$$
\begin{aligned}
&V_n^*(1) = MV_{n-1}^*(1), \\
&= \max\Big\{\frac{1}{2} \times \frac{1}{30} \times \int_0^{30} e^{-0.0035u} du \\
&+\frac{1}{2} \times \frac{1}{30} \times \int_0^{30} e^{-0.0035u} du \times V_{n-1}^*(2), \\
&\frac{2}{3} \times \frac{1}{40} \times \int_0^{40} e^{-0.011u} du + \frac{1}{3} \times \frac{1}{40} \times \int_0^{40} e^{-0.011u} du \times V_{n-1}^*(2)\Big\} \\
&V_n^*(2) = MV_{n-1}^*(2), \\
&= \max\Big\{\frac{3}{10} \times 0.11 \times \int_0^{+\infty} e^{-0.123u} du \\
&+\frac{7}{10} \times 0.11 \times \int_0^{+\infty} e^{-0.123u} du \times V_{n-1}^*(1), \\
&\frac{2}{5} \times 0.13 \times \int_0^{+\infty} e^{-0.145u} du + \frac{3}{5} \times 0.13 \times \int_0^{+\infty} e^{-0.145u} du \times V_{n-1}^*(1)\Big\}
\end{aligned}
$$

Step 3: When $|V_n^* - V_{n-1}^*| < 10^{-12}$, go to step 4, the value V_n^* is usually approximated as V^*; otherwise, go to step $n + 1$ and go back to step 2.

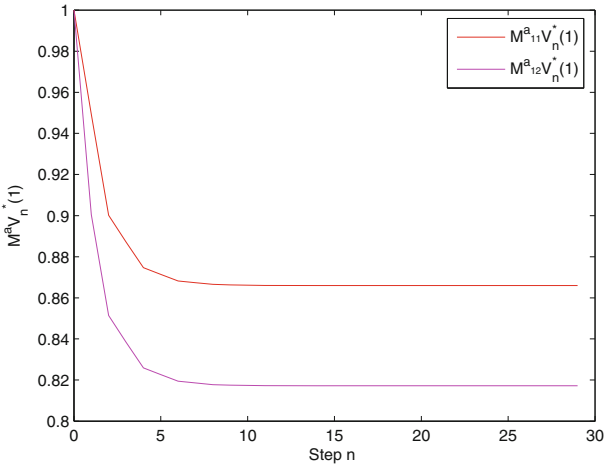

Fig. 1. The function $M^a V_n^*(1)$

Step 4: Plot out the graphs of the value functions $M^{a_{ij}} V_n^*(i)$ and $V_n^*(i), i = 1, 2; j = 1, 2$, see Figs. 1, 2 and 3.

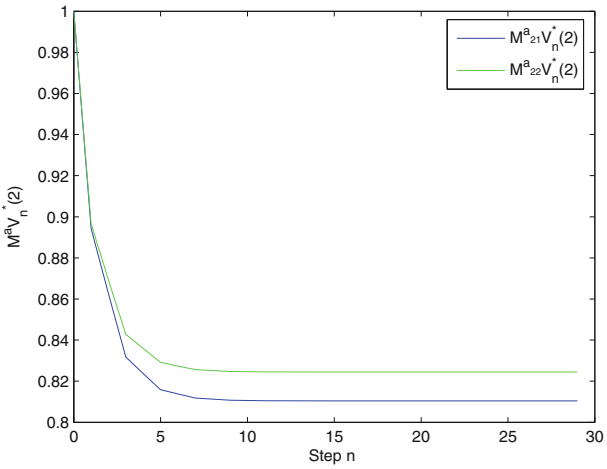

Fig. 2. The function $M^aV_n^*(2)$

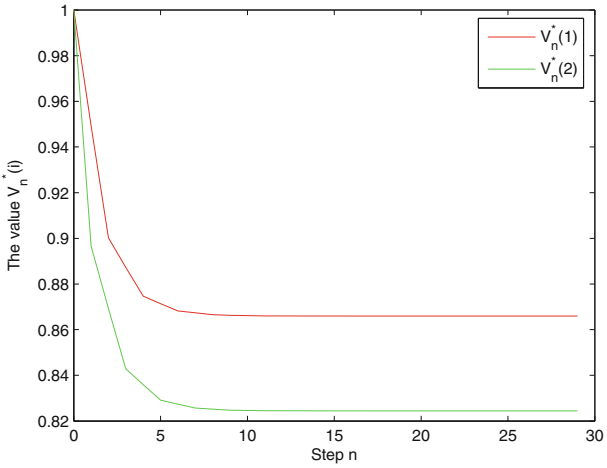

Fig. 3. The value function $V_n^*(i)$

Moreover, for $x = 1$, using Theorem 1, 2, Fig. 1 and Fig. 2, we know that

$$MV^*(1) = V^*(1) = M^{a_{11}}V^*(1).$$

For $x = 2$, we also obtain

$$MV^*(2) = V^*(2) = M^{a_{22}}V^*(2).$$

According to the above analysis and Theorem 2, we obtain the optimal stationary policy $f^*(1) = a_{12}, f^*(2) = a_{21}$ and the value function $V^*(1) = 0.8660, V^*(2) = 0.8245$.

Acknowledgement. This work was supported by National Natural Science Foundation of China (Grant No. 11961005, 11801590); Foundation of Guangxi Educational Committee (Grant No. KY2019YB0369); Ph.D. research startup foundation of Guangxi University of Science and Technology (Grant No. 18Z06); Guangxi Natural Science Foundation Program (Grant No. 2020GXNSFAA297196).

References

1. Baüerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Heidelberg (2011)
2. Baüerle, N., Rieder, U.: More risk-sensitive Markov decision processes. Math. Oper. Res. **39**, 105–120 (2014)
3. Cao, X.R.: Semi-Markov decision problems and performance sensitivity analysis. IEEE Trans. Autom. Control **48**, 758–769 (2003)
4. Cavazos-Cadena, R., Montes-De-Oca, R.: Optimal stationary policies in risk-sensitive dynamic programs with finite state space and nonnegative rewards. Appl. Math. (Warsaw) **27**, 167–185 (2000)
5. Cavazos-Cadena, R., Montes-De-Oca, R.: Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. Math. Meth. Oper. Res. **52**, 133–167 (2000)
6. Chung, K.J., Sobel, M.J.: Discounted MDP's: distribution functions and exponential utility maximization. SIAM J. Control Optim. **25**, 49–62 (1987)
7. Ghosh, M.K., Saha, S.: Risk-sensitive control of continuous time Markov chains. Stochastics **86**, 655–675 (2014)
8. Ghosh, M.K., Saha, S.: Non-stationary semi-Markov decision processes on a finite horizon. Stoch. Anal. Appl. **31**, 183–190 (2013)
9. Guo, X., Liu, Q.L., Zhang, Y.: Finite horizon risk-sensitive continuous-time Markov decision processes with unbounded transition and cost rates. 4OR **17**, 427–442 (2019)
10. Guo, X.P., Hernández-Lerma, O.: Continuous-Time Markov Decision Processes: Theory and Applications. Springer, Berlin (2009)
11. Hernández-Lerma, O., Lasserre, J.B.: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York (1996)
12. Howard, R.A., Matheson, J.E.: Risk-sensitive Markov decision processes. Manage. Sci. **18**, 356–369 (1972)
13. Huang, Y.H., Guo, X.P.: Discounted semi-Markov decision processes with nonnegative costs. Acta Math. Sin. (Chinese Ser.) **53**, 503–514 (2010)
14. Huang, Y.H., Guo, X.P.: Finite horizon semi-Markov decision processes with application to maintenance systems. Eur. J. Oper. Res. **212**, 131–140 (2011)
15. Huang, Y.H., Guo, X.P.: Mean-variance problems for finite horizon semi-Markov decision processes. Appl. Math. Optim. **72**, 233–259 (2015)
16. Huang, Y.H., Guo, X.P., Song, X.Y.: Performance analysis for controlled semi-Markov process. J. Optim. Theory Appl. **150**, 395–415 (2011)
17. Huang, Y.H., Lian, Z.T., Guo, X.P.: Risk-sensitive semi-Markov decision processes with general utilities and multiple criteria. Adv. Appl. Probab. **50**, 783–804 (2018)
18. Huang, X.X., Zou, X.L., Guo, X.P.: A minimization problem of the risk probability in first passage semi-Markov decision processes with loss rates. Sci. China Math. **58**, 1923–1938 (2015)

19. Huo, H.F., Zou, X.L., Guo, X.P.: The risk probability criterion for discounted continuous-time Markov decision processes. Discrete Event Dyn. Syst. **27**, 675–699 (2017)
20. Janssen, J., Manca, R.: Semi-Markov Risk Models for Finance, Insurance, and Reliability. Springer, New York (2006)
21. Jaquette, S.C.: A utility criterion for Markov decision processes. Manage. Sci. **23**, 43–49 (1976)
22. Jaśkiewicz, A.: A note on negative dynamic programming for risk-sensitive control. Oper. Res. Lett. **36**, 531–534 (2008)
23. Jaśkiewicz, A.: On the equivalence of two expected average cost criteria for semi Markov control processes. Math. Oper. Res. **29**, 326–338 (2013)
24. Limnios, N., Oprisan, G.: Semi-Markov Processes and Reliability. Birkhäuser, Boston (2001)
25. Luque-Vásquez, F., Minjárez-Sosa, J.A.: Semi-Markov control processes with unknown holding times distribution under a discounted criterion. Math. Meth. Oper. Res. **61**, 455–468 (2005)
26. Mamer, J.W.: Successive approximations for finite horizon semi-Markov decision processes with application to asset liquidation. Oper. Res. **34**, 638–644 (1986)
27. Nollau, V.: Solution of a discounted semi-Markovian decision problem by successive overrelaxation. Optimization **39**, 85–97 (1997)
28. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
29. Schäl, M.: Control of ruin probabilities by discrete-time investments. Math. Meth. Oper. Res. **70**, 141–158 (2005)
30. Wei, Q.D.: Continuous-time Markov decision processes with risk-sensitive finite-horizon cost criterion. Math. Meth. Oper. Res. **84**, 1–27 (2016)
31. Wei, Q.D., Guo, X.P.: New average optimality conditions for semi-Markov decision processes in Borel spaces. J. Optim. Theory Appl. **153**, 709–732 (2012)
32. Wei, Q.D., Guo, X.P.: Constrained semi-Markov decision processes with ratio and time expected average criteria in Polish spaces. Optimization **64**, 1593–1623 (2015)
33. Yushkevich, A.A.: On semi-Markov controlled models with average reward criterion. Theory Probab. Appl. **26**, 808–815 (1982)
34. Zhang, Y.: Continuous-time Markov decision processes with exponential utility. SIAM J. Control Optim. **55**, 1–24 (2017)