Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 6(10): 11516–11529.
DOI:10.3934/math.2021668
Received: 07 May 2021
Accepted: 04 August 2021
Published: 09 August 2021

Research article

Two-person zero-sum stochastic games with varying discount factors

Xiao Wu[1], **Qi Wang**[2] **and Yinying Kong**[3,*]

[1] School of Mathematics and Statistics, Zhaoqing University, Zhaoqing, 526061, China
[2] School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China
[3] School of Intelligence Financial Accounting Management, Guangdong University of Finance and Economics, Guangzhou, 510320, China

* **Correspondence:** Email: kongcoco@hotmail.com; Tel: +86-137-5178-1911.

Abstract: In this paper, two-person zero-sum Markov games with Borel state space and action space, unbounded reward function and state-dependent discount factors are studied. The optimal criterion is expected discount criterion. Firstly, sufficient conditions for the existence of optimal policies are given for the two-person zero-sum Markov games with varying discount factors. Then, the existence of optimal policies is proved by Banach fixed point theorem. Finally, we give an example for reservoir operations to illustrate the existence results.

Keywords: two-person zero-sum stochastic games; expected discount criterion; varying discount factors
Mathematics Subject Classification: 91A15, 60J05

1. Introduction

Since 1950s, people have begun to study discrete-time zero-sum stochastic games. The first work on this research topic is completed by Shapley in [1], in which Shapley considers the discount criteria of two-person zero-sum stochastic games with finite state space and action space. The results show that both of players have optimal policies. Then, Maitra extends the discount game to the case that the state space and the action space are uncountable in [2] and proves the existence of the optimal policies through the fixed point theorem. For the research on the relationship between the action space and the current state, we can refer to Parthasarathy [3]. In [2, 3], it is required that the state space be compact, which is relaxed in [4, 5]. The above work is carried out under the assumption that the state space is the Borel space, and Nowak [6] makes an in-depth study on the more common types of measurable policies in the context of stochastic games. Most of the articles mentioned above involve that the reward function is bounded or has only upper bound or only lower bound, which seems to be the first problem we consider in stochastic games. That is , when the state space is Borel space, the reward function can have neither upper bound nor lower bound, but it is limited by the drift condition. Interested readers can also refer to [7–10]. Such as, Guo et al. [8] studied the two-person zero-sum game of continuous time Markovian jump process under the discount criterion, in which the reward function is unbounded. By adding conditions to the initial data of the game, they guaranteed the existence of the solution of Shapley equation and obtained the existence of the optimal policies. Minjárez-Sosa and Luque-Vásquez [9] study the discount compensation criteria for zero-sum semi-Markov games with unbounded reward function.

The existing work on stochastic games can be roughly divided into the following four categories.

(1) Discrete-time Markov game: the state process of the game is a discrete-time Markov chain, the dwell time of adjacent states is a fixed constant, and the decision time is the time of state transition (i.e. equidistant fixed discrete time point), see [6, 7];

(2) Continuous-time Markov game: the state process of the game is a continuous time Markov chain, the trajectory of the system state is constant step by step, the state dwell time follows exponential distribution, and the decision time is any time point, see [8, 11];

(3) Semi-Markov game: the state process of the game is a semi Markov process, the state trajectory of the system is constant step by step, the state dwell time can obey any probability distribution, and the decision time is the random time point of state transition, see [9];

(4) Stochastic differential game: the state process of the game can be described by a stochastic differential equation.

As is well known, the discounted criterion of stochastic games with a constant discount factor have been widely studied as an important class of stochastic control problems. However, in financial systems, the discount is understood as $1/(1+\rho)$, where the interest rate ρ is usually not a constant, but depends on the states or actions of a underlying system, which is random in nature. Also, in investment, volatility is usually introduced into the stochastic discount factor as a state variable. Due to this interesting observation in practice, it is understandable to consider the varying discount factors, see, for example, Schäll [12], González-Hernández et al. [13, 14], Zhang [15] and their references therein.

Therefore, in this paper, we mainly discuss the discrete-time Markov games, with Borel state space and action space, unbounded reward function and state-dependent discount factors, give the expected discount criteria for two-person zero-sum Markov games. We construct two-person zero-sum Markov games model with varying discount factors, and give the expected discount criteria of the model. As far as we know, different from the existing literature, we consider the fact that the discount factor depends on the state of the system, which is more in line with the real world. This is also a generalization of the study of variable discount factor in MDPs model [16].

This paper contains two main contributions:

(a) We study the discrete-time two-person zero-sum stochastic games with Borel state space, and obtain the existence of optimal value and optimal policy pairs under suitable conditions, which provides a solid theoretical basis for us to calculate them.

(b) The discount factors are state-dependent, i.e., the discount factors $\alpha(x)$ is a state-dependent measurable function from the state space to $[0, 1)$, which is a generalization of the case of a constant discount factor.

This paper is organized as follows. In Section 2, we introduce the two-person zero-sum stochastic game model and its expected discount criterion. In Section 3, under suitable conditions, we prove the main result on the existence of optimal policy pairs. Finally, we give an example for reservoir operations to illustrate the existence results in Section 4.

2. Discrete-time stochastic game model and expected discount criterion

Before introducing the model of this paper, we first introduce the symbols throughout this paper. Given a Borel space X, that is, a subset of Borel of a complete separable metric space, we use $\mathscr{B}(x)$ to denote its Borel σ- algebra, and $P(X)$ to denote a probability measure on X and give it a weakly convergent topology. For any measurable function $\omega : X \to [1, \infty)$, we call the function u defined on X ω-bounded, If its ω-norm is finite, where ω-norm is defined as

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

Such a function ω can be considered as a weight function. In addition, for convenience, let $B_\omega(X)$ be a Banach space composed of all ω-bounded measurable functions defined on X.

Two-person zero-sum stochastic game model can be represented by the following model:

$$\{X, A, B, (A(x), B(x), x \in X), q(\cdot|x, a, b), \alpha(x), r(x, a, b)\}, \tag{2.1}$$

where,

- X is a state space, which is Borel space, and its Borel σ-algebra is $\mathscr{B}(X)$.
- A and B represent the action spaces of player 1 and player 2 respectively, which are Borel spaces, and their Borel σ- algebras are $\mathscr{B}(A)$ and $\mathscr{B}(B)$ respectively.
- $A(x)$ and $B(x)$ are the Borel subsets of A and B respectively, representing the allowed action sets of player 1 and player 2 in the state $x \in X$. Let

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

  be a measurable Borel subset of $X \times A \times B$.
- The transition probability $q(\cdot|x, a, b)$ is the random kernel on X given K, that is to say, for any $D \in \mathscr{B}(X)$, $q(D|x, a, b)$ is the Borel function defined on X, and then for any $x \in X, a \in A(x)$ and $b \in B(x)$, $q(\cdot|x, a, b)$ is the probability measure on K.
- The discount factor, $\alpha(x)$ is a state-dependent measurable function from X to $[0, 1)$.
- $r(x, a, b)$ is a real valued measurable function defined on K, which is the reward that the player 1 gets (i.e., the player 2 pays) when the current state is x, the player 1 takes action as a, and the player 2 takes action as b.

Next, we will give the definition of related policy classes.

Definition 1. *The random kernel sequence* $\pi^1 := (\pi^1_t, t = 0, 1, 2, \ldots)$ *is called a randomized Markov policy of player 1, if*

$$\pi^1_t(A(x)|x) = 1, \ \ \forall\ x \in X.$$

Let Π_1 be a collection of all randomized Markov policies of player 1.

Definition 2. *If there is a probability measure $\pi^1(\cdot|x) \in P(A(x))$ such that*

$$\pi_t^1(\cdot|x) = \pi^1(\cdot|x) \quad \forall\ x \in X\ ,\ t \geq 0,$$

Then, we call that the randomized Markov policy $\pi^1 := (\pi_t^1, t \geq 0)$ is stationary.

We denote all the stationary policies of player 1 as Π_1^s.

By replacing $A(x)$ in the above two definitions with $B(x)$, we can similarly define the randomized Markov policy class Π_2 and the stationary policy class Π_2^s of player 2.

For any initial state $x \in X$ and $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, by the famous Tulcea theorem (see [17]), there exists a unique probability space $(\Omega, \mathcal{F}, \mathbb{P}_x^{\pi^1,\pi^2})$ and the random process $\{(x_t, a_t, b_t), t \geq 0\}$ such that, for each $D \in \mathcal{B}(X)$ and $t \geq 0$, it holds that

$$\mathbb{P}_x^{\pi^1,\pi^2}(x_{t+1} \in D|h_t, a_t, b_t) = q(D|x_t, a_t, b_t),\ \forall h_t = (x, a_0, b_0, \cdots, x_{t-1}, a_{t-1}, b_{t-1}, x_t) \in H_t,$$

where, $H_0 := X$ and $H_t := K^t \times X = K \times H_{t-1}$, x_t, a_t and b_t represent the state variables and action variables of player 1 and player 2 at time t, respectively. Moreover, The expectation operator of $\mathbb{P}_x^{\pi^1,\pi^2}$ is given as $\mathbb{E}_x^{\pi^1,\pi^2}$.

Remark 1. *As mentioned above, each player selects actions independently, then, for any policy pair $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$ and the initial state $x \in X$, the action processes a_t and b_t are conditionally independent, that is,*

$$\mathbb{P}_x^{\pi^1,\pi^2}(a_t \in C, b_t \in E|h_t) = \pi_t^1(C|h_t)\pi_t^2(E|h_t), \forall h_t \in H_t, C \in \mathcal{B}(A), E \in \mathcal{B}(B).$$

Now, we give the expected discount compensation criteria for two-person zero-sum stochastic games.

Definition 3. *for any $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$, $x \in X$, and the discount factor $\alpha(x) \in [0, 1)$, the expected discount criteria of player 1 and player 2 are defined as follows:*

$$V_\alpha(x, \pi^1, \pi^2) := \mathbb{E}_x^{\pi^1,\pi^2}\Big[r(x_0, a_0, b_0) + \sum_{t=1}^{\infty}\prod_{s=0}^{t-1}\alpha(x_s)r(x_t, a_t, b_t)\Big]. \tag{2.2}$$

In addition, we call the functions defined on X as follow

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

as the lower value and the upper value of the discount compensation game, respectively. It is clear that $L(x) \leq U(x)$ for all $x \in X$. And, if $L(x) = U(x)$ for all $x \in X$, then we call it as the optimal value of a stochastic game, denoted as $V_\alpha^*(x)$.

Definition 4. *If the stochastic game have the optimal value $V_\alpha^*(x)$, then we call the policy π_1^* as the optimal policy of player 1, if*

$$\inf_{\pi^2\in\Pi_2} V_\alpha(x, \pi_1^*, \pi^2) = V_\alpha^*(x), \quad \forall x \in X.$$

Similarly, we call the policy π_2^* as the optimal policy of player 2, if

$$\sup_{\pi^1\in\Pi_1} V_\alpha(x, \pi^1, \pi_2^*) = V_\alpha^*(x), \quad \forall x \in X.$$

If π_k^* are the optimal policies of player k ($k = 1, 2$), then we call (π_1^*, π_2^*) as the optimal policy pair.

3. The existence of optimal policy pairs

In this section, inspired by [18], we will give sufficient conditions for the existence of optimal policy pairs in two-person zero-sum stochastic games. Some necessary symbols are given as follow.

for any $x \in X$, $h : K \to \mathbb{R}$, and the probability measures $\phi \in P(A(x))$ and $\psi \in P(B(x))$, let

$$h(x, \phi, \psi) := \int_{B(x)} \int_{A(x)} h(x, a, b)\phi(\mathrm{d}a)\psi(\mathrm{d}b),$$

and then,

$$r(x, \phi, \psi) := \int_{B(x)} \int_{A(x)} r(x, a, b)\phi(\mathrm{d}a)\psi(\mathrm{d}b), \tag{3.1}$$

$$q(D|x, \phi, \psi) := \int_{B(x)} \int_{A(x)} q(D|x, a, b)\phi(\mathrm{d}a)\psi(\mathrm{d}b), \quad \forall D \in \mathcal{B}(X). \tag{3.2}$$

Similarly, for any $\pi^1 = (\pi_t^1(\cdot|x)) \in \Pi_1$, $\pi^2 = (\pi_t^2(\cdot|x)) \in \Pi_2$, we write

$$r(x, \pi_t^1, \pi_t^2) := \int_{B(x)} \int_{A(x)} r(x, a, b)\pi_t^1(\mathrm{d}a|x)\pi_t^2(\mathrm{d}b|x), \tag{3.3}$$

and

$$q(D|x, \pi_t^1, \pi_t^2) := \int_{B(x)} \int_{A(x)} q(D|x, a, b)\pi_t^1(\mathrm{d}a|x)\pi_t^2(\mathrm{d}b|x), \quad \forall D \in \mathcal{B}(X). \tag{3.4}$$

In particular, if π^1 and π^2 are stationary, then (3.3) and (3.4) can be write as $r(x, \pi^1, \pi^2)$ and $q(D|x, \pi^1, \pi^2)$, respectively.

Because the reward function $r(x, a, b)$ may be unbounded, in order to guarantee the finiteness of $V_\alpha(x, \pi^1, \pi^2)$, we need to give the following assumptions, which is the so-called "expected growth" condition (see Assumption 3.1 in [18]).

Assumption 1. *(a) There exists a constant $\alpha \in (0, 1)$ such that $\sup_{x\in X} \alpha(x) \leq \alpha$.*

(b) There exist nonnegative constants β and γ (with $\gamma\alpha < 1$), and a weight function $\omega(x)$, such that for all $(x, a, b) \in K$, we have

$$|r(x, a, b)| \leq \beta\omega(x), \tag{3.5}$$

and

$$\int_X \omega(y)q(\mathrm{d}y|x, a, b) \leq \gamma\omega(x). \tag{3.6}$$

Remark 2. *By Assumption 1(a), it holds obviously for the case that the discount factor is constant. Assumption 1(b) shows that the reward function $r(x, a, b)$ can have neither upper bound nor lower bound. In addition, the purpose of (3.6) is to ensure that the Shapley operator (i.e. (3.8)) is a contractive operator.*

In order to guarantee the existence of optimal policies, we also need to give the following famous continuous-compact conditions, which can be referred to [1,8,11,17,19] and their references.

Assumption 2. *(a) For each $x \in X$, $A(x)$ and $B(x)$ are compact;*

(b) For each $x \in X$, $r(x,a,b)$ is continuous on $(a,b) \in A(x) \times B(x)$;

(c) For each $x \in X$ and any bounded measurable function u(x) on X, $\int_{y\in X} u(y)q(y|x,a,b)$ is continuous on $(a,b) \in A(x) \times B(x)$, and so is the weight function $\omega(x)$.

Now, for any $u \in B_\omega(X)$ and $(x,a,b) \in K$, we define

$$H(u,x,a,b) := r(x,a,b) + \alpha(x)\int_X u(y)q(\mathrm{d}y|x,a,b), \tag{3.7}$$

$$T_\alpha u(x) := \sup_{\phi\in P(A(x))} \inf_{\psi\in P(B(x))} H(u,x,\phi,\psi), \quad x \in X, \tag{3.8}$$

and

$$T_{\pi^1,\pi^2}u(x) := H(u,x,\pi^1,\pi^2), \quad \forall(\pi^1,\pi^2) \in \Pi_1^s \times \Pi_2^s. \tag{3.9}$$

Before stating our main conclusion, we need to give some lemmas to prove our main conclusion.

Lemma 1. *(a) Suppose that Assumptions 1 and 2 hold, then for any $u \in B_\omega(X)$, we have $T_\alpha u \in B_\omega(X)$, $T_{\pi^1,\pi^2}u \in B_\omega(X)$, and*

$$T_\alpha u(x) := \max_{\phi\in P(A(x))} \min_{\psi\in P(B(x))} H(u,x,\phi,\psi), \quad x \in X. \tag{3.10}$$

In addition, there exists a policy pair $(\pi_1^*,\pi_2^*) \in \Pi_1^s \times \Pi_2^s$ such that

$$\begin{aligned} T_\alpha u(x) = H(u,x,\pi_1^*,\pi_2^*) &= \max_{\pi^1\in P(A(x))} H(u,x,\pi^1,\pi_2^*) \\ &= \min_{\pi^2\in P(B(x))} H(u,x,\pi_1^*,\pi^2). \end{aligned} \tag{3.11}$$

(b) Both T_α and T_{π^1,π^2} are contraction operators.

Proof. (a) By Assumption 1, for any $u \in B_\omega(X), (x,a,b) \in K$, we have

$$\begin{aligned} |H(u,x,a,b)| &\le \beta\omega(x) + \alpha\|u\|_\omega \int_X \omega(y)q(\mathrm{d}y|x,a,b) \\ &\le \beta\omega(x) + \alpha\gamma\|u\|_\omega\omega(x). \end{aligned}$$

Since $T_\alpha u$ and $T_{\pi^1,\pi^2}u$ are measurable, then we obtain that $T_\alpha u \in B_\omega(X)$ and $T_{\pi^1,\pi^2}u \in B_\omega(X)$. On the other hand, by Assumption 2, $H(u,x,a,b)$ is continuous on $(a,b) \in A(x) \times B(x)$, which yields that $H(u,x,\phi,\psi)$ is continuous on $(\phi,\psi) \in P(A(x)) \times P(B(x))$. Moreover, by Fan's Minimax Theorem in [20], we have (3.10) holds. Furthermore, by Measurable Selection Theorem in [21], there exist $\pi_1^* \in \Pi_1^s$ and $\pi_2^* \in \Pi_2^s$ such that (3.11) holds.

(b) First, we show that T_{π^1,π^2} is a contraction operator. By Assumption 1, for all $x \in X$ and $u, v \in B_\omega(X)$, we can get

$$\begin{aligned} |T_{\pi^1,\pi^2}u(x) - T_{\pi^1,\pi^2}v(x)| &= |H(u, x, \pi^1, \pi^2) - H(v, x, \pi^1, \pi^2)| \\ &= |\alpha(x) \int_X (u(y) - v(y))q(\mathrm{d}y|x, \pi^1, \pi^2)| \\ &\le |\alpha(x)| \int_X \|u - v\|_\omega \cdot \omega(y) q(\mathrm{d}y|x, \pi^1, \pi^2) \\ &\le \gamma\alpha \|u - v\|_\omega \cdot \omega(x), \end{aligned} \tag{3.12}$$

which yields that

$$\|T_{\pi^1,\pi^2}u - T_{\pi^1,\pi^2}v\|_\omega \le \gamma\alpha\|u - v\|_\omega.$$

Thus, T_{π^1,π^2} is a contraction operator.

On the other hand, by (3.12), we have

$$T_{\pi^1,\pi^2}u(x) \le T_{\pi^1,\pi^2}v(x) + \gamma\alpha\|u - v\|_\omega \cdot \omega(x),$$

which yields that

$$\begin{aligned} &\max_{\pi^1 \in P(A(x))} \min_{\pi^2 \in P(B(x))} T_{\pi^1,\pi^2}u(x) \\ \le &\max_{\pi^1 \in P(A(x))} \min_{\pi^2 \in P(B(x))} T_{\pi^1,\pi^2}v(x) + \gamma\alpha\|u - v\|_\omega \cdot \omega(x), \end{aligned}$$

that is,

$$T_\alpha u(x) \le T_\alpha v(x) + \gamma\alpha\|u - v\|_\omega \cdot \omega(x).$$

Similarly, we can also obtain that

$$T_\alpha v(x) \le T_\alpha u(x) + \gamma\alpha\|v - u\|_\omega \cdot \omega(x),$$

and then,

$$|T_\alpha u(x) - T_\alpha v(x)| \le \gamma\alpha\|u - v\|_\omega \cdot \omega(x).$$

Furthermore, we can get

$$\|T_\alpha u - T_\alpha v\|_\omega \le \gamma\alpha\|u - v\|_\omega,$$

that is, T_α is also a contraction operator. □

Remark 3. *Since T_α and T_{π^1,π^2} are contraction operators, by Banach fixed point theorem, there exist unique functions v^* and $v^*_{\pi^1,\pi^2}$ in $B_\omega(X)$, such that, for all $x \in X$, we have $T_\alpha v^*(x) = v^*(x)$ and $T_{\pi^1,\pi^2}v^*_{\pi^1,\pi^2}(x) = v^*_{\pi^1,\pi^2}(x)$.*

Lemma 2. *For any $(\pi^1, \pi^2) \in \Pi_1^s \times \Pi_2^s$, the expected discount criteria $V_\alpha(\cdot, \pi^1, \pi^2)$ is the only fixed point of the operator T_{π^1,π^2} on $B_\omega(X)$.*

Proof. It needs only prove that $V_\alpha(x,\pi^1,\pi^2) = T_{\pi^1,\pi^2}V_\alpha(x,\pi^1,\pi^2)$ for all $x \in X$. In fact,

$$\begin{aligned}
V_\alpha(x,\pi^1,\pi^2) &= \mathbb{E}_x^{\pi^1,\pi^2}\Big[r(x_0,a_0,b_0) + \sum_{t=1}^{\infty}\prod_{s=0}^{t-1}\alpha(x_s)r(x_t,a_t,b_t)\Big] \\
&= r(x,\pi^1,\pi^2) + \mathbb{E}_x^{\pi^1,\pi^2}\Big[\sum_{t=1}^{\infty}\prod_{s=0}^{t-1}\alpha(x_s)r(x_t,a_t,b_t)\Big] \\
&= r(x,\pi^1,\pi^2) + \mathbb{E}_x^{\pi^1,\pi^2}\Big[\alpha(x)\mathbb{E}_x^{\pi^1,\pi^2}[r(x_1,a_1,b_1) + \sum_{t=2}^{\infty}\prod_{s=1}^{t-1}\alpha(x_s)r(x_t,a_t,b_t)|h_1]\Big] \\
&= r(x,\pi^1,\pi^2) + \mathbb{E}_x^{\pi^1,\pi^2}\Big[\alpha(x)V_\alpha(x_1,\pi^1,\pi^2)\Big] \\
&= T_{\pi^1,\pi^2}V_\alpha(x,\pi^1,\pi^2),
\end{aligned}$$

where the second and third equalities are due to the properties of conditional expectation, and the last equality is derived from the strong Markov property. □

Lemma 3. *Suppose that Assumptions 1 and 2 hold, then for π^1 and π^2, $x \in X$ and $t = 0,1,\ldots$, we have*

(a) $\mathbb{E}_x^{\pi^1,\pi^2}[\omega(x_t)] \le \gamma^t\omega(x)$,

(b) $\lim\limits_{t\to\infty}\mathbb{E}_x^{\pi^1,\pi^2}[\prod_{s=0}^{t-1}\alpha(x_s)u(x_t)] = 0$, $\forall\ u \in B_\omega(X)$.

Proof. (a) When $t = 0$, it holds obviously. Then, when $t = 1$, by (3.6), we can obtain

$$\begin{aligned}
\mathbb{E}_x^{\pi^1,\pi^2}[\omega(x_1)] =&\mathbb{E}_x^{\pi^1,\pi^2}[\mathbb{E}_x^{\pi^1,\pi^2}(\omega(x_1))|x,a_0,b_0] \\
&= \mathbb{E}_x^{\pi^1,\pi^2}[\int_X \omega(y)q(\mathrm{d}y|x,a_0,b_0)] \\
&\le \mathbb{E}_x^{\pi^1,\pi^2}[\gamma\omega(x)] = \gamma\omega(x).
\end{aligned}$$

When $t > 1$, we have

$$\begin{aligned}
\mathbb{E}_x^{\pi^1,\pi^2}[\omega(x_t)] &= \mathbb{E}_x^{\pi^1,\pi^2}\Big[\mathbb{E}_x^{\pi^1,\pi^2}(\omega(x_t))|h_{t-1},a_{t-1},b_{t-1}\Big] \\
&= \mathbb{E}_x^{\pi^1,\pi^2}\Big[\int_X \omega(y)q(\mathrm{d}y|x_{t-1},a_{t-1},b_{t-1})\Big] \\
&\le \gamma\mathbb{E}_x^{\pi^1,\pi^2}[\omega(x_{t-1})].
\end{aligned}$$

By the mathematical induction method and the iteration, part (a) holds.

(b) By Assumption 1(a) and Lemma 3(a), we can get

$$\begin{aligned}
\Big|\mathbb{E}_x^{\pi^1,\pi^2}[\prod_{s=0}^{t-1}\alpha(x_s)u(x_t)]\Big| &\le \alpha^t\mathbb{E}_x^{\pi^1,\pi^2}|u(x_t)| \le \alpha^t\|u\|_\omega\mathbb{E}_x^{\pi^1,\pi^2}[\omega(x_t)] \\
&\le (\alpha\gamma)^t\|u\|_\omega\omega(x).
\end{aligned}$$

Note that $\alpha\gamma < 1$, and let $t \to \infty$ in the equality above, then part (b) holds. □

Theorem 1. *Suppose that Assumptions 1 and 2 hold, then*

(a) The optimal value V_α^* of two-person zero-sum stochastic game exists and satisfies the following equation

$$V_\alpha^*(x) = T_\alpha V_\alpha^*(x), \quad \forall x \in X. \tag{3.13}$$

In addition, V_α^* is the unique solution of the above equation (3.13) on $B_\omega(X)$.

(b) The stationary policy pair $(\pi_1^*, \pi_2^*) \in \Pi_1^s \times \Pi_2^s$ is optimal if and only if $V_\alpha(\cdot, \pi_1^*, \pi_2^*)$ is the solution of equation (3.13).

Proof. (a) By Lemma 2, we can suppose V_α is a fixed point of T_α on $B_\omega(X)$, and then

$$V_\alpha(x) = T_\alpha V_\alpha(x).$$

By Lemma 1, there exists a stationary policy pair $(\pi_1^*, \pi_2^*) \in \Pi_1^s \times \Pi_2^s$ such that

$$V_\alpha(x) = H(V_\alpha, x, \pi_1^*, \pi_2^*) = \max_{\pi^1 \in P(A(x))} H(V_\alpha, x, \pi^1, \pi_2^*) = \min_{\pi^2 \in P(B(x))} H(V_\alpha, x, \pi_1^*, \pi^2),$$

which shows that V_α is also a fixed point of $T_{\pi_1^*, \pi_2^*}$ on $B_\omega(X)$. Thus, by Lemma 2,

$$V_\alpha(x) = V_\alpha(x, \pi_1^*, \pi_2^*), \quad \forall\ x \in X.$$

Next, we will state that V_α is the optimal value of the two-person zero-sum stochastic game, and (π_1^*, π_2^*) is the optimal policy pair. To do this, we just need to prove, for any $(\pi^1, \pi^2) \in \Pi_1 \times \Pi_2$ and all $x \in X$, it holds that

$$V_\alpha(x, \pi_1^*, \pi^2) \geq V_\alpha(x, \pi_1^*, \pi_2^*) \geq V_\alpha(x, \pi^1, \pi_2^*). \tag{3.14}$$

Here we only prove the second inequality of (3.14) and the first inequality can be similarly derived. By (2.2), we have

$$V_\alpha(x, \pi^1, \pi_2^*) = \mathbb{E}_x^{\pi^1, \pi_2^*}\Big[r(x_0, a_0, b_0) + \sum_{t=1}^{\infty} \prod_{s=0}^{t-1} \alpha(x_s) r(x_t, a_t, b_t)\Big].$$

From the nature of conditional expectation, for $t \geq 1$ and $h_t \in H_t$, we can conclude that

$$\begin{aligned}
&\mathbb{E}_x^{\pi^1, \pi_2^*}\Big[\prod_{s=0}^{t} \alpha(x_s) V_\alpha(x_{t+1}, \pi_1^*, \pi_2^*) | h_t, a_t, b_t\Big] \\
&= \prod_{s=0}^{t} \alpha(x_s) \mathbb{E}_x^{\pi^1, \pi_2^*}\Big[V_\alpha(x_{t+1}, \pi_1^*, \pi_2^*) | h_t, a_t, b_t\Big] \\
&= \prod_{s=0}^{t} \alpha(x_s) \sum_{y \in X} V_\alpha(y, \pi_1^*, \pi_2^*) q(y | x_t, \pi_t^1(h_t), \pi_2^*(x_t)) \\
&= \prod_{s=0}^{t-1} \alpha(x_s) \Big\{\alpha(x_t) \sum_{y \in X} V_\alpha(y, \pi_1^*, \pi_2^*) q(y | x_t, \pi_t^1(h_t), \pi_2^*(x_t))
\end{aligned}$$

$$
\begin{aligned}
&+ r(x_t,\pi_t^1(h_t),\pi_2^*(x_t)) - r(x_t,\pi_t^1(h_t),\pi_2^*(x_t))\Big\} \\
&\le \prod_{s=0}^{t-1}\alpha(x_s)[V_\alpha(x_t,\pi_1^*,\pi_2^*) - r(x_t,\pi_t^1(h_t),\pi_2^*(x_t))].
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\prod_{s=0}^{t-1}\alpha(x_s)V_\alpha(x_t,\pi_1^*,\pi_2^*) - \mathbb{E}_x^{\pi^1,\pi_2^*}\Big[\prod_{s=0}^{t}\alpha(x_s)V_\alpha(x_{t+1},\pi_1^*,\pi_2^*)|h_t,a_t,b_t\Big] \\
&\ge \prod_{s=0}^{t-1}\alpha(x_s)r(x_t,\pi_t^1(h_t),\pi_2^*(x_t)).
\end{aligned}
$$

Choose $t=0$, and then

$$
V_\alpha(x_0,\pi_1^*,\pi_2^*) - \mathbb{E}_x^{\pi^1,\pi_2^*}[\alpha(x_0)V_\alpha(x_1,\pi_1^*,\pi_2^*)] \ge r(x_0,\pi_0^1(x_0),\pi_2^*(x_0)).
$$

Take expectations $\mathbb{E}_x^{\pi^1,\pi_2^*}$ and sum t from 0 to T, we have

$$
\begin{aligned}
&V_\alpha(x_t,\pi_1^*,\pi_2^*) - \mathbb{E}_x^{\pi^1,\pi_2^*}[\prod_{s=0}^{t}\alpha(x_s)V_\alpha(x_{T+1},\pi_1^*,\pi_2^*)] \\
&\ge \mathbb{E}_x^{\pi^1,\pi_2^*}[r(x_0,a_0,b_0) + \sum_{t=1}^{T}\prod_{s=0}^{T-1}\alpha(x_s)r(x_t,a_t,b_t)]
\end{aligned}
$$

Let $T\to\infty$, and by Lemma 3(b), we can obtain that $V_\alpha(x,\pi_1^*,\pi_2^*) \ge V_\alpha(x,\pi^1,\pi_2^*)$. Then, part (a) holds.

(b) Firstly, we prove the 'only if' part. Suppose that $(\pi_1^*,\pi_2^*) \in \Pi_1^s\times\Pi_2^s$ is a optimal policy pair, then for any $x\in X, \pi^1\in\Pi_1$ and $\pi^2\in\Pi_2$, we have

$$
V_\alpha(x,\pi_1^*,\pi^2) \ge V_\alpha(x,\pi_1^*,\pi_2^*) \ge V_\alpha(x,\pi^1,\pi_2^*). \tag{3.15}
$$

Now, fix $x\in X$, and for any $\psi\in P(B(x))$, we can define $\hat\pi = (\hat\pi_t)$ as follow: $\hat\pi_0=\psi$ and $\hat\pi_t=\pi_2^*, \forall t\ge 1$. Then, by the first inequality in (3.15), we can get

$$
\begin{aligned}
V_\alpha(x,\pi_1^*,\pi_2^*) \le& V_\alpha(x,\pi_1^*,\hat\pi) \\
&= \int_{B(x)}\int_{A(x)}[r(x,a,b) + \alpha(x)\int_X V_\alpha(y,\pi_1^*,\pi_2^*)q(\mathrm{d}y|x,a,b)]\pi_1^*(\mathrm{d}a)\psi(\mathrm{d}b),
\end{aligned}
$$

which yields that

$$
V_\alpha(x,\pi_1^*,\pi_2^*) \le H(V_\alpha,x,\pi_1^*,\psi).
$$

Thus, by Lemma 1(a), we obtain

$$
V_\alpha(x,\pi_1^*,\pi_2^*) \le T_\alpha V_\alpha(x,\pi_1^*,\pi_2^*).
$$

For the same reason, we have

$$
V_\alpha(x,\pi_1^*,\pi_2^*) \ge T_\alpha V_\alpha(x,\pi_1^*,\pi_2^*).
$$

Then, $V_\alpha(x,\pi_1^*,\pi_2^*) = T_\alpha V_\alpha(x,\pi_1^*,\pi_2^*)$.

On the other hand, the 'if' part holds from the proof of part (a). □

4. An example for reservoir operations

Example 1. *As is well known, the reservoir systems have multiple purposes such as water supply for land irrigation, industrial or domestic use, hydropower generation, flood control, etc., which in some cases may be in conflict. If there is a well-defined priority between the purposes, the conflicting situation can be modeled as a constrained control optimal problem (with respect the purpose with the highest priority and imposing constraints to hedge the systems against the unsatisfied demand for the others). However, in many cases such a priority is very difficult or even impossible to establish. The game modeling provides an alternative to overcome this disadvantage of the control model formulation.*

Here, we study a single reservoir system with infinite capacity and two purposes modeled as a zero-sum game, in the sense that the water used for one purpose can be considered as water lost for the other. In addition, it is natural to include the economic state into the model, so that the discount factor is automatically state-dependent. Therefore, we model the reservoir as a zero-sum game in the following way. The inflows happen at nonnegative random times T_t, $t = 0, 1, 2, \ldots$, *with* $T_0 := 0$. *Let* Z_t *be the inflow at time* T_t ($t = 1, 2, \ldots$) *and assume it is a nonnegative random variable. At each time* T_t *the decision maker observes the stored water volume* $y_t \in [0, \infty) =: \mathbb{R}_+$ *and the economic state* $i_t \in \{1, 2, \ldots, N\}$ *(here N is an arbitrarily fixed positive integer), and chooses the consumption rate* $a_t \in A := [0, \overline{a}]$ *for purpose 1 and the consumption rate* $b_t \in B := [\underline{b}, \overline{b}]$ *for purpose 2, where* $\overline{a}, \underline{b}$ *and* $\overline{b}$ *are fixed positive constants. These consumption rates remain fixed until the next inflow time* T_{t+1} *occurs whenever the water available at the beginning of the period has not been depleted. If this is the case, the total withdrawal during the period* $(T_t, T_{t+1}]$ *is* $(a_t + b_t)L_t$, *where* $L_t := T_{t+1} - T_t$. *When the storage process reaches the zero volume it continues there until a positive inflow arrives.*

The storage process $\{y_t\}$ *evolves on* $\mathbb{R}_+$ *according to the recursive equation*

$$y_{t+1} = \max\{[y_t - (a_t + b_t)L_t] + Z_t,\ 0\}, \quad t = 0, 1, 2, \ldots, \tag{4.1}$$

where y_0 *is the initial water volume. Moreover, the economic state process is a time-homogeneous discrete-time Markov chain in* $\{1, 2, \ldots, N\}$ *with the initial state* i_0 *and the one-step transition probability* p_{ij}. *Suppose that the stored water volume is independent of the economic state, and we take the state space to be* $X = [0, \infty) \times \{1, 2, \ldots, N\}$. *Below in this example the generic denotation* $x = (y, i) \in X$ *is in frequent use, and the system starts with the initial state* $x_0 = (y_0, i_0)$. *Obviously, for each* $((y, i), a, b) \in K$ *and* $(z, l) \in \mathbb{R}_+ \times \mathbb{R}_+$, *it holds that the function* $\max\{y - (a + b)l + z,\ 0\}$ *is continuous in* $((y, i), a, b)$ *for all* $(z, l) \in \mathbb{R}_+ \times \mathbb{R}_+$.

To obtain the properties in the other assumptions we impose the following conditions:

(i) The sequences $\{Z_t\}$ *and* $\{L_t\}$ *are independent and each one of them is formed by independent and identically distributed random variables. Let* $\rho_1(\cdot)$ *be the density of* $\{Z_t\}$ *and* $\rho_2(\cdot)$ *be the density of* $\{L_t\}$. *Thus, denoting* ρ^* *the joint density of* (L_t, Z_t), *we have* $\rho^*(\cdot, \cdot) = \rho_1(\cdot)\rho_2(\cdot)$.

(ii) $\{Z_t\}$ *and* $\{L_t\}$ *have continuous bounded densities* ρ_1 *and* ρ_2, *respectively.*

(iii) We also assume that the mean values of inflow and the interarrival times are finite and also that they satisfy the inequality $E(Z_t) < \underline{b}E(L_t)$.

At each stage t, player 1 receives a payoff $r(x_t, a_t, b_t)$ *from player 2, and the game jumps to a new state* x_{t+1} *according to the transition law determined by (4.1):*

$$q(D \times \{j\} \big| (y, i), a, b) := \text{Prob}(x_{t+1} \in D \times \{j\} \big| x_t = (y, i), a_t = a, b_t = b)$$

$$= p_{ij} \iint_{\mathbb{R}_+\times\mathbb{R}_+} \boldsymbol{1}_D\{\max\{y-(a+b)l+z,\ 0\}\}\rho^*(z,l)\mathrm{d}z\mathrm{d}l,\quad \forall D\times\{j\}\in\mathscr{B}(X),$$

where $\boldsymbol{1}_D\{\cdot\}$ denotes the indicator function of the set D. The goal of player 1 (player 2, resp.) is to maximize (minimize, resp.) his/her reward flow (cost flow, resp.) $r(x_0,a_0,b_0), r(x_1,a_1,b_1),\ldots$ over an infinite horizon. Suppose that the discount factor $\alpha(x)=\alpha(y,i):=\alpha(i)$ is a measurable function from X to $[0,1)$, which depends on the economic state i, and then a discounted expected reward (or cost) criterion is well-defined as in section 2.

Now, we define the function

$$R(s)=E\exp(s(z-\underline{b}l))=\iint_{\mathbb{R}_+\times\mathbb{R}_+}\exp(s(z-\underline{b}l))\rho^*(z,l)\mathrm{d}z\mathrm{d}l$$

and observe that $R'(0)=E(Z_t)-\underline{b}E(L_t)<0$ by condition (iii). Note that $R(0)=1$, and then there exists $s_0\in(0,1)$ such that $\gamma_0:=R(s_0/2)<1$.

In addition, for the discount factor and reward function, we impose the following conditions:

(iv) Suppose that there exists a constant $\alpha\in(0,\frac{1}{\gamma_0+1})$ such that $\sup_{x\in X}\alpha(x)\le\alpha$.

(v) The reward function $r(x,a,b)$ is measurable on $X\times A\times B$ and continuous on $A(x)\times B(x)$, and satisfies that

$$|r((y,i),a,b)|\le\beta\exp(s_0y/2),\ \forall(y,i)\in X,(a,b)\in A(x)\times B(x),$$

where β is a positive constant.

Proposition 1. *Under the above conditions of (i)-(v), Example 1 satisfies Assumptions 1 and 2 in Theorem 1, then there exist the optimal policies of the two-person zero-sum stochastic game.*

Proof. It is clear that Assumption 1(a) holds. Now, let $\omega(x)=\omega(y,i):=\exp(s_0y/2)$ for each $x\in X$, $A_1:=\Big\{(z,l)\in\mathbb{R}_+\times\mathbb{R}_+\Big|y-(a+b)l+z\le0\Big\}$ and $A_1^c:=\mathbb{R}_+\times\mathbb{R}_+-A_1$, then we have

$$\begin{aligned}
&\int_X\omega(\hat{x})q(\mathrm{d}\hat{x}|(y,i),a,b)\\
=&E\Big[\omega(x_{t+1})\Big|x_t=(y,i),a_t=a,b_t=b\Big]\\
=&\sum_{j=1}^N\Big\{\iint_{A_1}\omega(0,j)\rho^*(z,l)\mathrm{d}z\mathrm{d}l+\iint_{A_1^c}\omega(y-(a+b)l+z,j)\rho^*(z,l)\mathrm{d}z\mathrm{d}l\Big\}p_{ij}\\
\le&\sum_{j=1}^N\Big\{\omega(0,j)+\exp(s_0y/2)\iint_{A_1^c}\exp\Big\{\frac{s_0}{2}[z-(a+b)l]\Big\}\rho^*(z,l)\mathrm{d}z\mathrm{d}l\Big\}p_{ij}\\
\le&\sum_{j=1}^N\Big\{1+\exp(s_0y/2)\iint_{A_1^c}\exp\Big\{\frac{s_0}{2}(z-\underline{b}l)\Big\}\rho^*(z,l)\mathrm{d}z\mathrm{d}l\Big\}p_{ij}\\
\le&\sum_{j=1}^N\Big\{1+\gamma_0\omega(x)\Big\}p_{ij}\\
\le&(1+\gamma_0)\omega(x),
\end{aligned}$$

which shows that Assumption 1(b) holds when $\gamma := \gamma_0 + 1$, and by condition (iv) it holds that $\alpha\gamma < 1$.

Note that, the joint density $\rho^*(\cdot,\cdot)$ is continuous and bounded by conditions (i) and (ii), and for any bounded measurable function $u \in \mathscr{B}(X)$, we have

$$\begin{aligned}\int_X u(\hat{x})q(\mathrm{d}\hat{x}|(y,i),a,b) &= E\Big[u(x_{t+1})\Big|x_t = (y,i), a_t = a, b_t = b\Big] \\ &= \sum_{j=1}^{N}\Big\{\iint_{A_1} u(0,j)\rho^*(z,l)\mathrm{d}z\mathrm{d}l + \iint_{A_1^c} u(y-(a+b)l+z,j)\rho^*(z,l)\mathrm{d}z\mathrm{d}l\Big\}p_{ij},\end{aligned}$$

which is obviously continuous on $(a,b) \in A(x) \times B(x)$, and so is the weight function $\omega(x)$. Therefore Assumptions 1 and 2 hold, by Theorem 1, there exist the optimal policies. □

5. Conclusions

This article considers two-person zero-sum Markov games with Borel state space and action space, unbounded reward function and varying discount factors and proves the existence of a value and associated equilibrium policies in these games. The relevant theories in this paper provide a solid theoretical basis to study the calculation of optimal value and optimal policy pairs. However, it does not give the complete algorithms to calculate the optimal value and optimal policy pairs of two-person zero-sum stochastic games, which is our important research work in the future.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 11961005), Opening Project of Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University (No. 2021021), Characteristic Innovation Projects of General Colleges and Universities in Guangdong Province (No. 2018KTSCX253), Guangzhou Science and Technology Plan Project (No. 202102080420) and Guangdong University (New Generation Information Technology) Key Field Project(No. 2020ZDZX3019). The authors also thank the editors and the referee for many valuable comments and suggestions which have improved this paper.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. L. S. Shapley, Stochastic games, *P. Natl. Acad. Sci. USA,* **39** (1953), 1095–1100.
2. A. Maitra, T. Parthasarathy, On stochastic games, *J. Appl. Probab.,* **5** (1970), 289–300.
3. T. Parthasarathy, Discounted, positive and noncooperative stochastic games, *Int. J. Game Theory,* **2** (1973), 25–37.
4. H. Couwenbergh, Stochastic games with metric state space, *Int. J. Game Theory,* **9** (1980), 25–36.
5. J. Filar, K. Vrieze, *Competitive Markov Decision Processes*, New York: Springer-Verlag, 1997.

6. A. S. Nowak, Universally measurable strategies in zero-sum stochastic games, *Ann. Probab.,* **13** (1985), 269–287.

7. A. Neyman, S. Sorin, *Stochastic Games and Applications*, Dordrecht: Kluwer Academic Publishers, 2003.

8. X. P. Guo, O. Hernández-Lerma, Zero-sum games for continuous-time jump Markov processes in Polish spaces: discounted payoffs, *Adv. Appl. Probab.,* **39** (2007), 645–668.

9. J. Minjárez-Sosa, F. Luque-Vásquez, Two person zero-sum semi-Markov games with unknown holding times distribution on one side: a discounted payoff criterion, *Appl. Math. Opt.,* **57** (2008), 289–305.

10. O. Hernández-Lerma, J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, New York:Springer-Verlag, 1996.

11. X. P. Guo, O. Hernández-Lerma, Zero-sum continuous-time Markov games with unbounded transition and discounted payoff rates, *Bernoulli,* **11** (2005), 1009–1029.

12. M. Schäll, Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal, *Z. Wahrscheinlichkeitstheor Verw. Geb.,* **32** (1975), 179–196.

13. J. González-Hernández, R. López-Martinez, J. Pérez-Hernández, Markov control processes with randomized discounted cost, *Math. Methods Oper. Res.,* **65** (2007), 27–44.

14. J. González-Hernández, R. López-Martinez, J. Minjárez-Sosa, Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion, *Kybernetika,* **45** (2009), 737–754.

15. Y. Zhang, Convex analytic approach to constrained discounted Markov decision processes with non-constant discount factors, *Top,* **21** (2013), 378–408.

16. X. Wu, X. P. Guo, First Passage Optimality and Variance Minimisation of Markov Decision Processes with Varying Discount Factors, *J. Appl. Probab.,* **52** (2015), 441–456.

17. L. I. Sennott, Nonzero-sum stochastic games with unbounded costs: discounted and average cost cases, *Math. Method Oper. Res.,* **40** (1994), 145–162.

18. X. P. Guo, Q. X. Zhu, Average optimality for Markov decision processes in Borel spaces: A new condition and approach, *J. Appl. Probab.,* **43** (2006), 318–334.

19. X. P. Guo, O. Hernández-Lerma, Nonzero-sum games for continuous-time Markov chains with unbounded discounted payoffs, *J. Appl. Probab.,* **42** (2005), 303–320.

20. K. Fan, Minimax theorems, *P. Natl. Acad. Sci. USA,* **39** (1953), 42–47.

21. A. S. Nowak, S. Andrzej, Measurable selection theorems for minimax stochastic optimization problems, *SIAM J. Control Optim.,* **23** (1985), 466–476.