Discrete Event Dynamic Systems (2022) 32:545–571
https://doi.org/10.1007/s10626-022-00366-4

Zero-sum semi-Markov games with state-action-dependent discount factors

Zhihui Yu[1] · **Xianping Guo**[2,3] · **Li Xia**[1,3]

Received: 28 April 2021 / Accepted: 4 August 2022 / Published online: 5 November 2022

Abstract

Semi-Markov model is one of the most general models for stochastic dynamic systems. This paper deals with a two-person zero-sum game for semi-Markov processes. We focus on the expected discounted criterion with state-action-dependent discount factors. The state and action spaces are both Polish spaces, and the reward rate function is ω-bounded. We first construct a fairly general model of semi-Markov games under a given semi-Markov kernel and a pair of strategies. Next, based on the standard regularity condition and the continuity-compactness condition for semi-Markov games, we derive a "drift condition" on the semi-Markov kernel and suppose that the discount factors have a positive lower bound, under which the existence of the value function and an optimal pair of stationary strategies of our semi-Markov game are proved by using a general Shapley equation. Moreover, in the scenario of finite state and action spaces, a value iteration-type algorithm for approximating the value function and an optimal pair of stationary strategies is developed. The convergence and the error bound of the algorithm are also proved. Finally, we conduct numerical examples to demonstrate the main results.

Keywords Semi-Markov game · State-action-dependent discount factor · Optimal pair of stationary strategies · Value iteration-type algorithm

1 Introduction

This paper deals with two-person zero-sum *semi-Markov games* (SMGs) with expected discounted criterion, which is a generalization of discrete-time Markov games (DTMGs) (Shapley 1953), since the sojourn time between two consecutive decision epochs follows any distribution rather than a constant. Such games have already been studied in the literature

✉ Li Xia
xiali5@sysu.edu.cn

Xianping Guo
mcsgxp@mail.sysu.edu.cn

[1] School of Business, Sun Yat-sen University, Guangzhou, China

[2] School of Mathematics, Sun Yat-sen University, Guangzhou, China

[3] Guangdong Province Key Laboratory of Computational Science, Guangzhou, China

(Chen et al. 2021; Lal and Sinha 1992; Luque-Vásquez 2002a, b; Minjárez-Sosa and Luque-Vásquez 2008; Mondal et al. 2016, 2020). However, the existing references are all restricted to the case where the discount factor is a constant, which may not always hold in practice.

In this paper, we study a more general case of SMGs with varying discount factors to meet the often demanding nature of real-world problems. For example, considering the application in economics, the discount factor (interest rate) may depend both on economy environments and decision-makers' actions. That is, the interest rate usually varies in different financial markets and monetary policies, where financial markets can be considered as states and monetary policies are actions taken by the government. Problems with non-constant discount factors have been studied for Markov decision processes (MDPs) (Feinberg and Shwartz 1994; González-Hernández et al. 2009; Schäl 1975; Wei and Guo 2011; Ye and Guo 2012; Minjárez-Sosa 2015; Wu and Zhang 2016) and two-person zero-sum DTMGs (González-Sánchez et al. 2019). Furthermore, in light of the application of SMGs to the dynamic oligopoly model (Kirman and Sobel 1974) and the dynamic overlapping generations model (Raut 1990), it is required and reasonable to investigate SMGs with *state-action-dependent discount factors* to compensate for the inadequacies of theory and application.

On the other hand, most of the literature work on SMGs focuses on the existence of optimal strategies. However, how to efficiently solve a stochastic dynamic game and compute an optimal pair of stationary strategies are especially important for practical implementation of game theory. The classic algorithmic study on game theory focuses on matrix games, which can be solved by linear programming (LP) (Barron 2013). This LP technique further combines the policy iteration in MDPs to solve discounted two-person zero-sum DTMGs (Hoffman and Karp 1966; Pollatschek and Avi-Itzhak 1969), where LP is used to improve strategy at every iteration. Recently, there are emerging investigations that aim to study the efficient computation for stochastic dynamic games using approximation or learning algorithms. Littman (1994) proposes a minimax-Q algorithm to solve discounted two-person zero-sum DTMGs, which is essentially motivated by the standard Q-learning algorithm with a minimax operator in Markov games replacing the max operator in reinforcement learning. In addition, finite algorithms for some interesting special classes of stochastic games are also widely studied, such as LP to solve DTMGs with single-controller, switching controller and separable reward and state independent transition (Filar and Vrieze 2012). Recently, the same results are extended to SMGs. Mondal et al. (2016) study the discounted two-person zero-sum SMGs with AR-AT-AITT (Additive Reward and Additive Transition and Action Independent Transition Times) structure. They prove that such game can be formulated as a vertical linear complementarity problem (VLCP), which can be solved by the Cottle-Dantzig's algorithm. They further propose a policy improvement algorithm for solving a mixture class of perfect information and AR-AT SMG (Mondal et al. 2020). Notice that the above algorithms for solving SMGs are only suitable for some special structures, which fail to effectively solve general discounted two-person zero-sum SMGs.

In this paper, we aim at studying the two-person zero-sum SMGs with expected discounted criterion in which the discount factors are state-action-dependent. The objective is to find an optimal pair of stationary strategies to maximize the reward of player 1 (P1) and minimize the payoff of player 2 (P2). More precisely, we deal with the SMGs specified by five primitive data: the state space X; the action spaces A, B for P1 and P2, respectively; the semi-Markov kernel $Q(t, y|x, a, b)$; the discount factor $\alpha(x, a, b)$; and the reward rate function $r(x, a, b)$. The state space X and action spaces A, B are all Polish spaces, and the reward rate function $r(x, a, b)$ is *ω-bounded*. The semi-Markov kernel $Q(t, y|x, a, b)$ is a joint distribution of sojourn time and state for any given $(x, a, b) \in X \times A \times B$, which is more general in comparison to the literature (Luque-Vásquez 2002a, b; Minjárez-Sosa and

Luque-Vásquez 2008; Mondal et al. 2016, 2020). To the best of our knowledge, the framework with the common state-time distribution was just used in previous works dealing with semi-Markov decision processes (for non-negative reward rate function (Huang and Guo 2011, 2010); for average criterion with unbounded reward rate function (Ross 1970)) and SMGs (for bounded reward rate function (Chen et al. 2021; Lal and Sinha 1992; Tanaka and Wakuta 1976); for average criterion with unbounded reward rate function (Vega-Amaya et al. 2022)), which has not been studied in discounted SMGs with unbounded reward rate function. This extension necessitates the addition of a new "drift condition", which leads to a general Shapley equation and a more complicated proof of optimality. With these new features, we construct an SMG model with a fairly general problem setting. Then we impose suitable conditions on the model parameters shown in Assumptions 1-4, under which we establish the Shapley equation and prove the existence of the value function and an optimal pair of stationary strategies of the game. Our proof is quite different from González-Sánchez et al. (2019) since we directly search for optimal strategies with respect to history-dependent strategies instead of restricted to Markov strategies. Furthermore, in the scenario of finite state and action spaces, we derive a *value iteration-type algorithm* to approach to the value function and an optimal pair of stationary strategies of the game based on the Shapley equation. The convergence and the error bound analysis of the algorithm are also derived. Finally, we conduct numerical examples on an investment problem to demonstrate the main results of our paper.

The contributions of this paper can be summarized as follows. (1) This paper studies a fairly general model of SMGs with state-action-dependent discount factors and joint probability transition functions. We derive a "drift condition" (see Assumption 3) on the generic semi-Markov kernel, which is more general than the counterparts in the literature work (Luque-Vásquez 2002a, b; González-Sánchez et al. 2019), as stated in Remark 3. (2) In order to find an ε-optimal pair of stationary strategies and an approximate value function, a value iteration-type algorithm is proposed, which can be viewed as a combination of the value iteration of MDPs and the LP of matrix games. Moreover, the convergence and the error bound of the algorithm are also analyzed.

The rest of this paper is organized as follows. In Section 2, we introduce the model of SMG as well as the optimality criterion. Our main optimality results are stated in Section 3 and studied with the proof in Section 4. A value iteration-type algorithm for approximating an optimal pair of stationary strategies is developed in Section 5, and some numerical examples are conducted to demonstrate our main results in Section 6. Finally, we conclude the paper and discuss some future research topics in Section 7.

2 Two-person zero-sum semi-Markov game model

Notation: If E is a Polish space (that is, a complete and separable metric space), its Borel σ-algebra is denoted by $\mathcal{B}(E)$, and $\mathbb{P}(E)$ denotes the family of probability measures on $\mathcal{B}(E)$, endowed with the topology of weak convergence.

In this section, we introduce a two-person zero-sum SMG model with expected discounted criterion and state-action-dependent discount factors, which is denoted by the collection

$$\{X, A, B, (A(x), B(x), x \in X), Q(t, y|x, a, b), \alpha(x, a, b), r(x, a, b)\},$$

where the symbols are explained as follows.

- X is the state space which is a Polish space, and A, B are action spaces for P1 and P2, respectively, which are also supposed to be Polish spaces.
- $A(x)$ and $B(x)$ are Borel subsets of A and B, which represent the sets of the admissible actions for P1 and P2 at state $x \in X$, respectively. Let

$$K := \{(x, a, b) | x \in X, a \in A(x), b \in B(x)\}$$

be a measurable Borel subset of $X \times A \times B$.
- $Q(t, y|x, a, b)$ is a semi-Markov kernel which satisfies the following properties.

(a) For each fixed $(x, a, b) \in K$, $Q(\cdot, \cdot|x, a, b)$ is a probability measure on $[0, \infty) \times X$, whereas for each fixed $t \in [0, \infty)$, $D \in \mathcal{B}(X)$, $Q(t, D|\cdot, \cdot, \cdot)$ is a real-valued Borel function on K.

(b) For each fixed $(x, a, b) \in K$ and $D \in \mathcal{B}(X)$, $Q(\cdot, D|x, a, b)$ is a non-decreasing right-continuous real-valued Borel function on $[0, \infty)$ such that $Q(0, D|x, a, b) = 0$.

(c) For each fixed $(x, a, b) \in K$, we denote by

$$H(\cdot|x, a, b) := Q(\cdot, X|x, a, b)$$

the distribution function of the sojourn time at state $x \in X$ when the actions $a \in A(x), b \in B(x)$ are chosen. For each $x \in X$ and $D \in \mathcal{B}(X)$, when P1 and P2 select actions $a \in A(x)$ and $b \in B(x)$, respectively, $Q(t, D|x, a, b)$ denotes the joint probability that the sojourn time in state x is not greater than $t \in \mathbb{R}_+$ and the next state belongs to D.

- $\alpha(x, a, b)$ is a measurable function from K to $(0, \infty)$, which denotes the state-action-dependent discount factor.
- $r(x, a, b)$ is a real-valued function on K, which represents the reward/payoff rate function for P1/P2.

Remark 1 In our SMG model, the semi-Markov kernel $Q(\cdot, \cdot|x, a, b)$ is a joint probability distribution with respect to sojourn time and state for given $(x, a, b) \in K$. Thus, our model is more general than the counterpart in the literature (Lal and Sinha 1992; Luque-Vásquez 2002a, b; González-Sánchez et al. 2019).

The evolution of SMGs with the expected discounted criterion carries on as follows.

Assume that the game starts at the initial state $x_0 \in X$ at the initial decision epoch $t_0 := 0$. The two players choose simultaneously actions $a_0 \in A(x_0), b_0 \in B(x_0)$ according to the variables t_0 and x_0, then P1 and P2 receive immediate reward $r(x_0, a_0, b_0)$ and immediate payoff $r(x_0, a_0, b_0)$, respectively. Consequently, after staying at state x_0 up to time $t_1 > t_0$, the system moves to a new state $x_1 \in D$ according to the transition law $Q(t_1 - t_0, D|x_0, a_0, b_0)$. Once the state transition to x_1 occurs at the 1st decision epoch t_1, the entire process repeats again and the game evolves in this way.

Thus, we obtain an admissible history at the n-th decision epoch

$$h_n := (t_0, x_0, a_0, b_0, t_1, x_1, a_1, b_1, \ldots, t_n, x_n).$$

When the game goes to infinity, we obtain the history

$$h := (t_0, x_0, a_0, b_0, t_1, x_1, a_1, b_1, \ldots),$$

where $t_n \leq t_{n+1}$, $(x_n, a_n, b_n) \in K$ for all $n \geq 0$. Moreover, let $\mathcal{H}_n$ be the class of all admissible histories h_n of the system up to the n-th decision epoch, endowed with a Borel σ-algebra.

Let $(\Omega, \mathcal{F})$ be the canonical measurable space consisting of the sample space $\Omega = (\mathbb{R}_+ \times X \times A \times B)^\infty$ and the corresponding product σ-algebra $\mathcal{F}$. For each $\omega = (t_0, x_0, a_0, b_0, t_1, x_1, \dots) \in \Omega$, we define a stochastic process $\{T_n, X_n, A_n, B_n, t \geq 0\}$ on $(\Omega, \mathcal{F})$ by

$$T_n(\omega) = t_n,\ X_n(\omega) = x_n,\ A_n(\omega) = a_n,\ B_n(\omega) = b_n.$$

Moreover, we denote H_n the processes of decision time, state and action until the n-th decision epoch by

$$H_n(\omega) = (T_0, X_0, A_0, B_0, \dots, T_n, X_n)(\omega) = h_n.$$

To introduce our expected discounted criterion discussed in this paper, we give the definitions of strategies as follows.

Definition 1 A randomized history-dependent strategy for P1 is a sequence of stochastic kernels $\pi^1 := (\pi_n^1, n \geq 0)$ that satisfies the following conditions:

(i) for each $D \in \mathcal{B}(X)$, $\pi_n^1(D|\cdot)$ is a Borel function on $\mathcal{H}_n$, and for each $h_n \in \mathcal{H}_n$, $\pi_n^1(\cdot|h_n)$ is a probability measure on A;
(ii) $\pi_n^1(\cdot|h_n)$ is concentrated on $A(x_n)$, that is

$$\pi_n^1(A(x_n)|h_n) = 1, \quad \forall h_n \in \mathcal{H}_n \text{ and } n \geq 0.$$

We denote by Π_1 the set of all the randomized history-dependent strategies for P1 for simplicity.

Definition 2 (1) A strategy $\pi^1 = (\pi_n^1, n \geq 0) \in \Pi_1$ is called a randomized Markov strategy if there exists a sequence of stochastic kernels $\phi_1 = (\varphi_n, n \geq 0)$ such that

$$\pi_n^1(\cdot|h_n) = \varphi_n(\cdot|x_n), \quad \forall h_n \in \mathcal{H}_n \text{ and } n \geq 0.$$

(2) A randomized Markov strategy $\phi_1 = (\varphi_n, n \geq 0)$ is called stationary if φ_n is independent of n; that is, if there exists a stochastic kernel φ on A given x such that

$$\varphi_n(\cdot|x) \equiv \varphi(\cdot|x), \quad \forall x \in X \text{ and } n \geq 0.$$

That is, $\phi_1 = (\varphi, \varphi, \dots) =: \varphi^\infty$. For convenience, we still use φ to denote the randomized stationary strategy without special declaration.

(3) Moreover, if $\varphi(\cdot|x)$ is a Dirac measure for all $x \in X$, then the stationary strategy φ is called a pure strategy.

We denote by Π_1^M, Φ_1 and Π_1^{MD} the sets of all the randomized Markov strategies, randomized stationary strategies and pure strategies for P1, respectively.

The sets of all randomized history-dependent strategies Π_2, randomized Markov strategies Π_2^M, randomized stationary strategies Φ_2, pure strategies Π_2^{MD} for P2 are defined similarly, with $B(x)$ in lieu of $A(x)$. Clearly, $\Pi_1^{MD} \subset \Phi_1 \subset \Pi_1^M \subset \Pi_1$ and $\Pi_2^{MD} \subset \Phi_2 \subset \Pi_2^M \subset \Pi_2$.

For each $x \in X, \pi^1 \in \Pi_1, \pi^2 \in \Pi_2$, by Theorem of C. Ionescu-Tulcea (Hernández-Lerma and Lasserre 2012a, P.178), there exist a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_x^{\pi^1,\pi^2})$ and a stochastic process $\{T_n, X_n, A_n, B_n, n \geq 0\}$ such that for each $D \in \mathcal{B}(X)$, $D_1 \in \mathcal{B}(A)$, $D_2 \in \mathcal{B}(B)$ and $n \geq 0$, we have

$$\mathbb{P}_x^{\pi^1,\pi^2}(X_0 = x) = 1,$$

$$\mathbb{P}_x^{\pi^1,\pi^2}(A_n \in D_1, B_n \in D_2|h_n) = \pi_n^1(D_1|h_n)\pi_n^2(D_2|h_n),$$

$$\mathbb{P}_x^{\pi^1,\pi^2}(T_{n+1} - T_n \le t, X_{n+1} \in D | h_n, a_n, b_n) = Q(t, D | x_n, a_n, b_n).$$

Here and in what follows, we denote by $\mathbb{E}_x^{\pi^1,\pi^2}$ the expectation operator with respect to $\mathbb{P}_x^{\pi^1,\pi^2}$.

To avoid the possibility of infinitely numerous decision epochs during the finite time interval, we take an assumption on the semi-Markov kernel, which is also used in Lal and Sinha (1992), Luque-Vásquez (2002a, b), and the references therein.

Assumption 1 *There exist constants $\theta > 0$ and $\delta > 0$ such that*

$$H(\theta | x, a, b) \le 1 - \delta, \quad \forall (x, a, b) \in K.$$

Proposition 1 *If Assumption 1 holds, then for each fixed $x \in X$ and $\pi^1 \in \Pi_1, \pi^2 \in \Pi_2$, we have*

$$\mathbb{P}_x^{\pi^1,\pi^2}(\lim_{n\to\infty} T_n = \infty) = 1.$$

The proof of Proposition 1 is provided in Appendix B.

Since $T_n \xrightarrow{P} \infty$, it is not required to consider the processes for $t > T_\infty = \lim_{n\to\infty} T_n$. Now, we establish an underlying continuous-time state-action process $\{X(t), A(t), B(t), t \ge 0\}$, which corresponds to the stochastic process $\{T_n, X_n, A_n, B_n, n \ge 0\}$ with probability space $(\Omega, \mathcal{F}, \mathbb{P}_x^{\pi^1,\pi^2})$,

$$X(t) = \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \le t < T_{n+1}\}} X_n,$$

$$A(t) = \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \le t < T_{n+1}\}} A_n,$$

$$B(t) = \sum_{n=0}^{\infty} \mathbb{1}_{\{T_n \le t < T_{n+1}\}} B_n,$$

where $\mathbb{1}_E$ is an indicator function on any set E.

Definition 3 The stochastic process $\{X(t), A(t), B(t), t \ge 0\}$ is called a semi-Markov game.

Next, we will show the definition of the expected discounted criterion in this paper.

Definition 4 For each $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, the initial state $x \in X$ and discount factor $\alpha(\cdot) > 0$, the expected discounted reward/payoff for P1/P2 is defined as follows:

$$V(x, \pi^1, \pi^2) := \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_0^\infty e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt\Big]. \quad (1)$$

P1 aims to maximize the reward while P2 aims to minimize the payoff. Both players aim to find an optimal strategy.

Definition 5 The upper value and lower value of the expected discounted SMG are defined as

$$U(x) := \inf_{\pi^2\in\Pi_2} \sup_{\pi^1\in\Pi_1} V(x, \pi^1, \pi^2) \text{ and } L(x) := \sup_{\pi^1\in\Pi_1} \inf_{\pi^2\in\Pi_2} V(x, \pi^1, \pi^2),$$

respectively. Obviously, $U(x) \geq L(x)$ for all $x \in X$. Moreover, if it holds that $L(x) = U(x)$ for all $x \in X$, then the common function is called the value function of the game and denoted by V^*.

Definition 6 Assume that the game has a value function $V^*(\cdot)$. Then a strategy $\pi_*^1 \in \Pi_1$ is said to be optimal for P1 if

$$\inf_{\pi^2 \in \Pi_2} V(x, \pi_*^1, \pi^2) = V^*(x), \quad \forall x \in X.$$

Similarly, $\pi_*^2 \in \Pi_2$ is said to be optimal for P2 if

$$\sup_{\pi^1 \in \Pi_1} V(x, \pi^1, \pi_*^2) = V^*(x), \quad \forall x \in X.$$

If π_*^i is optimal for player i $(i = 1, 2)$, then we call (π_*^1, π_*^2) an optimal pair of strategies.

Remark 2 (π_*^1, π_*^2) is an optimal pair of strategies if and only if

$$V(x, \pi^1, \pi_*^2) \leq V(x, \pi_*^1, \pi_*^2) \leq V(x, \pi_*^1, \pi^2), \quad \forall x \in X, \pi^1 \in \Pi_1, \pi^2 \in \Pi_2.$$

Remark 2 is an effective method to verify whether a pair of strategies (π^1, π^2) is an optimal pair of strategies, which is widely used in the literature of two-person zero-sum stochastic games; see, for instance, González-Sánchez et al. (2019), Luque-Vásquez (2002a, b), and the references therein.

3 Main results

This section focuses on the existence of the value function and an optimal pair of stationary strategies, which requires imposing suitable assumptions on the model parameters. We first give some notations for convenience.

Given a measurable function $\omega : X \to [1, \infty)$, a function u on X is said to be ω-bounded if it has finite ω-norm which is defined as

$$\|u\|_\omega := \sup_{x \in X} \frac{|u(x)|}{\omega(x)},$$

such a function ω can be referred to as a weight function. For convenience, we denote by $B_\omega(X)$ the Banach space of all ω-bounded measurable functions on X.

For each given function $u \in B_\omega(X)$ and $(x, a, b) \in K$, we write

$$G(u, x, a, b) := r(x, a, b) \int_0^\infty e^{-\alpha(x,a,b)t}(1 - H(t|x, a, b))dt + \int_0^\infty e^{-\alpha(x,a,b)t} \int_X u(y) Q(dt, dy|x, a, b). \tag{2}$$

For each fixed $x \in X$ and probability measures $\mu \in \mathbb{A}(x) := \mathbb{P}(A(x))$ and $\lambda \in \mathbb{B}(x) := \mathbb{P}(B(x))$, we denote

$$G(u, x, \mu, \lambda) := \int_{A(x)} \int_{B(x)} G(u, x, a, b)\mu(da)\lambda(db),$$

whenever the integral is well defined.

We define an operator T on $B_\omega(X)$ by

$$Tu(x) := \sup_{\mu\in\mathbb{A}(x)} \inf_{\lambda\in\mathbb{B}(x)} G(u, x, \mu, \lambda), \quad \forall x \in X, \tag{3}$$

which is called the Shapley operator. A function $v \in B_\omega(X)$ is said to be a solution of the Shapley equation if

$$Tv(x) = v(x), \quad \forall x \in X.$$

In order to explore the existence of an optimal pair of stationary strategies, we also need to define a stationary-strategy-dependent operator $T(\varphi_1, \varphi_2)$ on $B_\omega(X)$ by

$$T(\varphi_1, \varphi_2)u(x) := G(u, x, \varphi_1(\cdot|x), \varphi_2(\cdot|x)), \quad \forall x \in X,$$

where $(\varphi_1, \varphi_2) \in \Phi_1 \times \Phi_2$ is a pair of stationary strategies.

Next, we give some hypotheses to guarantee the existence of an optimal pair of stationary strategies. The framework settled by these hypotheses has became quite standard for the study of semi-Markov models (Chen et al. 2021; Huang and Guo 2011, 2010; Lal and Sinha, 1992; Luque-Vásquez, 2002a, b; Minjárez-Sosa and Luque-Vásquez 2008) and varying discount factors (Feinberg and Shwartz 1994; González-Hernández et al. 2009; González-Sánchez et al. 2019; Minjárez-Sosa 2015; Wei and Guo 2011; Ye and Guo 2012; Wu and Zhang 2016).

Assumption 2 *(a) There exists a constant $\alpha_0 > 0$ such that $\alpha(x, a, b) \geq \alpha_0$ for all $(x, a, b) \in K$. (b) There exists a measurable function $\omega : X \to [1, \infty)$ and a non-negative constant M such that for all $(x, a, b) \in K$,*

$$|r(x, a, b)| \leq M\omega(x).$$

The key point is that Assumption 2 entails a finiteness property of expected discounted reward. Below we give an important consequence of Assumption 1 and Assumption 2(a).

Lemma 1 *If Assumptions 1&2(a) hold, then there exists a constant $0 < \gamma < 1$ such that for each $(x, a, b) \in K$,*

$$\int_0^\infty e^{-\alpha(x,a,b)t} H(dt|x, a, b) \leqslant \gamma \tag{4}$$

The proof of Lemma 1 is provided in Appendix A.

Assumption 3 *There exists a constant η with $0 < \eta\gamma < 1$ such that for each fixed $t \geq 0$ and $(x, a, b) \in K$,*

$$\int_X \omega(y) Q(t, dy|x, a, b) \leq \eta\omega(x) H(t|x, a, b), \tag{5}$$

where $\omega(\cdot)$ is the function mentioned in Assumption 2.

Remark 3 (1) We call Assumption 3 the "drift condition", which is needed to ensure that the Shapley operator (defined in (3)) is a contraction operator with respect to a weighted norm as well as our main results. Obviously, Assumption 3 naturally holds when r is bounded by taking $\omega(\cdot) \equiv 1$ and $\eta = 1$.

(2) Particularly, if we set $Q(t, y|x, a, b) = H(t|x, a, b)P(y|x, a, b)$, where $P(y|x, a, b)$ denotes the state transition probability, then (5) degenerates into $\int_X \omega(y) P(dy|x, a, b) \leq \eta\omega(x)$, which is the same as the Assumption 3(b) of Luque-Vásquez (2002a) and Assumption 1(e) of González-Sánchez et al. (2019). Thus, our

Assumption 3 is more general than the counterpart in the literature Luque-Vásquez (2002a) and the Example 3 of González-Sánchez et al. (2019) about SMG is a special case of ours.

(3) Combining Lemma 1 with Assumption 3, it is easy to derive

$$\int_0^\infty e^{-\alpha(x,a,b)t} \int_X u(y) Q(dt, dy|x, a, b) \leq \eta\gamma \|u\|_\omega \omega(x), \quad \forall u \in B_\omega(X), (x, a, b) \in K. \tag{6}$$

Moreover, we impose the following continuity-compactness conditions to ensure the existence of an optimal pair of stationary strategies of our SMG model.

Assumption 4 (a) *For each fixed $x \in X$, both $A(x)$ and $B(x)$ are compact sets.*

(b) *For each fixed $(x, a, b) \in K$, $r(x, \cdot, b)$ is upper semi-continuous on $A(x)$ and $r(x, a, \cdot)$ is lower semi-continuous on $B(x)$.*

(c) *For each fixed $(x, a, b) \in K$, $t \geq 0$ and $v \in B_\omega(X)$, the functions*

$$a \longmapsto \int v(y) Q(t, dy|x, a, b) \quad \text{and} \quad b \longmapsto \int v(y) Q(t, dy|x, a, b)$$

are continuous on $A(x)$ and $B(x)$, respectively.

(d) *For each fixed $t \geq 0$, $H(t|\cdot, \cdot)$ is continuous on K.*

(e) *The function $\alpha(x, a, b)$ is continuous on K.*

Remark 4 (1) Assumption 4 is similar to the standard continuity-compactness hypotheses for Markov control processes; see, for instance, Hernández-Lerma and Lasserre (2012b), and the references therein. It is commonly used for the existence of minimax points of games.

(2) By Lemma 1.11 in Nowak (1984), if Assumption 4(a) holds, then the probability spaces $\mathbb{A}(x)$ and $\mathbb{B}(x)$ are also compact for each $x \in X$.

(3) These continuity-compactness conditions are specifically applied to infinite state and action spaces, which obviously hold when S and A, B are finite.

Now, we present our main results, Theorem 1 below, which extends to SMGs with common state-time distribution and state-action dependent discount factors the analysis given in Luque-Vásquez (2002a) for SMGs with constant discount factor.

Theorem 1 *Suppose that Assumptions 1-4 hold, then*

(a) *The SMG has a value function $V^*(\cdot)$, which is the unique function in $B_\omega(X)$ that satisfies the Shapley equation, i.e.,*

$$V^*(x) = TV^*(x), \quad \forall x \in X,$$

and furthermore, there exists an optimal pair of stationary strategies.

(b) *A pair of stationary strategies $(\varphi_1^*, \varphi_2^*) \in \Phi_1 \times \Phi_2$ is optimal if and only if its expected discounted reward satisfies the Shapley equation, i.e., $TV(x, \varphi_1^*, \varphi_2^*) = V(x, \varphi_1^*, \varphi_2^*)$ for all $x \in X$.*

4 Preliminaries and proofs

In this section, we present some results needed to prove Theorem 1. Some of these results are already known in the literature (Luque-Vásquez 2002a, b), but we state them here for completeness and ease of reference.

Lemma 2 *Suppose that Assumptions 1-4 hold, then for each given function $u \in B_\omega(X)$, the function Tu is in $B_\omega(X)$ and*

$$Tu(x) := \min_{\lambda\in\mathbb{B}(x)}\max_{\mu\in\mathbb{A}(x)} G(u,x,\mu,\lambda). \tag{7}$$

Moreover, there exists a pair of stationary strategies $(\varphi_1,\varphi_2) \in \Phi_1\times\Phi_2$ such that

$$\begin{aligned} Tu(x) &= G(u,x,\varphi_1(\cdot|x),\varphi_2(\cdot|x)) \\ &= \max_{\mu\in\mathbb{A}(x)} G(u,x,\mu,\varphi_2(\cdot|x)) \\ &= \min_{\lambda\in\mathbb{B}(x)} G(u,x,\varphi_1(\cdot|x),\lambda). \end{aligned} \tag{8}$$

Lemma 3 *Both T and $T(\varphi_1,\varphi_2)$ are contraction operators with modulus less than 1.*

Since T and $T(\varphi_1,\varphi_2)$ are both contraction operators, there exist unique functions $u^* \in B_\omega(X)$ and $u^*_{\varphi_1,\varphi_2} \in B_\omega(X)$ such that $Tu^*(\cdot) = u^*(\cdot)$ and $T(\varphi_1,\varphi_2)u^*_{\varphi_1,\varphi_2}(\cdot) = u^*_{\varphi_1,\varphi_2}(\cdot)$ by the Banach's fixed point theorem.

Before stating the next important result, we give the definition of the m-shift strategy (Hernández-Lerma and Lasserre 2012b, P.96).

Definition 7 Given a strategy $\pi^i = \{\pi^i_n, n = 0,1,\ldots\} \in \Pi_i, i = 1,2$, and an integer $m \geq 0$, the corresponding m-shift strategy ${}^{(m)}\pi^i = \{{}^{(m)}\pi^i_n, n = 0,1,\ldots\}$ is given by

$${}^{(m)}\pi^i_0(\cdot|x_m) := \pi^i_m(\cdot|h_m),$$

and for $n = 1,2,\ldots,$

$${}^{(m)}\pi^i_n(\cdot|x_m,a_m,b_m,\ldots,x_{m+n}) := \pi^i_{m+n}(\cdot|h_m,a_m,b_m,\ldots,x_{m+n}),$$

where $h_m := (t_0,x_0,a_0,b_0,\ldots,t_{m-1},x_{m-1},a_{m-1},b_{m-1},t_m,x_m)$ denotes the admissible history at the m-th decision epoch.

Lemma 4 *For each $(\pi^1,\pi^2) \in \Pi_1\times\Pi_2$ and $x \in X$,*

$$V(x,\pi^1,\pi^2) = T(\pi^{1\infty}_0,\pi^{2\infty}_0)V(x,{}^{(1)}\pi^1,{}^{(1)}\pi^2)$$

where ${}^{(1)}\pi^i := (\pi^i_n, n \geq 1)$ is the 1-shift strategy of $\pi^i := (\pi^i_n, n \geq 0)$, $\pi^{i\infty}_0 = (\pi^i_0,\pi^i_0,\cdots)$, $i = 1,2$, and $(\pi^{1\infty}_0,\pi^{2\infty}_0)$ is a pair of stationary strategies.

Now, if we set $\pi^1 = \varphi_1$ and $\pi^2 = \varphi_2$ specially, which are both stationary strategies, from Lemma 4, we have

$$V(x,\varphi_1,\varphi_2) = T(\varphi_1,\varphi_2)V(x,\varphi_1,\varphi_2), \quad \forall x \in X,$$

which implies that the function $V(x,\varphi_1,\varphi_2)$ is the unique fixed point of the contraction operator $T(\varphi_1,\varphi_2)$.

Lemma 5 *Suppose that Assumptions 1-3 hold, let* $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, *then for each* $x \in X$, $u \in B_\omega(X)$, *we have*

$$\lim_{n\to\infty} \mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\int_0^{T_n} \alpha(X(s),A(s),B(s))ds} u(X_n)\Big] = 0.$$

Proofs of Lemmas 2-5 are provided in Appendix A. Next, a complete proof of Theorem 1 is given by applying Lemmas 2-5.

Proof of Theorem 1 (a) Let u^* be the unique fixed point of T in $B_\omega(X)$, that is

$$u^*(x) = Tu^*(x), \quad \forall x \in X.$$

By Lemma 2, there exists a pair of stationary strategies $(\varphi_1, \varphi_2) \in \Phi_1 \times \Phi_2$ such that for each $x \in X$,

$$\begin{aligned} Tu^*(x) &= G(u^*, x, \varphi_1(\cdot|x), \varphi_2(\cdot|x)) \\ &= \max_{\mu\in\mathbb{A}(x)} G(u^*, x, \mu, \varphi_2(\cdot|x)) \\ &= \min_{\lambda\in\mathbb{B}(x)} G(u^*, x, \varphi_1(\cdot|x), \lambda), \end{aligned} \tag{9}$$

which implies that

$$u^*(x) = G(u^*, x, \varphi_1(\cdot|x), \varphi_2(\cdot|x)) = T(\varphi_1, \varphi_2)u^*(x), \quad \forall x \in X.$$

Moreover, by Lemma 4,

$$V(x, \varphi_1, \varphi_2) = T(\varphi_1, \varphi_2)V(x, \varphi_1, \varphi_2), \quad \forall x \in X,$$

from which we can derive

$$u^*(x) = V(x, \varphi_1, \varphi_2), \quad \forall x \in X.$$

Next, we prove that u^* is the value function of the game and (φ_1, φ_2) is an optimal pair of stationary strategies, that is

$$V(x, \varphi_1, \pi^2) \geq V(x, \varphi_1, \varphi_2) \geq V(x, \pi^1, \varphi_2)\,, \quad \forall (\pi^1, \pi^2) \in \Pi_1 \times \Pi_2. \tag{10}$$

We just prove the first inequality in (10). Then a similar proof can follow for the second inequality. By (9), we have

$$u^*(x) \leq G(u^*, x, \varphi_1(\cdot|x), \lambda)\,, \quad \forall \lambda \in \mathbb{B}(x).$$

Particularly, let λ be an indicator function such that $\lambda(db) = 1$. Then for each $b \in B(x)$, we have

$$\begin{aligned} u^*(x) \leq \int_{A(x)} \Bigg\{ & r(x, a, b) \int_0^\infty e^{-\alpha(x,a,b)t}\Big[1 - H(t|x, a, b)\Big]dt \\ & + \int_0^\infty e^{-\alpha(x,a,b)t}\Big[\int_X u^*(y)Q(dt, dy|x, a, b)\Big]\Bigg\} \varphi_1(da|x). \end{aligned}$$

For each given $h_n = (t_0, x_0, a_0, b_0, \ldots, t_n, x_n) \in \mathcal{H}_n$ and $b_n \in B(x_n)$, we have

$$\begin{aligned} u^*(x_n) \leq \int_{A(x_n)} \Bigg\{ & r(x_n, a_n, b_n) \int_0^\infty e^{-\alpha(x_n,a_n,b_n)t}\Big[1 - H(t|x_n, a_n, b_n)\Big]dt \\ & + \int_0^\infty e^{-\alpha(x_n,a_n,b_n)t}\Big[\int_X u^*(y)Q(dt, dy|x_n, a_n, b_n)\Big]\Bigg\} \varphi_1(da_n|X_n = x_n). \end{aligned}$$

For $\forall \pi^2 \in \Pi_2$, integrating b_n on both sides in the above inequality, we have

$$\begin{aligned} u^*(x_n) \le & \int_{B(x_n)} \int_{A(x_n)} \left\{ \int_0^\infty e^{-\alpha(x_n,a_n,b_n)t} \Big[\int_X u^*(y) Q(dt, dy|x_n, a_n, b_n) \Big] + \right. \\ & \left. r(x_n, a_n, b_n) \int_0^\infty e^{-\alpha(x_n,a_n,b_n)t} \Big[1 - H(t|x_n, a_n, b_n) \Big] dt \right\} \varphi_1(da_n|X_n = x_n) \pi_n^2(db_n|H_n = h_n) \\ = & \ \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_{T_n}^{T_{n+1}} \alpha(X(s),A(s),B(s))ds} u^*(X_{n+1})|H_n = h_n \Big] \\ & + \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_{T_n}^{T_{n+1}} e^{-\int_{T_n}^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt|H_n = h_n \Big]. \end{aligned}$$

Notice that the above inequality holds for all $h_n \in \mathcal{H}_n$, taking x_n as a random variable X_n and we have

$$\begin{aligned} u^*(X_n) \le & \ \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_{T_n}^{T_{n+1}} \alpha(X(s),A(s),B(s))ds} u^*(X_{n+1})|H_n \Big] \\ & + \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_{T_n}^{T_{n+1}} e^{-\int_{T_n}^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt|H_n \Big]. \end{aligned}$$

Multiplying $e^{-\int_0^{T_n} \alpha(X(s),A(s),B(s))ds}$ on both sides in the above inequality and using the properties of the conditional expectation, we have

$$\begin{aligned} e^{-\int_0^{T_n} \alpha(X(s),A(s),B(s))ds} u^*(X_n) \le & \ \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_0^{T_{n+1}} \alpha(X(s),A(s),B(s))ds} u^*(X_{n+1})|H_n \Big] \\ & + \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_{T_n}^{T_{n+1}} e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt|H_n \Big]. \end{aligned}$$

Then, taking the expectation $\mathbb{E}_x^{\varphi_1,\pi^2}$, we have

$$\begin{aligned} \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_0^{T_n} \alpha(X(s),A(s),B(s))ds} u^*(X_n) \Big] \le & \ \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_0^{T_{n+1}} \alpha(X(s),A(s),B(s))ds} u^*(X_{n+1}) \Big] \\ & + \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_{T_n}^{T_{n+1}} e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt \Big]. \end{aligned}$$

Now, summing over $n = 0, 1, 2, \ldots, N$, we obtain

$$\begin{aligned} u^*(x) \le & \ \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_0^{T_{N+1}} e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt \Big] \\ & + \mathbb{E}_x^{\varphi_1,\pi^2} \Big[e^{-\int_0^{T_{N+1}} \alpha(X(s),A(s),B(s))ds} u^*(X_{N+1}) \Big]. \end{aligned}$$

Letting $N \to \infty$, according to Lemma 5, we derive

$$u^*(x) \le \mathbb{E}_x^{\varphi_1,\pi^2} \Big[\int_0^\infty e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t), A(t), B(t))dt \Big],$$

which means that the first inequality in (10) holds.

(b) ($\Rightarrow$)

Suppose that $(\varphi_1^*, \varphi_2^*) \in \Phi_1 \times \Phi_2$ is an optimal pair of stationary strategies, then for each $x \in X$, $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, we have

$$V(x, \varphi_1^*, \pi^2) \ge V(x, \varphi_1^*, \varphi_2^*) \ge V(x, \pi^1, \varphi_2^*).$$

For each fixed $\lambda \in \mathbb{B}(x)$, let $\pi^2 = \{\pi_n^2, n \geq 0\}$ with $\pi_0^2 = \lambda$ and $\pi_n^2 = \varphi_2^*, n \geq 1$, then by Lemma 4, for each $x \in X$, we have

$$V(x, \varphi_1^*, \varphi_2^*) \leq V(x, \varphi_1^*, \pi^2) = T(\varphi_1^*, \lambda)V(x, \varphi_1^*, \varphi_2^*),$$

which yields

$$V(x, \varphi_1^*, \varphi_2^*) \leq \min_{\lambda \in \mathbb{B}(x)} T(\varphi_1^*, \lambda)V(x, \varphi_1^*, \varphi_2^*) \leq TV(x, \varphi_1^*, \varphi_2^*).$$

Similarly, we can prove

$$V(x, \varphi_1^*, \varphi_2^*) \geq TV(x, \varphi_1^*, \varphi_2^*).$$

Combining the last two inequalities, we obtain the desired result.

($\Leftarrow$)

This part holds, which has been proved in part (a).

5 Algorithm

In this section, for the feasible implementation of algorithms, the state and action spaces are supposed to be finite. To avoid excessive symbols, we still use ω-norm by taking $\omega(\cdot) \equiv 1$, which represents the infinity norm, i.e., $\|u\|_\omega = \sup_{x \in X} |u(x)| = \|u\|_\infty$, $\forall u \in B_\omega(X)$. We develop an iterative algorithm to approach to the value function and an optimal pair of stationary strategies of our two-person zero-sum SMG, where numerically solving matrix games is iteratively utilized at every state in a form of value iteration.

Without loss of generality, we assume that $A(x) := \{a_1, a_2, \ldots, a_{m_1}\}$ and $B(x) := \{b_1, b_2, \ldots, b_{m_2}\}$, for any $x \in X := \{x_0, x_1, \ldots, x_{n-1}\}$. Under Assumptions 1-4 presented in Section 3, we rewrite the Shapley equation in a matrix form as follows

$$\begin{aligned} V^*(x) &= \min_{\varphi_2(\cdot|x) \in \mathbb{B}(x)} \max_{\varphi_1(\cdot|x) \in \mathbb{A}(x)} G(V^*, x, \varphi_1(\cdot|x), \varphi_2(\cdot|x)) \\ &= \min_{\boldsymbol{\varphi}_2(x) \in \mathbb{B}(x)} \max_{\boldsymbol{\varphi}_1(x) \in \mathbb{A}(x)} \boldsymbol{\varphi}_1^T(x) \boldsymbol{\mathcal{G}}(V^*, x) \boldsymbol{\varphi}_2(x) \\ &= \max_{\boldsymbol{\varphi}_1(x) \in \mathbb{A}(x)} \min_{\boldsymbol{\varphi}_2(x) \in \mathbb{B}(x)} \boldsymbol{\varphi}_1^T(x) \boldsymbol{\mathcal{G}}(V^*, x) \boldsymbol{\varphi}_2(x), \quad \forall x \in X, \end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{G}}(u, x)$ denotes an $m_1 \times m_2$-dimensional matrix for each fixed $x \in X$ and given function $u \in B_\omega(X)$, with elements defined as

$$\boldsymbol{\mathcal{G}}(u, x)_{ij} := G(u, x, a_i, b_j), \quad i = 1, 2, \ldots, m_1; \ j = 1, 2, \ldots, m_2,$$

and $\boldsymbol{\varphi}_k(x) := (\varphi_k(1|x), \ldots, \varphi_k(m_k|x))^T$ is an m_k-dimensional column vector for $k = 1, 2$. It is obviously that the last equation of (11) can be viewed as a matrix game with reward matrix $\boldsymbol{\mathcal{G}}(V^*, x)$ at each state $x \in X$.

However, we cannot directly solve (11) since the value function V^* is unknown. Below, we develop Algorithm 1 to iteratively compute a series of matrix games whose values can asymptotically approach to $V^*(x)$ at each state x. From the lines 11-12 of Algorithm 1, we can see that at the n-th iteration, $V_n(x)$ and $(\boldsymbol{\varphi}_1^n(x), \boldsymbol{\varphi}_2^n(x))$ are obtained by using LP to solve the game with reward matrix $\boldsymbol{\mathcal{G}}(V_{n-1}, x)$. This iterative procedure of computing a series of V_n is similar to the classic value iteration algorithm in the MDP theory. Considering that iterative algorithms usually converge to approximate solutions in finite steps, we give the definition of ε-optimal.

Algorithm parameter: a small threshold $\epsilon > 0$ determining accuracy of estimation; model parameters θ, δ given by Assumption 1, α_0 given by Assumption 2(a), and $\eta = 1$ given by Assumption 3, with $\gamma = 1 - \delta + \delta e^{-\alpha_0 \theta}$ and $\varepsilon = \frac{\epsilon}{1-\eta\gamma}$; a measurable function $\omega(\cdot) \equiv 1$ given by Assumption 2(b)
Initialize: $V(x) \in \mathbb{R}$ for all $x \in X$ arbitrarily
repeat
  $\Delta \leftarrow 0$
  Loop for *each* $x \in X$ **do**
    $v \leftarrow V(x)$
    for $i = 1; i < m_1; i++$ **do**
      for $j = 1; j < m_2; j++$ **do**
        $\mathcal{G}(v, x)_{ij} \leftarrow G(v, x, a_i, b_j)$
    Solving the game with matrix $\mathcal{G}(v, x)$
    $V(x) \leftarrow \max_{\boldsymbol{\varphi}_1(x) \in \mathbb{A}(x)} \min_{\boldsymbol{\varphi}_2(x) \in \mathbb{B}(x)} \boldsymbol{\varphi}_1^T(x) \mathcal{G}(v, x) \boldsymbol{\varphi}_2(x)$
    $(\boldsymbol{\varphi}_1(x), \boldsymbol{\varphi}_2(x)) \leftarrow \arg \max_{\boldsymbol{\varphi}_1(x) \in \mathbb{A}(x)} \min_{\boldsymbol{\varphi}_2(x) \in \mathbb{B}(x)} \boldsymbol{\varphi}_1^T(x) \mathcal{G}(v, x) \boldsymbol{\varphi}_2(x)$
    $\Delta \leftarrow \max\{\Delta, \frac{|v - V(x)|}{\omega(x)}\}$
until $\Delta < \epsilon$;
Output:
  $V_\varepsilon(x) = V(x)$ and $(\boldsymbol{\varphi}_1^\varepsilon(x), \boldsymbol{\varphi}_2^\varepsilon(x)) = (\boldsymbol{\varphi}_1(x), \boldsymbol{\varphi}_2(x))$

Algorithm 1 Value iteration-type algorithm to solve the two-person zero-sum SMG.

Definition 8 Assume that the game has a value function V^*. Then a pair of strategies $(\pi_\varepsilon^1, \pi_\varepsilon^2) \in \Pi_1 \times \Pi_2$ is said to be ε-optimal of the game if

$$\|V(\cdot, \pi_\varepsilon^1, \pi_\varepsilon^2) - V^*(\cdot)\|_\omega < \varepsilon.$$

Moreover, $V_\varepsilon(\cdot) := V(\cdot, \pi_\varepsilon^1, \pi_\varepsilon^2)$ is called the ε-value function of the game.

Furthermore, we derive Theorem 2 to guarantee the convergence of Algorithm 1.

Theorem 2 *Under Algorithm 1, for any given $\epsilon > 0$ and initial value $V_0 \in \mathbb{R}$, there exists a non-negative integer $N_\epsilon = \left(1 + \lfloor log_{\eta\gamma}(\frac{\epsilon}{\|TV_0 - V_0\|_\omega}) \rfloor\right) \mathbb{1}_{TV_0 \neq V_0}$ such that $\|V_{N_\epsilon+1} - V_{N_\epsilon}\|_\omega < \epsilon$, which implies that Algorithm 1 can converge within N_ϵ iterations. Moreover, the strategy pair $(\varphi_1^\varepsilon, \varphi_2^\varepsilon)$ output by Algorithm 1 is ε-optimal, where $\varepsilon = \frac{\epsilon}{1-\eta\gamma}$.*

Proof According to the iterative formula of Algorithm 1, we have

$$\|V_{n+1} - V_n\|_\omega = \|TV_n - TV_{n-1}\|_\omega \leq \eta\gamma \|V_n - V_{n-1}\|_\omega, \quad \forall n \geq 1,$$

which by iteration yields

$$\|V_{n+1} - V_n\|_\omega \leq (\eta\gamma)^n \|TV_0 - V_0\|_\omega, \quad \forall n \geq 0.$$

For each given $\epsilon > 0$ and initial value $V_0 \in \mathbb{R}$, if $TV_0 = V_0$, choose $N_\epsilon = 0$, and we have

$$\|V_{N_\epsilon+1} - V_{N_\epsilon}\|_\omega = 0 < \epsilon,$$

otherwise, if $TV_0 \neq V_0$, choose $N_\epsilon = 1 + \lfloor log_{\eta\gamma}(\frac{\epsilon}{\|TV_0 - V_0\|_\omega})\rfloor$, and we have

$$\|V_{N_\epsilon+1} - V_{N_\epsilon}\|_\omega \leq (\eta\gamma)^{N_\epsilon}\|TV_0 - V_0\|_\omega < \epsilon.$$

Combining the two cases above, choose $N_\epsilon = \left(1 + \lfloor log_{\eta\gamma}(\frac{\epsilon}{\|TV_0 - V_0\|_\omega})\rfloor\right)\mathbb{1}_{TV_0 \neq V_0}$ and we have $\|V_{N_\epsilon+1} - V_{N_\epsilon}\|_\omega < \epsilon$, which implies that Algorithm 1 can converge within N_ϵ iterations.

Moreover, since V^* is the unique solution of the Shapley equation, we have

$$\|V_n - V^*\|_\omega \leq \|V_{n+1} - V^*\|_\omega + \|V_n - V_{n+1}\|_\omega \leq \eta\gamma\|V_n - V^*\|_\omega + \|V_n - V_{n+1}\|_\omega,$$

thus,

$$\|V_n - V^*\|_\omega \leq \frac{\|V_n - V_{n+1}\|_\omega}{1 - \eta\gamma},$$

taking $n = N_\epsilon$, and we have

$$\|V_{N_\epsilon} - V^*\|_\omega < \frac{\epsilon}{1 - \eta\gamma} = \varepsilon,$$

which implies that $(\varphi_1^\varepsilon, \varphi_2^\varepsilon)$ is ε-optimal by Definition 8. □

Therefore, with Algorithm 1, we can iteratively approach to the value function and an optimal pair of stationary strategies of our SMG problem through recursively solving a matrix game at each state x. Theorem 2 guarantees the convergence of Algorithm 1. We can implement Algorithm 1 to solve practical problems, as illustrated in the next section.

6 Numerical experiment

In this section, we conduct numerical examples to illustrate our main results derived in Sections 3-5. First, we give an example to demonstrate that Assumptions 1-4 ensuring the existence of the value function and an optimal pair of stationary strategies of SMGs are easy to verify in practice.

Example 1 Consider a system with a model of SMG which is defined as follows:

The state space $X := (-\infty, \infty)$ and the action spaces $A := [\underline{a}, \bar{a}]$, $B := [\underline{b}, \bar{b}]$ with admissible action sets $A(x) := A$, $B(x) := B$ for each $x \in X$. The semi-Markov kernel is given by:

$$Q(t, y|x, a, b) = \Phi\left(\frac{1+t}{2+t}y\right)F(t), \quad \forall(t, y, x, a, b) \in [0, +\infty) \times X \times K,$$

where $\Phi(\cdot)$ and $F(\cdot)$ denote the cumulative distribution functions of a normal distribution with mean $\mu(x, a, b)$ and variance $\sigma^2(x, a, b)$ and an exponential distribution with parameter $\beta(x, a, b)$, respectively. The reward rate function is denoted by $r(x, a, b) := x^2 + a + b$. Moreover, the discount factor is defined as $\alpha(x, a, b) := e^{|x|+a+b}$.

Now, we verify that the conditions on the existence of an optimal pair of stationary strategies described in Assumptions 1-4 are satisfied in this example. To this end, we need the following hypothesis:

Assumption 5 *(a) The function $\beta(x,a,b)$ is continuous on K and has both a positive lower bound $\underline{\beta}$ and a positive upper bound $\bar{\beta}$; (b) $\mu^2(x,a,b)+\sigma^2(x,a,b) \le \frac{1}{4}x^2 + \frac{e^{\underline{a}+\underline{b}-1}}{8(\bar{\beta}+e^{\underline{a}+\underline{b}})}$, $\forall (x,a,b)\in K$.*

With this hypothesis, we directly have the following result.

Proposition 2 *Suppose that Assumption 5 holds, then Example 1 satisfies Assumptions 1-4, which means that the SMG has an optimal pair of stationary strategies.*

The proof of Proposition 2 is provided in Appendix B.

Next, we give another example about investment problem to demonstrate the numerical computation of Algorithm 1 to solve the value function and an optimal pair of stationary strategies of the game.

Example 2 Consider an investment problem with three states $\{1,2,3\}$, which denotes the benefit, medium and loss economy environments, respectively. At each state, the investor will buy some assets while the market-maker will sell. The interest rate depends on the economy environments as well as the number of assets that investor buys and market-maker sells. In state $i\in\{1,2\}$, the investor buys a certain amount of assets from $\{a_{i1},a_{i2}\}$ and the market-maker sells from $\{b_{i1},b_{i2}\}$, which leads to a reward rate $r(i,a,b)$ to the investor and $-r(i,a,b)$ to the market-maker, where $a\in\{a_{i1},a_{i2}\}, b\in\{b_{i1},b_{i2}\}$. Then the system moves to a new state j with probability $p(j|i,a,b)$ after staying at state i for a random time which follows exponential-distribution with parameter $\beta(i,a,b)$. In state 3, the investor buys a certain amount of assets from $\{a_{31},a_{32}\}$ and the market-maker sells from $\{b_{31},b_{32}\}$, which leads to a reward rate $r(3,a,b)$ to the investor and $-r(3,a,b)$ to the market-maker, where $a\in\{a_{31},a_{32}\}, b\in\{b_{31},b_{32}\}$. Then the system moves to a new state j with probability $p(j|3,a,b)$ after staying at state 3 for a random time uniformly distributed in $[0,\beta(3,a,b)]$ with parameter $\beta(3,a,b)>0$. For this system, the decision makers aim to find an optimal pair of stationary strategies.

We establish an SMG model to solve this investment problem. The state space is $X=\{1,2,3\}$, action spaces are $A(i)=\{a_{i1},a_{i2}\}$, $B(i)=\{b_{i1},b_{i2}\}$ for each $i\in X$ and the semi-Markov kernel Q is given by:

$$Q(t,j|i,a,b)=\begin{cases}(1-e^{-\beta(i,a,b)t})p(j|i,a,b) & \text{if} \quad i\in\{1,2\},\\ \frac{t}{\beta(i,a,b)}p(j|i,a,b) & \text{if} \quad i=3,\ 0\le t\le\beta(i,a,b),\\ p(j|i,a,b) & \text{otherwise,}\end{cases}$$

from which we can obtain

$$Q(dt,j|i,a,b)=\begin{cases}p(j|i,a,b)\beta(i,a,b)e^{-\beta(i,a,b)t}dt & \text{if} \quad i\in\{1,2\},\\ \frac{dt}{\beta(i,a,b)}p(j|i,a,b) & \text{if} \quad i=3,\ 0\le t\le\beta(i,a,b),\\ 0 & \text{otherwise,}\end{cases}$$

and

$$H(t|i,a,b)=\begin{cases}1-e^{-\beta(i,a,b)t} & \text{if} \quad i\in\{1,2\},\\ \frac{t}{\beta(i,a,b)} & \text{if} \quad i=3,\ 0\le t\le\beta(i,a,b),\\ 1 & \text{otherwise.}\end{cases}$$

Then by (2), we have

$$G(u,i,a,b)=\begin{cases}\frac{r(i,a,b)}{\alpha(i,a,b)+\beta(i,a,b)}+\frac{\beta(i,a,b)}{\alpha(i,a,b)+\beta(i,a,b)}\sum_{j=1}^{3}p(j|i,a,b)u(j) & \text{if} \quad i\in\{1,2\},\\ \frac{r(3,a,b)}{(\alpha(3,a,b))^2\beta(3,a,b)}\left[\alpha(3,a,b)\beta(3,a,b)-1+e^{-\alpha(3,a,b)\beta(3,a,b)}\right] & \\ +\frac{1-e^{-\alpha(3,a,b)\beta(3,a,b)}}{\alpha(3,a,b)\beta(3,a,b)}\sum_{j=1}^{3}p(j|3,a,b)u(j) & \text{if} \quad i=3.\end{cases}$$

To take numerical calculation for this example, we assume that the values of model parameters are shown in Table 1.

With this data setting, by taking $\alpha_0=0.25$, $\delta=0.1$, $\theta=0.023$, Assumptions 1-4 obviously hold since both the state and action spaces are finite. Thus, the existence of the value function and an optimal pair of stationary strategies of the SMG are ensured by Theorem 1. Next, we use Algorithm 1 to find the approximate value function and an ε-optimal pair of stationary strategies of the game. The detailed steps are listed as follows.

Step 1: Initialization.

Let $n=0$, and $V_0(\cdot)=1$; set a small threshold $\epsilon:=10^{-4}$, and we have $\varepsilon=\frac{\epsilon}{1-\eta\gamma}=0.33$.

Step 2: Iteration.

For $n\geq 0$, $(a,b)\in A(i)\times B(i)$, we have

$$u_n(i,a,b)=\frac{r(i,a,b)}{\alpha(i,a,b)+\beta(i,a,b)}+\frac{\beta(i,a,b)}{\alpha(i,a,b)+\beta(i,a,b)}\sum_{j=1}^{3}p(j|i,a,b)V_n(j),\quad i=1,2,$$

$$\begin{aligned}u_n(3,a,b)=&\frac{r(3,a,b)}{(\alpha(3,a,b))^2\beta(3,a,b)}\left[\alpha(3,a,b)\beta(3,a,b)-1+e^{-\alpha(3,a,b)\beta(3,a,b)}\right]\\&+\frac{1-e^{-\alpha(3,a,b)\beta(3,a,b)}}{\alpha(3,a,b)\beta(3,a,b)}\sum_{j=1}^{3}p(j|3,a,b)V_n(j).\end{aligned}$$

Then, for each state $i\in\{1,2,3\}$, we solve the linear program

$$\max_{f(i,a_{i1}),f(i,a_{i2}),v}\quad v$$
$$\text{s.t.}\quad\begin{cases}v\leq u_n\left(i,a_{i1},b_{ij}\right)f\left(i,a_{i1}\right)+u_n\left(i,a_{i2},b_{ij}\right)f\left(i,a_{i2}\right),\quad j=1,2\\ f\left(i,a_{i1}\right)+f\left(i,a_{i2}\right)=1\\ f\left(i,a_{i1}\right)\geq 0,\ f\left(i,a_{i2}\right)\geq 0,\end{cases}\tag{12}$$

with the solution denoted by $\pi_n^1(\cdot|i)$ where $\pi_n^1(a_{i1}|i)=f(i,a_{i1})$, $\pi_n^1(a_{i2}|i)=f(i,a_{i2})$.

Also we solve the dual program of (12)

$$\min_{g(i,b_{i1}),g(i,b_{i2}),z}\quad z$$
$$\text{s.t.}\quad\begin{cases}z\geq u_n\left(i,a_{ij},b_{i1}\right)g\left(i,b_{i1}\right)+u_n\left(i,a_{ij},b_{i2}\right)g\left(i,b_{i2}\right),\quad j=1,2\\ g\left(i,b_{i1}\right)+g\left(i,b_{i2}\right)=1\\ g\left(i,b_{i1}\right)\geq 0,\ g\left(i,b_{i2}\right)\geq 0,\end{cases}$$

with the solution denoted by $\pi_n^2(\cdot|i)$ where $\pi_n^2(b_{i1}|i)=g(i,b_{i1})$, $\pi_n^2(b_{i2}|i)=g(i,b_{i2})$. We set

$$V_{n+1}(i)=\sum_{a\in A(i),b\in B(i)}u_n(i,a,b)\pi_n^1(a|i)\pi_n^2(b|i).$$

Table 1 The values of model parameters

state	1				2				3			
action	(a_{11}, b_{11})	(a_{11}, b_{12})	(a_{12}, b_{11})	(a_{12}, b_{12})	(a_{21}, b_{21})	(a_{21}, b_{22})	(a_{22}, b_{21})	(a_{22}, b_{22})	(a_{31}, b_{31})	(a_{31}, b_{32})	(a_{32}, b_{31})	(a_{32}, b_{32})
$\alpha(x, a, b)$	0.98	0.96	0.92	0.9	0.78	0.76	0.73	0.7	0.86	0.84	0.89	0.82
$r(x, a, b)$	40	24	18	33	12	8	10	17	3	5	2	6
$\beta(x, a, b)$	20	30	11	13	7	8	6.5	4	0.34	0.44	0.55	0.15
$p(1\|x, a, b)$	0	0	0	0	0.46	0.48	0.39	0.3	0.45	0.24	0.43	0.4
$p(2\|x, a, b)$	0.5	0.43	0.32	0.62	0	0	0	0	0.55	0.76	0.57	0.6
$p(3\|x, a, b)$	0.5	0.57	0.68	0.38	0.54	0.52	0.61	0.7	0	0	0	0

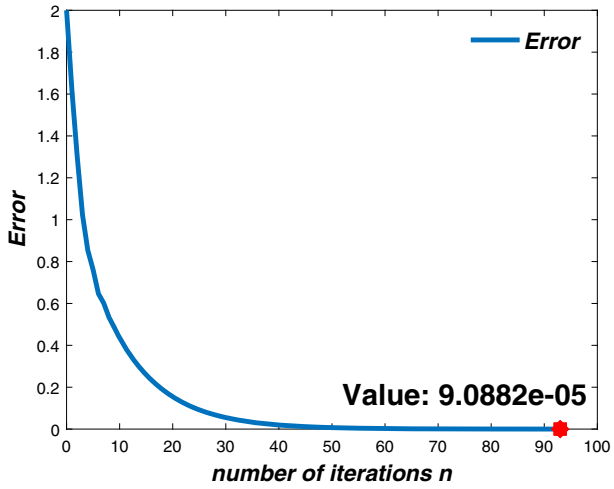


Fig. 1 The error of two successive iterations

Step 3: Termination judgement.

If $\max_{i=1,2,3} |V_{n+1}(i) - V_n(i)| < \epsilon$, then the iteration stops, V_n is the ε-value function and (π_n^1, π_n^2) is an ε-optimal pair of stationary strategies of the SMG. Otherwise, set $n = n + 1$ and go to Step 2.

We use Matlab to implement the iteration algorithm for this example. It takes about 10 seconds to stop at the 93rd iteration. The curves of the error of two successive iterations, the value function, and the optimal pair of stationary strategies of players with respect to the iteration times are illustrated by Figs. 1, 2 and 3.

Based on the experimental results, we have the following observations:

1. When the state is benefit, the investor should take action a_{11} with probability 0.60217 and a_{12} with probability 0.39783, while the market-maker should take action b_{11} with probability 0.55737 and b_{12} with probability 0.44263;

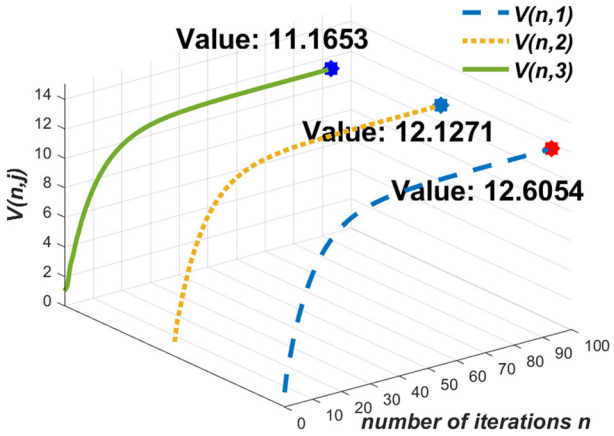


Fig. 2 The value function of the game

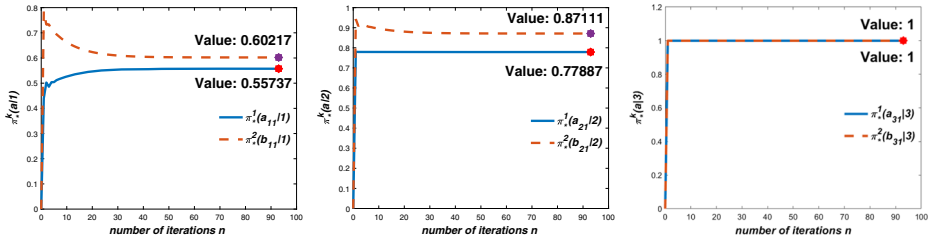


Fig. 3 The optimal pair of stationary strategies (π_*^1, π_*^2)

2. When the state is medium, the investor should take action a_{21} with probability 0.87111 and a_{22} with probability 0.12889, while the market-maker should take action b_{21} with probability 0.77887 and b_{22} with probability 0.22113;
3. When the state is loss, the investor should always take action a_{31} while the market-maker should always take action b_{31};
4. If both investor and market-maker use the optimal strategies, the investor will obtain a profit 12.6054 at benefit state, 12.1271 at medium state and 11.1653 at loss state, while the market-maker will lose the same amount, respectively.

Remark 5 In this example, we choose a uniformly distributed sojourn time at state 3 to show that arbitrary distributions are permitted for the sojourn time of semi-Markov processes. Other distributions can also be chosen for the sojourn time according to practical situations. Moreover, if all the sojourn times are exponentially distributed, the semi-Markov games degenerate into discrete-time Markov games.

7 Conclusion

In this paper, we concentrate on the two-person zero-sum SMG with expected discounted criterion in which the discount factors are state-action-dependent. We first construct the SMG model with a fairly general definition setting. Then we impose suitable conditions on the model parameters, under which we establish the Shapley equation whose unique solution is the value function and prove the existence of an optimal pair of stationary strategies of the game. While the state and action spaces are finite, a value iteration-type algorithm for approaching to the value function and an optimal pair of stationary strategies is developed. Finally, we apply our results to an investment problem, which demonstrates that our algorithm performs well.

One of the future research topics is to deal with the nonzero-sum case of this game model. We wish to find sufficient conditions under which we use the similar arguments to establish the Shapley equation and prove the existence of Nash equilibrium for such game. In practice, many decision problems are continuous in states or actions. It is of interest to further study the discretization technology of continuous SMGs in order to apply our value iteration algorithm. Moreover, considering the limitations of computing resources, the dynamic programming algorithm is difficult to implement in reality when the scale of the game becomes huge. Another future research topic is to develop data-driven learning algorithms to approximately solve the game problems, such as the combination with multi-agent reinforcement learning approaches.

Appendix A: Proofs of Lemmas 1-5

Proof **(Lemma 1)** For each fixed $(x, a, b) \in K$, integrating by parts and we have

$$\begin{aligned}
\int_0^\infty e^{-\alpha(x,a,b)t} H(dt|x,a,b) &= \alpha(x,a,b)\int_0^\infty e^{-\alpha(x,a,b)t} H(t|x,a,b)dt \\
&= \alpha(x,a,b)\Big[\int_0^\theta e^{-\alpha(x,a,b)t} H(t|x,a,b)dt + \int_\theta^\infty e^{-\alpha(x,a,b)t} H(t|x,a,b)dt\Big] \\
&\le \alpha(x,a,b)\Big[(1-\delta)\int_0^\theta e^{-\alpha(x,a,b)t}dt + \int_\theta^\infty e^{-\alpha(x,a,b)t}dt\Big] \\
&= 1-\delta\left(1-e^{-\alpha(x,a,b)\theta}\right) \\
&\leqslant 1-\delta+\delta e^{-\alpha_0\theta} < 1.
\end{aligned}$$

Let $\gamma = 1-\delta+\delta e^{-\alpha_0\theta}$, which yields (4). □

Proof **(Lemma 2)**

By Assumption 2 and formulation (6), for each given function $u \in B_\omega(X)$ and $(x, a, b) \in K$, we can easily get

$$|G(u,x,a,b)| \le \frac{M}{\alpha_0}\cdot\omega(x) + \eta\gamma\|u\|_\omega\cdot\omega(x).$$

The above inequality yields $\|G(u,\cdot,a,b)\|_\omega \le \frac{M}{\alpha_0} + \eta\gamma\|u\|_\omega$, which implies $G(u,x,a,b)$ is in $B_\omega(X)$, and so $Tu \in B_\omega(X)$.

On the one hand, by Assumption 4, it follows that $G(u,x,\cdot,b)$ is upper semi-continuous on $A(x)$, then for each fixed $\lambda \in \mathbb{B}(x)$, by the Fatou's theorem, the function

$$a \longmapsto \int_{B(x)} G(u,x,a,b)\lambda(db)$$

is also upper semi-continuous on $A(x)$. Moreover, since the probability measures on $\mathcal{B}(X)$ endowed with the topology of weak convergence, by Theorem 2.8.1 in Ash et al. (2000), the function $G(u,x,\cdot,\lambda)$ is upper semi-continuous on $\mathbb{A}(x)$. Similarly, $G(u,x,\mu,\cdot)$ is lower semi-continuous on $\mathbb{B}(x)$. Thus, by Theorem A.2.3 in Ash et al. (2000), the supremum and the infimum are indeed attained in (3), which means

$$Tu(x) = \max_{\mu\in\mathbb{A}(x)}\min_{\lambda\in\mathbb{B}(x)} G(u,x,\mu,\lambda).$$

Then, by the Fan's minimax Theorem (Fan 1953), we obtain (7).

On the other hand, since $\mathbb{A}(x)$ and $\mathbb{B}(x)$ are compact, by the well-known measurable selection theorem for minimax problems (Nowak 1984), there exists a pair of stationary strategies $(\varphi_1, \varphi_2) \in \Phi_1 \times \Phi_2$ that satisfies (8). □

Proof **(Lemma 3)**

First, it is easy to verify that the operator $T(\varphi_1, \varphi_2)$ is monotonically increasing. Let $u, v \in B_\omega(X)$, by the definition of ω-norm, $u(\cdot) \leq v(\cdot) + \|u - v\|_\omega \omega(\cdot)$, it follows that for each fixed $x \in X$, we have

$$\begin{aligned} T(\varphi_1, \varphi_2)u(x) &\leqslant T(\varphi_1, \varphi_2)(v + \omega\|u - v\|_\omega)(x) \\ &= T(\varphi_1, \varphi_2)v(x) \\ &\quad + \|u - v\|_\omega \int_{A(x)} \int_{B(x)} \Big[\int_0^\infty e^{-\alpha(x,a,b)t} \int_X \omega(y) Q(dt, dy|x, a, b) \Big] \varphi_1(da|x)\varphi_2(db|x) \\ &\leqslant T(\varphi_1, \varphi_2)v(x) + \eta\gamma\|u - v\|_\omega \omega(x), \end{aligned} \tag{13}$$

where the last inequality is followed by formulation (6). Furthermore, taking maximum of $\varphi_1 \in \Phi_1$ and minimum of $\varphi_2 \in \Phi_2$ on both sides of the inequality (13), we have

$$\max_{\varphi_1 \in \Phi_1} \min_{\varphi_2 \in \Phi_2} T(\varphi_1, \varphi_2)u(x) \leq \max_{\varphi_1 \in \Phi_1} \min_{\varphi_2 \in \Phi_2} T(\varphi_1, \varphi_2)v(x) + \eta\gamma\|u - v\|_\omega \omega(x),$$

i.e.,

$$Tu(x) \leq Tv(x) + \eta\gamma\|u - v\|_\omega \omega(x).$$

Similarly, interchanging u and v, we obtain

$$Tv(x) \leq Tu(x) + \eta\gamma\|v - u\|_\omega \omega(x).$$

Combining the two inequalities above, we have

$$|Tu(x) - Tv(x)| \leq \eta\gamma\|u - v\|_\omega \omega(x), \quad \forall x \in X,$$

i.e.,

$$\|Tu - Tv\|_\omega \leq \eta\gamma\|u - v\|_\omega,$$

which implies T is a contraction operator with modulus $\eta\gamma < 1$. Using the same arguments, we can prove that $T(\varphi_1, \varphi_2)$ is also a contraction operator with modulus $\eta\gamma < 1$. □

Proof **(Lemma 4)**

$$
\begin{aligned}
V(x,\pi^1,\pi^2) &= \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_0^\infty e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t),A(t),B(t))dt\Big] \\
&= \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_0^{T_1} e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t),A(t),B(t))dt\Big] \\
&\quad + \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_{T_1}^\infty e^{-\int_0^t \alpha(X(s),A(s),B(s))ds} r(X(t),A(t),B(t))dt\Big] \\
&= \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_0^\infty \mathbb{1}_{\{T_1>t\}} e^{-\alpha(x,A_0,B_0)t} r(x,A_0,B_0)dt\Big] \\
&\quad + \mathbb{E}_x^{\pi^1,\pi^2}\Big[\mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_{T_1}^\infty e^{-\alpha(x,A_0,B_0)T_1} e^{-\int_{T_1}^t \alpha(X(s),A(s),B(s))ds} r(X(t),A(t),B(t))dt|H_1\Big]\Big] \\
&= \int_{A(x)}\int_{B(x)}\Big[\int_0^\infty e^{-\alpha(x,a,b)t}\big[1-H(t|x,a,b)\big]r(x,a,b)dt\Big]\pi_0^1(da|x)\pi_0^2(db|x) \\
&\quad + \mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\alpha(x,A_0,B_0)T_1}\mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_{T_1}^\infty e^{-\int_{T_1}^t \alpha(X(s),A(s),B(s))ds} r(X(t),A(t),B(t))dt|H_1\Big]\Big] \\
&= \int_{A(x)}\int_{B(x)}\Big[\int_0^\infty e^{-\alpha(x,a,b)t}\big[1-H(t|x,a,b)\big]r(x,a,b)dt\Big]\pi_0^1(da|x)\pi_0^2(db|x) \\
&\quad + \mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\alpha(x,A_0,B_0)T_1} V(X_1,{}^{(1)}\pi^1,{}^{(1)}\pi^2)\Big] \\
&= \int_{A(x)}\int_{B(x)} r(x,a,b)\Big[\int_0^\infty e^{-\alpha(x,a,b)t}\big[1-H(t|x,a,b)\big]dt\Big]\pi_0^1(da|x)\pi_0^2(db|x) \\
&\quad + \int_{A(x)}\int_{B(x)}\Big[\int_0^\infty e^{-\alpha(x,a,b)t}\int_X V(y,{}^{(1)}\pi^1,{}^{(1)}\pi^2)Q(dt,dy|x,a,b)\Big]\pi_0^1(da|x)\pi_0^2(db|x) \\
&= \int_{A(x)}\int_{B(x)}\Big\{r(x,a,b)\Big[\int_0^\infty e^{-\alpha(x,a,b)t}\big[1-H(t|x,a,b)\big]dt\Big] + \\
&\quad \int_0^\infty e^{-\alpha(x,a,b)t}\Big[\int_X V(y,{}^{(1)}\pi^1,{}^{(1)}\pi^2)Q(dt,dy|x,a,b)\Big]\Big\}\pi_0^1(da|x)\pi_0^2(db|x),
\end{aligned}
$$

where the third and fourth equalities are ensured by the property of conditional expectation. The fifth equality follows from the strong Markovian property. Hence,

$$V(x,\pi^1,\pi^2) = T(\pi_0^{1\infty},\pi_0^{2\infty})V(x,{}^{(1)}\pi^1,{}^{(1)}\pi^2),$$

which is required. □

Proof **(Lemma 5)**

For $\forall n \geq 1$ and $x \in X$, we have

$$
\begin{aligned}
&\Big|\mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\int_0^{T_n}\alpha(X(s),A(s),B(s))ds}\omega(X_n)\Big]\Big| \\
=\ &\Big|\mathbb{E}_x^{\pi^1,\pi^2}\Big[\mathbb{E}_x^{\pi^1,\pi^2}\big[e^{-\int_0^{T_n}\alpha(X(s),A(s),B(s))ds}\omega(X_n)|H_{n-1},A_{n-1},B_{n-1}\big]\Big]\Big| \\
=\ &\Big|\mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\int_0^{T_{n-1}}\alpha(X(s),A(s),B(s))ds}\mathbb{E}_x^{\pi^1,\pi^2}\big[e^{-\int_{T_{n-1}}^{T_n}\alpha(X(s),A(s),B(s))ds}\omega(X_n)|H_{n-1},A_{n-1},B_{n-1}\big]\Big]\Big| \\
=\ &\Big|\mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\int_0^{T_{n-1}}\alpha(X(s),A(s),B(s))ds}\big[\int_0^\infty e^{-\alpha(X_{n-1},A_{n-1},B_{n-1})t}\int_X \omega(y)Q(dt,dy|X_{n-1},A_{n-1},B_{n-1})\big]\Big]\Big| \\
\leq\ &\eta\gamma\Big|\mathbb{E}_x^{\pi^1,\pi^2}\Big[e^{-\int_0^{T_{n-1}}\alpha(X(s),A(s),B(s))ds}\omega(X_{n-1})\Big]\Big|,
\end{aligned}
$$

where the first and second equalities are ensured by the property of conditional expectation. The last inequality follows from formulation (6). Through iteration we have

$$\begin{aligned}\left|\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-\int_0^{T_n}\alpha(X(s),A(s),B(s))ds}u(X_n)\right]\right| &\leq \|u\|_\omega\left|\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-\int_0^{T_n}\alpha(X(s),A(s),B(s))ds}\omega(X_n)\right]\right| \\ &\leq (\eta\gamma)^n\|u\|_\omega\omega(x),\end{aligned}$$

which yields Lemma 5. □

Appendix B: Proofs of Propositions 1-2

Proof **(Proposition 1)**

By the property of conditional expectation and Lemma 1, we have

$$\begin{aligned}\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_n}\right] &= \mathbb{E}_x^{\pi^1,\pi^2}\left[\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_n}|H_{n-1},A_{n-1},B_{n-1}\right]\right] \\ &= \mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_{n-1}}\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-(T_n-T_{n-1})}|H_{n-1},A_{n-1},B_{n-1}\right]\right] \\ &= \mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_{n-1}}\left[\int_0^\infty e^{-t}H(dt|X_{n-1},A_{n-1},B_{n-1})\right]\right] \\ &\leq (1-\delta+\delta e^{-\theta})\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_{n-1}}\right],\end{aligned}$$

where the last inequality follows directly from the proof of Lemma 1 by taking $\alpha_0 := 1$. Through iteration we have,

$$\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_n}\right] \leq (1-\delta+\delta e^{-\theta})^n.$$

For any given $t > 0$, by the Chebychev inequality, we obtain

$$\mathbb{P}_x^{\pi^1,\pi^2}(T_n \leq t) = \mathbb{P}_x^{\pi^1,\pi^2}(e^{-T_n} \geq e^{-t}) \leq e^t\mathbb{E}_x^{\pi^1,\pi^2}\left[e^{-T_n}\right] \leq e^t(1-\delta+\delta e^{-\theta})^n,$$

notice that $1-\delta+\delta e^{-\theta} < 1$, we have

$$\mathbb{P}_x^{\pi^1,\pi^2}(\lim_{n\to\infty} T_n \leq t) = \lim_{n\to\infty}\mathbb{P}_x^{\pi^1,\pi^2}(T_n \leq t) = 0,$$

since t is arbitrary, Proposition 1 holds. □

Proof **(Proposition 2)**

Using the marginal distribution formula, the corresponding distribution of sojourn time is given as follows,

$$H(t|x,a,b) = Q(t,+\infty|x,a,b) = 1-e^{-\beta(x,a,b)t}.$$

Easy to show that Assumption 2 holds by choosing $\alpha_0 = e^{\underline{a}+\underline{b}}$, $\omega(x) = x^2+1$ and $M = 3\max\{|\bar{a}|,|\underline{a}|,|\bar{b}|,|\underline{b}|,1\}$. Now, let $\theta = \frac{1}{\alpha_0}\ln\left(1+\frac{\alpha_0}{\bar{\beta}}\right)$ and $\delta = \left(1+\frac{\alpha_0}{\bar{\beta}}\right)^{-\bar{\beta}/\alpha_0}$, then for each $(x,a,b) \in K$,

$$H(\theta|x,a,b) = 1-e^{-\beta(x,a,b)\theta} \leq 1-e^{-\bar{\beta}\theta} = 1-\delta,$$

thus, Assumption 1 holds.

Next we verify Assumption 3. According to Lemma 1 and its proof, we can choose $\gamma = 1 - \delta + \delta e^{-\alpha_0 \theta}$ and further take $\eta = \frac{2}{1+\gamma}$, then

$$\begin{aligned}
\int_X \omega(y) Q(t, dy|x, a, b) &= \int_X (y^2+1)\Phi\left(\frac{1+t}{2+t}y\right) F(t) d\left(\frac{1+t}{2+t}y\right) \\
&= \left[\left(\frac{2+t}{1+t}\right)^2 (\mu^2(x,a,b) + \sigma^2(x,a,b)) + 1\right] F(t) \\
&\leq \left[4\left(\frac{1}{4}x^2 + \frac{e^{\underline{a}+\underline{b}-1}}{8(\bar{\beta} + e^{\underline{a}+\underline{b}})}\right) + 1\right] F(t) \\
&\leq \left[x^2 + \frac{\delta\alpha_0}{2(\bar{\beta}+\alpha_0)} + 1\right] F(t) \\
&< \left[x^2 + \frac{1-\gamma}{1+\gamma} + 1\right] F(t) \\
&\leq \eta\omega(x) H(t|x,a,b),
\end{aligned}$$

which implies that Assumption 3 holds. Finally, since the reward rate $r(x, a, b)$, discount factor $\alpha(x, a, b)$ and semi-Markov kernel $Q(t, y|, x, a, b)$ are continuous on K, Assumption 4 holds. Hence, the SMG of Example 1 has an optimal pair of stationary strategies. □

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (62073346, 11931018, U1811462), the Guangdong Basic and Applied Basic Research Foundation (2021A1515011984), and the Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2020B1212060032).

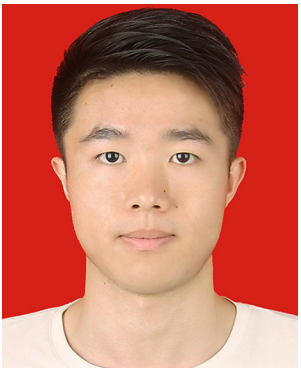
Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Ash RB, Robert B, Doleans-Dade CA, Catherine A (2000) Probability and measure theory. Academic Press
Barron EN (2013) Game Theory: An Introduction, 2nd Edn. Wiley, New York
Chen F, Guo X, Liao ZW (2021) Discounted semi-Markov games with incomplete information on one side. arXiv:210707499
Fan K (1953) Minimax theorems. Proc Natl Acad Sci U S A 39(1):42–47
Feinberg EA, Shwartz A (1994) Markov decision models with weighted discounted criteria. Math Oper Res 19(1):152–168
Filar J, Vrieze K (2012) Competitive Markov Decision Processes. Springer Science & Business Media, Berlin
González-Hernández J, López-martínez RR, Minjárez-Sosa JA (2009) Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion. Kybernetika 45(5):737–754
González-Sánchez D, Luque-Vásquez F, Minjárez-Sosa JA (2019) Zero-Sum Markov games with random State-Actions-Dependent discount factors: Existence of optimal strategies. Dyn Games Appl 9(1):103–121
Hernández-Lerma O, Lasserre JB (2012a) Discrete-time markov control processes: basic optimality criteria, vol 30. Springer Science & Business Media, Berlin
Hernández-Lerma O, Lasserre JB (2012b) Further topics on discrete-time markov control processes, vol 42. Springer Science & Business Media
Hoffman AJ, Karp RM (1966) On nonterminating stochastic games. Manag Sci 12(5):359–370
Huang Y, Guo X (2011) Finite horizon semi-Markov decision processes with application to maintenance systems. Eur J Oper Res 212(1):131–140
Huang YH, Guo XP (2010) Discounted semi-Markov decision processes with nonnegative costs. Acta Math Sinica (Chinese Series) 53:503–514

Kirman AP, Sobel MJ (1974) Dynamic oligopoly with inventories. Econometrica: Journal of the Econometric Society, 279–287
Lal AK, Sinha S (1992) Zero-sum two-person semi-Markov games. J Appl Probab 29(1):56–72
Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning
Luque-Vásquez F (2002a) Zero-sum semi-Markov game in Borel spaces with discounted payoff. Morfismos 6(1):15–29
Luque-Vásquez F (2002b) Zero-sum semi-Markov games in Borel spaces: discounted and average payoff. Bol Soc Mat Mexicana 8:227–241
Minjárez-Sosa JA (2015) Markov control models with unknown random state–action-dependent discount factors. Top 23(3):743–772
Minjárez-Sosa JA, Luque-Vásquez F (2008) Two person zero-sum semi-Markov games with unknown holding times distribution on one side: a discounted payoff criterion. Appl Math Optim 57(3):289–305
Mondal P, Sinha S, Neogy SK, Das AK (2016) On discounted AR-AT semi-Markov games and its complementarity formulations. Int J Game Theory 45(3):567–583
Mondal P, Neogy S, Gupta A, Ghorui D (2020) A policy improvement algorithm for solving a mixture class of perfect information and AR-AT semi-Markov games. Int Game Theory Rev 22(02):2040008
Nowak AS (1984) On zero-sum stochastic games with general state space I. Probab Math Stat 4(1):13–32
Pollatschek M, Avi-Itzhak B (1969) Algorithms for stochastic games with geometrical interpretation. Manag Sci 15(7):399–415
Raut LK (1990) Two-sided altruism lindahl equilibrium and pareto efficiency in overlapping generations models. Department of Economics, University of California
Ross SM (1970) Average cost semi-Markov decision processes. J Appl Probability 7(3):649–656
Schäl M (1975) Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 32(3):179–196
Shapley LS (1953) Stochastic games. Proc Natl Acad Sci 39(10):1095–1100
Tanaka K, Wakuta K (1976) On semi-Markov games. Science Reports of Niigata University Series A, Mathematics 13:55–64
Vega-Amaya Ó, Luque-Vásquez F, Castro-Enríquez M (2022) Zero-sum average cost semi-markov games with weakly continuous transition probabilities and a minimax semi-markov inventory problem. Acta Appl Math 177(1):1–27
Wei Q, Guo X (2011) Markov decision processes with state-dependent discount factors and unbounded rewards/costs. Oper Res Lett 39(5):369–374
Wu X, Zhang J (2016) Finite approximation of the first passage models for discrete-time Markov decision processes with varying discount factors. Discrete Event Dyn Syst 26(4):669–683
Ye L, Guo X (2012) Continuous-time Markov decision processes with state-dependent discount factors. Acta Applicandae Math 121(1):5–27



Zhihui Yu is a Ph.D. student with the School of Business, Sun Yat-Sen University, Guangzhou 510275, China. He received the Master degree in probability and statistics from the School of Mathematics, Sun Yat-Sen University, Guangzhou, China, in 2021. His research interests include the methodology research in stochastic learning and optimization, Markov decision processes, stochastic games, and reinforcement learning, etc.

Xianping Guo received the Ph.D. degree in probability and statistics from the Central South University, Changsha, China, in 1996. He has published many papers in international journals such as The Annals of Applied Probability, SIAM Journal on Optimization, SIAM Journal on Control and Optimization, IEEE Transactions on Automatic Control, and Automatica, and written (with O. Hernandez-Lerma) a book entitled Continuous-Time Markov Decision Processes (Springer, 2009). He serves/served as an editor of Advances in Applied Probability, Journal of Applied Probability, and of Science China Mathematics. He is currently with the Sun Yat-Sen University, Guangzhou, China, where he was appointed as a professor in 2002. His research interests include Markov decision processes and stochastic games.

Li Xia is a professor with the School of Business, Sun Yat-Sen University, Guangzhou 510275, China. He received the Bachelor and the Ph.D. degree in control theory both from Tsinghua University, Beijing, China, in 2002 and 2007, respectively. He worked at the IBM Research China (2007-2009) and at the King Abdullah University of Science and Technology (KAUST) Saudi Arabia (2009-2011). Then he returned to Tsinghua University as a lecturer in 2011 and was promoted as an associate professor in 2013. In 2019, he joined Sun Yat-Sen University as a full professor. He was a visiting scholar at Stanford University, the Hong Kong University of Science and Technology, etc. He serves/served as an associate editor for the IEEE Transactions on Automation Science and Engineering, Discrete Event Dynamic Systems, Energy Informatics, etc. His research interests include the methodology research in stochastic learning and optimization, Markov decision processes, reinforcement learning, queueing theory, and the application research in financial technology, energy systems, etc.