Dyn Games Appl (2019) 9:103–121
https://doi.org/10.1007/s13235-018-0248-8

Zero-Sum Markov Games with Random State-Actions-Dependent Discount Factors: Existence of Optimal Strategies

David González-Sánchez[1] · **Fernando Luque-Vásquez**[2] · **J. Adolfo Minjárez-Sosa**[2]

Published online: 3 March 2018

Abstract This work deals with a class of discrete-time zero-sum Markov games under a discounted optimality criterion with random state-actions-dependent discount factors of the form $\tilde{\alpha}(x_n, a_n, b_n, \xi_{n+1})$, where x_n, a_n, b_n, and ξ_{n+1} are the state, the actions of players, and a random disturbance at time n, respectively, taking values in Borel spaces. Assuming possibly unbounded payoff, we prove the existence of a value of the game as well as a stationary pair of optimal strategies.

Keywords Markov games · Discounted optimality · Nonconstant discount factor

Mathematics Subject Classification 91A15 · 91A50 · 60J05

1 Introduction

The paper deals with a class of discrete-time zero-sum discounted Markov games with non-constant discount factors of the form

$$\tilde{\alpha}(x_n, a_n, b_n, \xi_{n+1}), \tag{1}$$

Work supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) under Grant CB2015/254306.

✉ J. Adolfo Minjárez-Sosa
aminjare@gauss.mat.uson.mx

David González-Sánchez
david.glzsnz@gmail.com

Fernando Luque-Vásquez
fluque@gauss.mat.uson.mx

1 CONACYT–Universidad de Sonora, Rosales s/n, 83000 Hermosillo, Sonora, Mexico

2 Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, 83000 Hermosillo, Sonora, Mexico

where x_n is the state of the game, a_n and b_n represent the actions of players 1 and 2, respectively, at time n, and $\{\xi_n\}$ is a sequence of independent and identically distributed random variables with common distribution θ representing a random disturbance at each time. The one-stage payoff (or utility) $r(x_n, a_n, b_n)$ accumulates in an infinite horizon by means of the functional

$$E\left[\sum_{n=0}^{\infty}\prod_{k=0}^{n-1}\tilde{\alpha}(x_k, a_k, b_k, \xi_{k+1})r(x_n, a_n, b_n)\right], \tag{2}$$

which defines the total expected discounted payoff with random discount factors depending on the state and the actions. Thus, in this scenario, our main objective is to prove the existence of a value of the game and a pair of optimal strategies.

Among the optimality criteria to study zero-sum and nonzero-sum Markov games, the discounted payoff with a constant discount factor is the best understood. It has been analyzed under several approaches, for instance, dynamic programming via the Shapley's equation, linear programming, estimation, and control procedures (see, e.g., [9,18–20,24–28,36,37]), and Nash equilibrium [5,13,31]. Moreover, its main applications are in economic and financial models where the discount factor is a function of the interest rate. Hence, considering a constant discount factor could be restrictive in problems where such an interest rate is random. It is in these situations where the need arises to consider a function as (1) representing the discount factor.

Even though the usual applications of the discounted criterion is in economic and financial models, there are other problems where a discount factor as (1) appears naturally. For instance, consider a game that is played as follows. At stage n, when the game is in state x_n and once the players choose the actions (a_n, b_n), player 2 pays $r(x_n, a_n, b_n)$ to player 1. Then, there is a positive probability the game stops which is influenced by (x_n, a_n, b_n), otherwise the game moves to a new state x_{n+1} according to a transition law, and the process is repeated. Under these circumstances, the performance of the zero-sum game is measured by the total expected payoff criterion with a random horizon τ of the form

$$E\left[\sum_{n=0}^{\tau} r(x_n, a_n, b_n)\right]. \tag{3}$$

However, we prove that (3) can be written as (2) with

$$\tilde{\alpha}(x_n, a_n, b_n, \xi_{n+1}) := 1 - \gamma(x_n, a_n, b_n),$$

where $\gamma(x_n, a_n, b_n)$ is the probability the game stops at stage n.

Another example with random state-actions-dependent discount factors is the following zero-sum semi-Markov game. Let $\{\xi_n\}$ be a sequence of independent and identically distributed random variables with exponential distribution representing the sojourn (or holding) times. In addition, let $\gamma(x_n, a_n, b_n)$ be the discount factor imposed at stage n. Then, by defining $\tilde{\alpha}(x_n, a_n, b_n, \xi_{n+1}) := \exp(-\gamma(x_n, a_n, b_n)\xi_{n+1})$, the total expected discounted payoff takes the form (2).

Typically, the existence of optimal strategies in zero-sum Markov games is studied via Shapley's equation. Such an approach has the advantage that it allows to apply the nice contractive properties of the minimax (maximin) operator. In our case, since the discount factor is a nonconstant function $\tilde{\alpha}$ which explicitly depends on the random disturbance process $\{\xi_n\}$, it is not possible to obtain, at least directly, a Shapley-like equation. To obtain the advantages that such an equation entails, we first need to establish a representation of the performance index related to (2) in terms of the common distribution θ of the random variables

$\{\xi_n\}$. However, due to measurability issues, such a representation is only possible when the players are restricted to use Markov strategies, not arbitrary strategies (see Proposition 2). Taking into account this fact, we then prove the sufficiency of Markov (stationary) strategies in the sense that if a pair of stationary strategies is optimal with respect to the Markov strategies, so is with respect to all strategies (see Proposition 1). It is worth remarking that this fact is a well-known result in zero-sum games under the standard discounted criterion (see, e.g., [21,30]). In our game model, from the fact that the discount factor is a function of the game process $\{(x_n, a_n, b_n, \xi_{n+1})\}$, such a result is not a direct consequence from [21], and therefore, some modifications should be made. Hence, for completeness of our work, we have included its proof. Once ensured the sufficiency of Markov strategies, we establish, in Theorem 1, the existence of a value of the game and an optimal pair of stationary strategies.

Problems with nonconstant discount factors have been extensively studied for Markov decision processes from several point of views (see, e.g., [3,8,10–12,23,34,38]). In particular, control processes with random state-action-dependent discount factors are analyzed in [23], which is close to our context. Nonetheless, in addition to the usual difficulties of dealing with stochastic games, proving Proposition 1 requires different arguments from those followed in the single-controller case.

The paper is organized as follows. In Sect. 2 we present the game model we deal with. Next, in Sect. 3 we introduce the optimality criterion and the main properties related with the sufficiency of Markov strategies. The existence of the value of the game and the pair of optimal strategies is established in Sect. 4, whereas the proofs are remitted to Sect. 6. In order to illustrate our results, in Sect. 5 we present some examples with nonconstant discount factors. The first one is a financial model where the discount factor is function of a random interest rate. Next we present an example of a game with random horizon and nondiscounted payoff criterion which is equivalent to a game where the discount factor is a state-actions-dependent function representing the probability of continuing the game. The third example is a semi-Markov game where the payoffs are exponentially discounted according to a random state-actions-dependent discount factor. We also present some insights into the fulfillment of our hypotheses.

Notation As usual, $\mathbb{N}$ (respectively $\mathbb{N}_0$) denotes the set of positive (resp. nonnegative) integers. On the other hand, given a Borel space X (that is, a Borel subset of a complete and separable metric space) its Borel sigma-algebra is denoted by $\mathcal{B}(X)$, and "measurable", for either sets or functions, means "Borel measurable". Let X and Y be Borel spaces. Then a stochastic kernel $\gamma(\mathrm{d}x \mid y)$ on X given Y is a function such that $\gamma(\cdot \mid y)$ is a probability measure on X for each fixed $y \in Y$, and $\gamma(B \mid \cdot)$ is a measurable function on Y for each fixed $B \in \mathcal{B}(X)$. The space of probability measures on X is denoted by $\mathbb{P}(X)$, which is endowed with the weak topology. In addition, we denote by $\mathbb{P}(X \mid Y)$ the family of stochastic kernels on X given Y.

2 The Game Model

A zero-sum Markov game model with random state-actions-dependent discount factors is defined by the collection

$$\mathcal{GM} := (\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbb{K}_{\mathbf{A}}, \mathbb{K}_{\mathbf{B}}, \mathbf{S}, Q, \tilde{\alpha}, r), \tag{4}$$

satisfying the following conditions. The state space $\mathbf{X}$ and the action sets $\mathbf{A}$ and $\mathbf{B}$ for players 1 and 2, respectively, as well as the discount factors disturbance space $\mathbf{S}$, are assumed to be Borel spaces. The constraint sets $\mathbb{K}_{\mathbf{A}}$ and $\mathbb{K}_{\mathbf{B}}$ are Borel subsets of $\mathbf{X} \times \mathbf{A}$ and $\mathbf{X} \times \mathbf{B}$, respectively. For each $x \in \mathbf{X}$, the x-sections

$$A(x) := \{a \in \mathbf{A} : (x, a) \in \mathbb{K}_{\mathbf{A}}\}$$

and

$$B(x) := \{b \in \mathbf{B} : (x, a) \in \mathbb{K}_{\mathbf{B}}\}$$

represent the admissible actions or controls sets for players 1 and 2, respectively, and the set

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

of admissible state-actions triplets is a Borel subset of $\mathbf{X} \times \mathbf{A} \times \mathbf{B}$. The transition law $Q(\cdot|x, a, b)$ is a stochastic kernel on $\mathbf{X}$ given $\mathbb{K}$, and $\tilde{\alpha} : \mathbb{K} \times \mathbf{S} \to (0, 1)$ is a measurable function which gives the discount factor $\tilde{\alpha}(x_n, a_n, b_n, \xi_{n+1})$ at stage $n \in \mathbb{N}$, where $\{\xi_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ taking values in $\mathbf{S}$ with common distribution $\theta \in \mathbb{P}(S)$. That is

$$\theta(S) = P(\xi_n \in S), \quad S \in \mathcal{B}(\mathbf{S}), n \in \mathbb{N}.$$

Finally, $r(\cdots)$ is a real-valued measurable function on $\mathbb{K}$ that represents the one-stage payoff function.

The game is played as follows. At the initial state $x_0 \in \mathbf{X}$, the players independently choose actions $a_0 \in A(x_0)$ and $b_0 \in B(x_0)$. Then player 1 receives a payoff $r(x_0, a_0, b_0)$ from player 2, and the game jumps to a new state x_1 according to the transition law $Q(\cdot|x_0, a_0, b_0)$, and the random disturbance ξ_1 comes in. Once the system is in state x_1, the players select actions $a_1 \in A(x_1)$ and $b_1 \in B(x_1)$ and player 1 receives a discounted payoff $\tilde{\alpha}(x_0, a_0, b_0, \xi_1) r(x_1, a_1, b_1)$ from player 2. Next the system moves to a state x_2, and the process is repeated over and over again. In general, at stage $n \in \mathbb{N}$, player 1 receives from player 2 a discounted payoff of the form

$$\tilde{\Gamma}_n r(x_n, a_n, b_n) \tag{5}$$

where

$$\tilde{\Gamma}_n := \prod_{k=0}^{n-1} \tilde{\alpha}(x_k, a_k, b_k, \xi_{k+1}) \text{ if } n \in \mathbb{N}, \text{ and } \tilde{\Gamma}_0 = 1. \tag{6}$$

Thus, the goal of player 1 (player 2, resp.) is to maximize (minimize, resp.) the total expected discounted payoff defined by the accumulation of the one-stage payoffs (5) over an infinite horizon.

The actions chosen by players at each stage are selected by rules known as strategies which are defined as follows.

Let $\mathbb{H}_0 := \mathbf{X}$ and $\mathbb{H}_n := \mathbb{K} \times \mathbf{S} \times \mathbb{H}_{n-1}$ for $n \in \mathbb{N}$. For each $n \in \mathbb{N}_0$, an element $h_n \in \mathbb{H}_n$ takes the form

$$h_n := (x_0, a_0, b_0, s_1, \ldots, x_{n-1}, a_{n-1}, b_{n-1}, s_n, x_n),$$

which represents the history of the game up to time n. A *strategy* for player 1 is a sequence $\pi^1 = \{\pi_n^1\}$ of stochastic kernels $\pi_n^1 \in \mathbb{P}(\mathbf{A}|\mathbb{H}_t)$ such that $\pi_n^1(A(x_n)|h_n) = 1$ for every $h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$. We denote by Π^1 the family of all strategies for player 1.

For each $x \in \mathbf{X}$, let $\mathbb{A}(x) := \mathbb{P}(A(x))$ and $\mathbb{B}(x) := \mathbb{P}(B(x))$. We denote by Φ^1 the class of all stochastic kernels $\varphi^1 \in \mathbb{P}(\mathbf{A}|\mathbf{X})$ such that $\varphi^1(\cdot|x) \in \mathbb{A}(x)$, $x \in \mathbf{X}$, and by Φ^2

the class of all stochastic kernels $\varphi^2 \in \mathbb{P}(\mathbf{B}|\mathbf{X})$ such that $\varphi^2(\cdot|x) \in \mathbb{B}(x)$, $x \in \mathbf{X}$. Hence, a strategy $\pi^1 = \{\pi_n^1\} \in \Pi^1$ is called a *Markov strategy* if there exists $\varphi_n^1 \in \Phi^1$ such that $\pi_n^1(\cdot|h_n) = \varphi_n^1(\cdot|x_n)$ for every h_n, $n \in \mathbb{N}_0$. The class of all Markov strategies for player 1 is denoted by Π_M^1. Now, a Markov strategy is called stationary if $\varphi_n^1 = \varphi^1$ for every $n \in \mathbb{N}_0$ and some stochastic kernel φ^1 in Φ^1. The set of stationary strategies for player 1 is denoted by Π_S^1. The sets Π^2, Π_M^2, and Π_S^2 corresponding to player 2 are defined similarly.

According to the previous definitions, and by using a standard convention, a Markov strategy $\varphi^i \in \Pi_M^i$ takes the form $\varphi^i = \{\varphi_0^i, \varphi_1^i, \ldots\} =: \{\varphi_n^i\}$, for $i = 1, 2$. In particular, for stationary strategies, we have $\varphi^i = \{\varphi^i, \varphi^i, \ldots\} = \{\varphi^i\}$.

The game process. Let $(\Omega', \mathcal{F}')$ be the measurable space consisting of the sample space $\Omega' = (\mathbb{K} \times \mathbf{S})^\infty$ and its product σ-algebra $\mathcal{F}'$. Following standard arguments (see, e.g., [6]), we have that for each pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and initial state $x_0 = x \in \mathbf{X}$, there exists a unique probability measure $P_x^{\pi^1,\pi^2}$ and a stochastic process $\{(x_n, a_n, b_n, \xi_{n+1})\}$, where x_n, a_n, b_n, and ξ_{n+1} represent the state, the actions of players, and the random disturbance in the discount factor, respectively, at stage $n \in \mathbb{N}_0$, satisfying

$$P_x^{\pi^1,\pi^2}[x_0 \in X] = \delta_x(X), \quad X \in \mathcal{B}(\mathbf{X}); \tag{7}$$

$$P_x^{\pi^1,\pi^2}[a_n \in A, b_n \in B|h_n] = \pi_n^1(A|h_n)\,\pi_n^2(B|h_n), \quad A \in \mathcal{B}(\mathbf{A}), B \in \mathcal{B}(\mathbf{B}); \tag{8}$$

$$P_x^{\pi^1,\pi^2}\left[x_{n+1} \in X|h_n, a_n, b_n, \xi_{n+1}\right] = Q(X|x_n, a_n, b_n), \quad X \in \mathcal{B}(\mathbf{X}); \tag{9}$$

$$P_x^{\pi^1,\pi^2}\left[\xi_{n+1} \in S|h_n, a_n, b_n\right] = \theta(S), \quad S \in \mathcal{B}(\mathbf{S}), \tag{10}$$

where $\delta_x(\cdot)$ is the Dirac measure concentrated at x. We denote by $E_x^{\pi^1,\pi^2}$ the expectation operator with respect to $P_x^{\pi^1,\pi^2}$. The stochastic process $\{x_n\}$ defined on $(\Omega', \mathcal{F}', P_x^{\pi^1,\pi^2})$ is called *game process*.

3 The Optimality Criterion

According to (5) and (6), given the initial state $x_0 = x \in \mathbf{X}$ and a pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, the total expected discounted payoff—with random state-actions-dependent discount factors—is defined as

$$\tilde{V}(x, \pi^1, \pi^2) := E_x^{\pi^1,\pi^2}\left[\sum_{n=0}^{\infty} \tilde{\Gamma}_n r(x_n, a_n, b_n)\right]. \tag{11}$$

The lower and the upper value of the game are

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

respectively, for each initial state $x \in \mathbf{X}$. Of course, $U(\cdot) \geq L(\cdot)$; however, if $U(\cdot) = L(\cdot)$ holds, then the common function is called the *value of the game* and is denoted by $V^*(\cdot)$.

Suppose the game has a value V^*. A strategy $\pi_*^1 \in \Pi^1$ is said to be *optimal for player 1* if

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

Similarly, a strategy $\pi_*^2 \in \Pi^2$ is said to be *optimal for the player 2* if

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

Hence, the pair (π_*^1, π_*^2) is called an *optimal pair* of strategies. Observe that $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ is an optimal pair if and only if

$$\tilde{V}(x, \pi^1, \pi_*^2) \leq \tilde{V}(x, \pi_*^1, \pi_*^2) \leq \tilde{V}(x, \pi_*^1, \pi^2), \quad \forall (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2, \ x \in \mathbf{X}. \tag{12}$$

An important fact in our analysis on the existence of a value of the game is the sufficiency of Markov strategies in the following sense.

Proposition 1 *Let $(\varphi_*^1, \varphi_*^2) \in \Pi_S^1 \times \Pi_S^2$ be an optimal pair with respect to the Markov strategies, i.e.,*

$$\tilde{V}(x, \varphi^1, \varphi_*^2) \leq \tilde{V}(x, \varphi_*^1, \varphi_*^2) \leq \tilde{V}(x, \varphi_*^1, \varphi^2), \quad \forall (\varphi^1, \varphi^2) \in \Pi_M^1 \times \Pi_M^2, \ x \in \mathbf{X}. \tag{13}$$

Then $(\varphi_*^1, \varphi_*^2)$ is an optimal pair with respect to all strategies, i.e., (12) holds.

By virtue of Proposition 1 we can restrict our study to the set of Markov strategies. Furthermore, over the Markov strategies, we can express the performance index (11) in terms of the distribution of the discount factor random disturbance θ. We proceed to establish this fact in a precise way.

We define the mean discount factor function $\alpha_\theta : \mathbb{K} \to (0, 1)$ as

$$\alpha_\theta(x, a, b) := \int_S \tilde{\alpha}(x, a, b, s)\theta(\mathrm{d}s), \quad (x, a, b) \in \mathbb{K}, \tag{14}$$

and denote

$$\Gamma_n = \prod_{k=0}^{n-1} \alpha_\theta(x_k, a_k, b_k) \text{ if } n \in \mathbb{N}, \text{ and } \Gamma_0 = 1. \tag{15}$$

For each pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and initial state $x \in \mathbf{X}$, we define

$$V(x, \pi^1, \pi^2) := E_x^{\pi^1, \pi^2} \left[\sum_{n=0}^{\infty} \Gamma_n r(x_n, a_n, b_n) \right]. \tag{16}$$

Proposition 2 *For each initial state $x \in \mathbf{X}$ and pair of strategies $(\varphi^1, \varphi^2) \in \Pi_M^1 \times \Pi_M^2$,*

$$V(x, \varphi^1, \varphi^2) = \tilde{V}(x, \varphi^1, \varphi^2). \tag{17}$$

4 Existence of Optimal Strategies

To ease notation, the probability measures $\varphi^1(\cdot|x) \in \mathbb{A}(x)$ and $\varphi^2(\cdot|x) \in \mathbb{B}(x)$, $x \in \mathbf{X}$, are written $\varphi^i(x) = \varphi^i(\cdot|x)$, $i = 1, 2$. In addition, for a measurable function $u : \mathbb{K} \to \mathbb{R}$,

$$u(x, \varphi^1, \varphi^2) = u(x, \varphi^1(x), \varphi^2(x)) := \int_{B(x)} \int_{A(x)} u(x, a, b)\varphi^1(da|x)\varphi^2(db|x). \tag{18}$$

For instance, for $x \in \mathbf{X}$, we have

$$r(x, \varphi^1, \varphi^2) := \int_{B(x)} \int_{A(x)} r(x, a, b)\varphi^1(da|x)\varphi^2(db|x),$$

and

$$Q(X|x,\varphi^1,\varphi^2) := \int_{B(x)}\int_{A(x)} Q(X|x,a,b)\varphi^1(da|x)\varphi^2(db|x), \quad X \in \mathcal{B}(\mathbf{X}).$$

The existence of a value of the game as well as a pair of optimal strategies is analyzed under the following conditions.

Assumption 1 The game model (4) satisfies the following:
(a) For each $x \in \mathbf{X}$, the sets $A(x)$ and $B(x)$ are compact.
(b) For each $(x,a,b) \in \mathbb{K}$, $r(x,\cdot,b)$ is upper semicontinuous (usc) on $A(x)$, and $r(x,a,\cdot)$ is lower semicontinuous (lsc) on $B(x)$. Moreover, there exists a constant $r_0 > 0$ and a function $W : \mathbf{X} \to [1,\infty)$ such that

$$|r(x,a,b)| \le r_0 W(x), \tag{19}$$

and the functions

$$\int_{\mathbf{X}} W(y)Q(\mathrm{d}y|x,\cdot,b) \quad \text{and} \quad \int_{\mathbf{X}} W(y)Q(\mathrm{d}y|x,a,\cdot) \tag{20}$$

are continuous on $A(x)$ and $B(x)$, respectively.
(c) For each $(x,a,b) \in \mathbb{K}$ and each bounded measurable function u on $\mathbf{X}$, the functions

$$\int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,\cdot,b) \quad \text{and} \quad \int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,a,\cdot)$$

are continuous on $A(x)$ and $B(x)$, respectively.
(d) The function $\tilde{\alpha}(x,a,b,s)$ is continuous on $\mathbb{K} \times \mathbf{S}$, and

$$\alpha^* := \sup_{(x,a,b)\in\mathbb{K}} \alpha_\theta(x,a,b) < 1. \tag{21}$$

(e) There exists a positive constant β such that $1 \le \beta < (\alpha^*)^{-1}$, and for every $(x,a,b) \in \mathbb{K}$

$$\int_{\mathbf{X}} W(y)Q(\mathrm{d}y \mid x,a,b) \le \beta W(x). \tag{22}$$

For each measurable function $u : \mathbf{X} \to \mathbb{R}$, we define the W-norm as

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

and let $\mathbb{B}_W$ be Banach space of all real-valued measurable functions defined on $\mathbf{X}$ with finite W-norm. It is easy to prove that under Assumption 1, the Shapley operator

$$Tu(x) := \inf_{\varphi^2\in\mathbb{B}(x)} \sup_{\varphi^1\in\mathbb{A}(x)} \hat{T}(u,x,\varphi^1,\varphi^2), \quad x \in \mathbf{X}, \tag{23}$$

maps $\mathbb{B}_W$ into itself, where

$$\hat{T}(u,x,a,b) := r(x,a,b) + \alpha_\theta(x,a,b)\int_{\mathbf{X}} u(y)Q(\mathrm{d}y|x,a,b), \quad (x,a,b) \in \mathbb{K}. \tag{24}$$

Moreover, as will be established later, the interchange of inf and sup in (23) holds.

We now state our main results as follows.

Theorem 1 *Suppose that Assumption 1 holds. Then*

(a) the game $\mathcal{GM}$ (4) has a value $V^* \in \mathbb{B}_W$,

(b) the value V^* is the unique function in $\mathbb{B}_W$ such that $TV^* = V^*$, and
(c) there exist $\varphi_*^1(x) \in \mathbb{A}(x)$ and $\varphi_*^2 \in \mathbb{B}(x)$ such that

$$V^*(x) = \hat{T}(V^*, x, \varphi_*^1, \varphi_*^2) \tag{25}$$

$$= \max_{\varphi^1 \in \mathbb{A}(x)} \hat{T}(V^*, x, \varphi^1, \varphi_*^2) \tag{26}$$

$$= \min_{\varphi^2 \in \mathbb{B}(x)} \hat{T}(V^*, x, \varphi_*^1, \varphi^2), \quad \forall x \in \mathbf{X}. \tag{27}$$

In addition, the stationary strategies $\varphi_*^1 = \{\varphi_*^1\} \in \Pi_S^1$ and $\varphi_*^2 = \{\varphi_*^2\} \in \Pi_S^2$ form an optimal pair of strategies respect to the Markov strategies. Hence, from Proposition 1, $(\varphi_*^1, \varphi_*^2)$ is an optimal pair of strategies for the game $\mathcal{GM}$.

5 Examples

In order to illustrate the theory developed above, we present two classes of examples. In the first one, Examples 1–3, we describe the potential applications of indices with nonconstant discount factors. Specifically, in Example 1 we present an application of this kind of optimality criteria in games involving monetary units in which the discount factor is a function of a random interest rate and/or inflation rate, and in Examples 2 and 3, the state-actions-dependent discount factors appear in a natural manner. Finally, Examples 4 and 5, which constitute the second class, are devoted to illustrate the assumptions imposed on the game model.

Example 1 (Monetary payoffs) Consider the game model (4). In general, r is a *utility function* which represents the preferences over the outcomes (x, a, b) in $\mathbb{K}$, and so money is not necessarily involved ([22, p. 9] or [32, p. 13]). In this example, we assume that $r(x, a, b)$ is indeed measured in monetary units. Let

$$\tilde{\alpha}(x, a, b, \xi) = \frac{1}{1 + \rho - \xi},$$

where $\rho > 0$ is the (constant) *nominal* interest rate and ξ represents the *inflation* rate between two consecutive periods; thus $\rho - \xi$ is the *real* interest rate which is random. Assume that ξ takes values in $S = [\underline{s}, \overline{s}]$, with $0 < \underline{s} < \overline{s} < \rho$. Hence, Assumption 1 (d) trivially follows since

$$\alpha^* = \frac{1}{1 + \rho - \overline{s}} < 1.$$

Example 2 (A nondiscounted payoff game with random horizon) In Sect. 2 we described how the discounted game with infinite horizon is played. Let us consider the game model

$$(\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbb{K}_{\mathbf{A}}, \mathbb{K}_{\mathbf{B}}, Q, \alpha, r) \tag{28}$$

with the following alternative playing where the horizon is random. For simplicity, we are not considering the disturbance space $\mathbf{S}$. At state x_n, players 1 and 2 choose actions (a_n, b_n) and respectively receive $r(x_n, a_n, b_n)$ and $-r(x_n, a_n, b_n)$, then with probability $1-\alpha(x_n, a_n, b_n)$ the game stops; otherwise, the system moves to another state x_{n+1} according to $Q(\cdot \mid x_n, a_n, b_n)$, where the nondiscounted payoff $r(x_{n+1}, a_{n+1}, b_{n+1})$ is determined by the actions (a_{n+1}, b_{n+1}). We assume that there is $\gamma \in (0, 1)$ such that $1 - \alpha(x_n, a_n, b_n) \geq \gamma$. Thus

$$\alpha^* := \sup_{(x,a,b) \in \mathbb{K}} \alpha(x, a, b) \leq 1 - \gamma < 1.$$

We will show that the total expected payoff in this game takes the form (16). For this purpose, let x^* and (a^*, b^*) be artificial state and actions. We define the game model

$$\mathcal{GM}^* = (X^*, A^*, B^*, \mathbb{K}_{A^*}, \mathbb{K}_{B^*}, Q^*, r^*)$$

where $X^* = X \cup \{x^*\}$, $A^* = A \cup \{a^*\}$, $B^* = B \cup \{b^*\}$, and the corresponding x-sections are the sets

$$A^*(x) := \begin{cases} \{a^*\} \ if \ x = x^*, \\ A(x) \ if \ x \in X; \end{cases}$$

$$B^*(x) := \begin{cases} \{b^*\} \ if \ x = x^*, \\ B(x) \ if \ x \in X; \end{cases}$$

The transition law Q^* among the states in X^* is a stochastic kernel on X^* given the set

$$\mathbb{K}^* := \left\{(x, a, b) : x \in X^*, a \in A^*(x), b \in B^*(x)\right\}$$

defined as follows: For $(x, a, b) \in \mathbb{K}$,

$$\begin{aligned} Q^*(D \mid x, a, b) &:= \alpha(x, a, b) Q(D \mid x, a, b), \quad D \in \mathcal{B}(X), \\ Q^*(\{x^*\} \mid x, a, b) &:= 1 - \alpha(x, a, b), \\ Q^*(\{x^*\} \mid x^*, a^*, b^*) &:= 1. \end{aligned}$$

Finally, the payoff function $r^* : \mathbb{K}^* \to \mathbb{R}$ is given by

$$r^*(x, a, b) := \begin{cases} r(x, a, b) & \text{if } (x, a, b) \in \mathbb{K}, \\ 0 & \text{if } (x, a, b) = (x^*, a^*, b^*). \end{cases}$$

On the other hand, let $(\Omega', \mathcal{F}')$ be the measurable space associated with the game model $\mathcal{GM}^*$ (see Sect. 2) and define the first passage time $\tau : \Omega' \to \mathbb{N}_0 \cup \{+\infty\}$ as

$$\tau(x_0, a_0, b_0, \ldots) := \inf\{n \in \mathbb{N}_0 : x_n = x^*\},$$

where, as usual, $\inf \emptyset = +\infty$. For each pair of strategies $(\varphi^1, \varphi^2) \in \Pi_M^1 \times \Pi_M^2$ and initial state $x \in X$, the total expected payoff with random horizon τ takes the form

$$V_\tau(x, \varphi^1, \varphi^2) := E_x^{\varphi^1, \varphi^2} \sum_{n=0}^{\tau} r(x_n, a_n, b_n). \tag{29}$$

Then a straightforward calculation shows that the performance index (29) can be written as a performance index with state-actions-dependent discount factors. Specifically, by following similar arguments as the proof of Proposition 2, it is possible to prove the equality

$$V_\tau(x, \varphi^1, \varphi^2) = V(x, \varphi^1, \varphi^2) = E_x^{\varphi^1, \varphi^2} \sum_{n=0}^{\infty} \prod_{k=0}^{n-1} \alpha(x_k, a_k, b_k) r^*(x_n, a_n, b_n).$$

Hence, provided that Assumption 1 holds, the game (28) with random horizon has a value and there exists a pair of optimal stationary strategies due to Theorem 1.

This game model with random horizon is in the spirit of Shapley's [35] seminal paper where finite stochastic games were introduced. Similar games but considering continuous and bounded payoff functions in the performance index (16) and countable state space were studied by Rieder [33]. On the other hand, Markov decision models with random horizon and Borel spaces have also been studied under several settings (see, for instance, [2,4]); however, such control processes are assumed to be stopped with constant probability. Therefore, our example generalizes many results in the existing literature.

Example 3 (A semi-Markov game) Consider a zero-sum semi-Markov game (see, e.g., [17, 20,24]) where the sojourn (or holding) times $\xi_1, \xi_2, \ldots$ are i.i.d. random variables with common exponential distribution with parameter $\lambda > 0$. Suppose that the discount factor is a continuous function $\gamma : \mathbb{K} \to (d, \infty)$ where $d > 0$. Then the expected discounted payoff is

$$\bar{V}(x, \pi^1, \pi^2) := E_x^{\pi^1, \pi^2}\left[r(x_0, a_0, b_0) + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \mathrm{e}^{-\gamma(x_k, a_k, b_k)\xi_{k+1}} r(x_n, a_n, b_n) \right].$$

If we define the function $\tilde{\alpha} : \mathbb{K} \times \mathbf{S} \to (0, 1)$ as

$$\tilde{\alpha}(x, a, b, \xi) = \mathrm{e}^{-\gamma(x,a,b)\xi},$$

where $\mathbf{S} = (0, \infty)$, then the performance index $\bar{V}$ takes the form (16). In addition, observe that $\tilde{\alpha}$ is continuous on $\mathbb{K} \times \mathbf{S}$, and for all (x, a, b),

$$\alpha_\theta(x, a, b) = \lambda \int_0^\infty \mathrm{e}^{-\gamma(x,a,b)s} \mathrm{e}^{-\lambda s} \mathrm{d}s = \frac{\lambda}{\lambda + \gamma(x, a, b)} < \frac{\lambda}{\lambda + d}.$$

Thus

$$\alpha^* < \frac{\lambda}{\lambda + d} < 1. \tag{30}$$

To the best of our knowledge, semi-Markov models with state-action-dependent discount factors have been considered only for decision processes in [15,16].

We conclude by presenting some insights into the fulfillment of continuity and W-growth conditions imposed in Assumption 1. Such conditions are standard in the literature (see, e.g., [14,18,20,24–26]) and satisfied by several zero-sum game models.

As stated in [14, Appendix C], Assumption 1(c) holds if the transition kernel Q on $\mathbf{X}$ given $\mathbb{K}$ has a continuous density $q(y|x, a, b)$ in $(x, a, b) \in \mathbb{K}$ with respect to a σ-finite measure m on X, that is

$$Q(X|x, a, b) = \int_X q(y|x, a, b) m(\mathrm{d}y), \quad X \in \mathcal{B}(\mathbf{X}), \ (x, a, b) \in \mathbb{K}.$$

Furthermore, Assumption 1(c) also holds for games that evolve on $\mathbf{X} = \mathbb{R}$ according to noise-additive difference equations of the form

$$x_{n+1} = G(x_n, a_n, b_n) + w_n, \quad n \in \mathbb{N}_0,$$

with $\mathbf{A} = \mathbf{B} = \mathbb{R}$, where G is a continuous function and $\{w_n\}$ is a sequence of i.i.d. random variables with continuous density g on $\mathbb{R}$. In this case the kernel Q takes the form

$$Q(X|x, a, b) = \int_{\mathbb{R}} I_X[G(x, a, b) + w] \, g(w) dw,$$

where $I_X(\cdot)$ stands for the indicator function of the set $X \in \mathcal{B}(\mathbf{X})$.

In general, the conditions related to the weighted function W are easier to illustrate in difference-equation game models as we show in the following examples.

Example 4 (A linear quadratic game (see [7])) Consider a game whose dynamics is defined by the linear equation

$$x_{n+1} = x_n + a_n + b_n + w_n, \quad n \in \mathbb{N}_0,$$

where $\mathbf{X} = \mathbf{A} = \mathbf{B} = \mathbb{R}$ and $\{w_n\}$ is a sequence of i.i.d. random variables with standard normal distribution

$$g(w) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right), \quad w \in \mathbb{R}.$$

We assume that the admissible action sets are $A(x) = B(x) = [-|x|/2, |x|/2]$. In addition, the one-stage payoff r is a quadratic function such that

$$|r(x, a, b)| \leq r_0(x^2 + 1),$$

for some positive constant r_0.

Since the dynamics is defined by a continuous noise-additive function and g is a continuous density, from [14, Appendix C], Assumption 1(c) holds. Moreover, if we take $W(x) := x^2+1$, by applying the same arguments, the continuity of the functions defined in (20) follows.

On the other hand, for all $(x, a, b) \in \mathbb{K}$,

$$\begin{aligned}\int_{\mathbb{R}} W(y)Q(\mathrm{d}y|x, a, b) &= \int_{\mathbb{R}} [(x + a + b + w)^2 + 1]\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw \\ &= \int_{\mathbb{R}} (x + a + b + w)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw + 1 \\ &= \int_{\mathbb{R}} y^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - (x + a + b))^2}{2}\right) \mathrm{d}y + 1 \\ &= (x + a + b)^2 + 2 \leq 4x^2 + 2 \leq 4W(x).\end{aligned}$$

Hence, Assumption 1 (d) and (e) are satisfied with $\beta = 4$ and any continuous function such that $\tilde{\alpha}(x, a, b, s) < \frac{1}{4}$.

Example 5 (A semi-Markov storage system) We consider a storage system with controlled input/output, whose evolution is as follows. At time T_n when an amount of certain product $M > 0$ accumulates for admission in the system, player 1 selects an action $a \in [a_*, 1] =: \mathbf{A}$, $a_* \in (0, 1)$, representing the portion of M to be admitted. In addition, there is a continuous consumption of the admitted product which is controlled by the player 2 by selecting $b \in [b_*, b^*] =: \mathbf{B}$ $(0 < b_* < b^*)$ which represents the consumption rate per unit time. Thus, if $x_n \in \mathbf{X} := [0, \infty)$ is the stock level, a_n and b_n are the decisions of players 1 and 2, respectively, at the time of the nth decision epoch T_n, the process $\{x_n\}$ can be modeled as a semi-Markov game evolving according to the equation

$$x_{n+1} = (x_n + a_n M - b_n \xi_{n+1})^+$$

with holding times $\xi_{n+1} := T_{n+1} - T_n$. In the context of Example 3, we suppose that $\{\xi_n\}$ is a sequence of i.i.d. random variables, exponentially distributed with parameter $\lambda > 0$. Moreover, the discount factor is a continuous function $\gamma : \mathbb{K} \to (d, \infty)$ where $d > 0$. It is reasonable to assume that

$$b_* E(\xi) < b^* E(\xi) = \frac{b^*}{\lambda} < M. \tag{31}$$

Let Ψ be the moment generating function of the random variable $M - b_*\xi$, that is:

$$\Psi(t) = E[\exp(t(M - b_*\xi))] = \frac{\lambda \exp(Mt)}{b_* t + \lambda}.$$

Then, computing the derivative Ψ' and using the fact that $b_* < M\lambda$ (see (31)), it is easy to prove that $\Psi'(t) > 0, t > 0$. Moreover, taking the constant $d > \lambda$, from the continuity of Ψ and because $\Psi(0) = 1$, there exists $\lambda^* > 0$ such that

$$\beta_0 := \Psi(\lambda^*) = \frac{d}{\lambda}. \tag{32}$$

Now we assume that the one-stage payoff r is an arbitrary function satisfying Assumption 1(b) such that

$$|r(x,a,b)| \le r_0 \mathrm{e}^{\lambda^* x},$$

for some constant $r_0 > 0$. Hence, defining $W(x) := \mathrm{e}^{\lambda^* x}$, relation (19) is satisfied. Moreover, for $(x,a,b) \in \mathbb{K}$,

$$\begin{aligned}\int_{\mathbf{X}} \mathrm{e}^{\lambda^* y} Q(\mathrm{d}y|x,a,b) &= \int_0^\infty \mathrm{e}^{\lambda^*(x+aM-bs)^+} \lambda \mathrm{e}^{-\lambda s}\mathrm{d}s \\ &= P[x+aM-b\xi \le 0] + \mathrm{e}^{\lambda^* x}\int_0^\infty \mathrm{e}^{\lambda^*(M-bs)}\lambda \mathrm{e}^{-\lambda s}\mathrm{d}s \\ &\le 1 + W(x)E[\mathrm{e}^{\lambda^*(M-b_*\xi)}] \\ &\le (\beta_0+1)W(x).\end{aligned}$$

Hence, combining (30) and (32), we obtain

$$1 < \beta_0 + 1 = \frac{d}{\lambda} + 1 = \frac{\lambda + d}{\lambda} < (\alpha^*)^{-1},$$

and defining $\beta := \beta_0 + \bar{r}$, Assumptions 1(d), (e) are satisfied.

Finally, to verify Assumption 1(c), let u be a bounded measurable function on $\mathbf{X}$ and $\rho_{(a,b)}$ be the density of the random variable $aM - b\delta$, for every fixed $a \in \mathbf{A}$ and $b \in \mathbf{B}$. Observe that

$$\rho_{(a,b)}(y) = \frac{1}{b}\lambda \mathrm{e}^{-\lambda\left(\frac{aM-y}{b}\right)}, \quad -\infty < y \le aM,$$

and therefore, for each $y \in \mathbf{R}$, $(a,b) \longmapsto \rho_{(a,b)}(y)$ is continuous function on $\mathbf{A}\times\mathbf{B}$. Hence,

$$\begin{aligned}\int_X u(y)Q(\mathrm{d}y \mid x,a,b) &= \int_0^\infty u[(x+y)^+]\rho_{(a,b)}(y)\mathrm{d}y \\ &= u(0)\int_{-\infty}^{-x}\rho_{(a,b)}(y)\mathrm{d}y + \int_{-x}^{\infty} u(x+y)\rho_{(a,b)}(y)\mathrm{d}y \\ &= u(0)\int_{-\infty}^{-x}\rho_{(a,b)}(y)\mathrm{d}y + \int_0^\infty u(y)\rho_{(a,b)}(y-x)\mathrm{d}y.\end{aligned}$$

Thus by Scheffé's Theorem,

$$(a,b) \longmapsto \int_X u(y)Q(\mathrm{d}y \mid x,a,b)$$

defines a continuous function on $\mathbf{A}\times\mathbf{B}$, which proves that Assumption 1(c) holds. Similarly is shown the continuity of the functions in (20).

6 Proofs

6.1 Proof of Proposition 1

The proof is a consequence of the following facts.

Let us fix $\varphi^2 \in \Pi_S^2$. Define the stochastic kernel Q_{φ^2} on $\mathbf{X}$ given $\mathbb{K}_\mathbf{A}$ as

$$Q_{\varphi^2}(X|x,a) := \int_\mathbf{B} Q(X|x,a,b)\varphi^2(db|x), \quad X \in \mathcal{B}(\mathbf{X}), \tag{33}$$

$r_{\varphi^2} : \mathbb{K}_\mathbf{A} \longmapsto \mathbb{R}$ and $\tilde{\alpha}_{\varphi^2} : \mathbb{K}_\mathbf{A} \times \mathbf{S} \to (0,1)$ are the measurable functions defined as

$$r_{\varphi^2}(x,a) := \int_\mathbf{B} r(x,a,b)\varphi^2(db|x), \tag{34}$$

$$\tilde{\alpha}_{\varphi^2}(x,a,s) := \int_\mathbf{B} \tilde{\alpha}(x,a,b,s)\varphi^2(db|x). \tag{35}$$

In addition, let $\pi^1 \in \Pi^1$ be an arbitrary strategy, and for $x \in \mathbf{X}$, we define the performance index

$$\tilde{V}_{\varphi^2}(x,\pi^1) := E_x^{\pi^1}\left[\sum_{n=0}^{\infty} \tilde{\Gamma}_n^{\varphi^2} r_{\varphi^2}(x_n,a_n)\right], \tag{36}$$

where

$$\tilde{\Gamma}_n^{\varphi^2} = \prod_{k=0}^{n-1} \tilde{\alpha}_{\varphi^2}(x_k,a_k,\xi_{k+1}), \quad \tilde{\Gamma}_0^{\varphi^2} = 1,$$

and $E_x^{\pi^1}$ is the expectation operator with respect to the probability measure $P_x^{\pi^1} \equiv P_x^{\pi^1,\varphi^2}$ induced by $(\pi^1,\varphi^2) \in \Pi^1 \times \Pi_S^2$ and $x_0 = x$. Then, from (7)–(10), $P_x^{\pi^1}$ satisfies the following properties:

$$P_x^{\pi^1}[x_0 \in X] = \delta_x(X), \quad X \in \mathcal{B}(\mathbf{X}); \tag{37}$$

$$\begin{aligned} P_x^{\pi^1}[a_n \in A|h_n] &= P_x^{\pi^1}[a_n \in A, b_n \in \mathbf{B}|h_n] \\ &= \pi_n^1(A|h_n)\,\varphi_n^2(\mathbf{B}|x_n) \\ &= \pi_n^1(A|h_n), \quad A \in \mathcal{B}(\mathbf{A}); \end{aligned} \tag{38}$$

$$P_x^{\pi^1}\left[x_{n+1} \in X|h_n,a_n,b_n,\xi_{n+1}\right] = Q_{\varphi^2}(X|x_n,a_n), \quad X \in \mathcal{B}(\mathbf{X}); \tag{39}$$

$$P_x^{\pi^1}\left[\xi_{n+1} \in S|h_n,a_n,b_n\right] = \theta(S), \quad S \in \mathcal{B}(\mathbf{S}). \tag{40}$$

Similarly, for a fixed $\varphi^1 \in \Pi_S^1$, define Q_{φ^1}, r_{φ^1}, $\tilde{\alpha}_{\varphi^1}$ and the performance index

$$\tilde{V}_{\varphi^1}(x,\pi^2) := E_x^{\pi^2}\left[\sum_{n=0}^{\infty} \tilde{\Gamma}_n^{\varphi^1} r_{\varphi^1}(x_n,b_n)\right], \quad \pi^2 \in \Pi^2, \ x \in \mathbf{X}, \tag{41}$$

where

$$\tilde{\Gamma}_n^{\varphi^1} = \prod_{k=0}^{n-1} \tilde{\alpha}_{\varphi^1}(x_k,b_k,\xi_{k+1}), \quad \tilde{\Gamma}_0^{\varphi^1} = 1.$$

The next result is an adaptation of [23, Lemma 15] to our context. The proof follows by applying similar arguments and making the appropriate changes.

Lemma 1 *For each $x \in \mathbf{X}$, $\varphi^2 \in \Pi_S^2$, and $\pi^1 \in \Pi^1$ there exists $\varphi^1 \in \Pi_M^1$ such that*

$$\tilde{V}_{\varphi^2}(x,\pi^1) = \tilde{V}_{\varphi^2}(x,\varphi^1). \tag{42}$$

Remark 1 Let us fix $\varphi^1 \in \Pi_S^1$. Then we can also prove that for each $\pi^2 \in \Pi^2$ there exists $\varphi^2 \in \Pi_M^2$ such that

$$\tilde{V}_{\varphi^1}(x, \pi^2) = \tilde{V}_{\varphi^1}(x, \varphi^2), \quad x \in \mathbf{X}, \tag{43}$$

where $\tilde{V}_{\varphi^1}$ is the performance index defined in (41).

Lemma 2 *(a) For each $\pi^1 \in \Pi^1$ and $\varphi^2 \in \Pi_S^2$, there exists $\varphi^1 \in \Pi_M^1$ such that*

$$\tilde{V}(x, \pi^1, \varphi^2) = \tilde{V}(x, \varphi^1, \varphi^2), \quad x \in \mathbf{X}. \tag{44}$$

(b) For each $\pi^2 \in \Pi^2$ and $\varphi^1 \in \Pi_S^1$, there exists $\varphi^2 \in \Pi_M^2$ such that

$$\tilde{V}(x, \varphi^1, \pi^2) = \tilde{V}(x, \varphi^1, \varphi^2), \quad x \in \mathbf{X}. \tag{45}$$

Proof Let $\pi^1 \in \Pi^1$ and $\varphi^2 \in \Pi_S^2$ be arbitrary strategies and consider the corresponding performance index $\tilde{V}_{\varphi^2}(x, \pi^1)$, $x \in \mathbf{X}$. From Lemma 1, there exists $\varphi^1 \in \Pi_M^1$ such that $\tilde{V}_{\varphi^2}(x, \pi^1) = \tilde{V}_{\varphi^2}(x, \varphi^1)$, $x \in \mathbf{X}$. Hence, to obtain (44), it is enough to prove

$$\tilde{V}_{\varphi^2}(x, \pi^1) = \tilde{V}(x, \pi^1, \varphi^2), \quad x \in \mathbf{X}, \tag{46}$$

which is obtained by comparing the corresponding terms in the sums (36) and (11).

Indeed, for the first term, from (34)

$$\begin{aligned} E_x^{\pi^1} r_{\varphi^2}(x_0, a_0) &= \int_{\mathbf{A}} r_{\varphi^2}(x, a_0) \pi_0^1(da_0|x) \\ &= \int_{\mathbf{A}} \int_{\mathbf{B}} r(x, a_0, b_0) \varphi_0^2(db_0|x) \pi_0^1(da_0|x) \\ &= E_x^{\pi^1, \varphi^2} r(x_0, a_0, b_0). \end{aligned}$$

Furthermore, from (35)

$$\begin{aligned} E_x^{\pi^1} \tilde{\Gamma}_1^{\varphi^2} r_{\varphi^2}(x_1, a_1) &= E_x^{\pi^1} \tilde{\alpha}_{\varphi^2}(x_0, a_0, \xi_1) r_{\varphi^2}(x_1, a_1) \\ &= \int_{\mathbf{A}\times\mathbf{S}\times\mathbf{X}\times\mathbf{A}} \tilde{\alpha}_{\varphi^2}(x, a_0, \xi_1) r_{\varphi^2}(x_1, a_1) \pi_1^1(da_1|h_1) \\ &\quad Q_{\varphi^2}(\mathrm{d}x_1|x_0, a_0) \theta(\mathrm{d}\xi_1) \pi_0^1(da_0|x) \\ &= \int_{\mathbf{A}\times\mathbf{B}\times\mathbf{S}\times\mathbf{X}\times\mathbf{A}\times\mathbf{B}} \tilde{\alpha}(x, a_0, b_0, \xi_1) r(x_1, a_1, b_1) \\ &\quad \varphi_1^2(db_1|x) \pi_1^1(da_1|h_1) Q(\mathrm{d}x_1|x_0, a_0, b_0) \theta(\mathrm{d}\xi_1) \varphi_0^2(db_0|x) \pi_0^1(da_0|x) \\ &= E_x^{\pi^1, \varphi^2} \tilde{\alpha}(x_0, a_0, b_0, \xi_1) r(x_1, a_1, b_1) \\ &= E_x^{\pi^1, \varphi^2} \tilde{\Gamma}_1 r(x_1, a_1, b_1). \end{aligned}$$

An induction argument shows that

$$E_x^{\pi^1} \tilde{\Gamma}_n^{\varphi^2} r_{\varphi^2}(x_n, a_n) = E_x^{\pi^1, \varphi^2} \tilde{\Gamma}_n r(x_n, a_n, b_n), \quad \forall n \in \mathbb{N}_0.$$

Hence, from (36) and (11), we obtain (46).

Part (b) is proved similarly. □

Proof of Proposition 1 Let $(\varphi^1_*, \varphi^2_*) \in \Pi^1_S \times \Pi^2_S$ be a pair that satisfies (13). From Lemma 2, we have, for each $\varphi^2 \in \Pi^2_S$,

$$\max_{\pi^1 \in \Pi^1} \tilde{V}(x, \pi^1, \varphi^2) = \max_{\varphi^1 \in \Pi^1_M} \tilde{V}(x, \varphi^1, \varphi^2), \quad x \in \mathbf{X}, \tag{47}$$

and for each $\varphi^1 \in \Pi^1_S$

$$\min_{\pi^2 \in \Pi^2} \tilde{V}(x, \varphi^1, \pi^2) = \min_{\varphi^2 \in \Pi^2_M} \tilde{V}(x, \varphi^1, \varphi^2), \quad x \in \mathbf{X}. \tag{48}$$

Now, from (13) and (47)

$$\begin{aligned} \tilde{V}(x, \varphi^1_*, \varphi^2_*) &\geq \max_{\varphi^1 \in \Pi^1_M} \tilde{V}(x, \varphi^1, \varphi^2_*) \\ &= \max_{\pi^1 \in \Pi^1} \tilde{V}(x, \pi^1, \varphi^2_*) \\ &\geq \tilde{V}(x, \pi^1, \varphi^2_*), \quad \forall \pi^1 \in \Pi^1, \ x \in \mathbf{X}. \end{aligned} \tag{49}$$

Similarly, from (13) and (48)

$$\begin{aligned} \tilde{V}(x, \varphi^1_*, \varphi^2_*) &\leq \min_{\varphi^2 \in \Pi^2_M} \tilde{V}(x, \varphi^1_*, \varphi^2) \\ &= \min_{\pi^2 \in \Pi^2} \tilde{V}(x, \varphi^1_*, \pi^2) \\ &\leq \tilde{V}(x, \varphi^1_*, \pi^2), \quad \forall \pi^2 \in \Pi^2, \ x \in \mathbf{X}. \end{aligned} \tag{50}$$

Therefore, (49) and (50) yield the desired inequality (12). This completes the proof of the proposition. □

6.2 Proof of Proposition 2

Proof The proof follows by applying similar arguments as those in the proof of Lemma 2. For instance, observe that for each $x \in \mathbf{X}$ and $(\varphi^1, \varphi^2) \in \Pi^1_M \times \Pi^2_M$, from (14)

$$\begin{aligned} E^{\varphi^1,\varphi^2}_x \tilde{\Gamma}_1 r(x_1, a_1, b_1) &= E^{\varphi^1,\varphi^2}_x \tilde{\alpha}(x_0, a_0, b_0, \xi_1) r(x_1, a_1, b_1) \\ &= \int_{\mathbf{A}\times\mathbf{B}\times\mathbf{S}\times\mathbf{X}\times\mathbf{A}\times\mathbf{B}} \tilde{\alpha}(x_0, a_0, b_0, \xi_1) r(x_1, a_1, b_1) \\ &\qquad \varphi^2_1(db_1|x_1)\varphi^1_1(da_1|x_1) Q(\mathrm{d}x_1|x_0, a_0, b_0)\theta(\mathrm{d}\xi_1)\varphi^2_0(db_0|x)\varphi^1_0(da_0|x) \\ &= \int_{\mathbf{A}\times\mathbf{B}} \int_{\mathbf{S}} \tilde{\alpha}(x_0, a_0, b_0, \xi_1)\theta(\mathrm{d}\xi_1) \int_{\mathbf{X}} \int_{\mathbf{A}\times\mathbf{B}} r(x_1, a_1, b_1) \\ &\qquad \varphi^2_1(db_1|x_1)\varphi^1_1(da_1|x_1) Q(\mathrm{d}x_1|x_0, a_0, b_0)\varphi^2_0(db_0|x)\varphi^1_0(da_0|x) \\ &= E^{\varphi^1,\varphi^2}_x \alpha_\theta(x_0, a_0, b_0) r(x_1, a_1, b_1) \\ &= E^{\varphi^1,\varphi^2}_x \Gamma_1 r(x_1, a_1, b_1). \end{aligned}$$

It is shown, by induction, that

$$E^{\varphi^1,\varphi^2}_x \tilde{\Gamma}_n r(x_n, a_n, b_n) = E^{\varphi^1,\varphi^2}_x \Gamma_n r(x_n, a_n, b_n), \quad \forall n \in \mathbb{N}_0.$$

Therefore, from (11) and (16), we get (17). □

6.3 Proof of Theorem 1

Before presenting the proof, we establish some important facts on minimax theorems and the W-norm, as well as the Shapley operator (23). All these facts are summarized in the following remark.

Remark 2 (a) Provided that Assumption 1 holds, for $u \in \mathbb{B}_W$ and $(x, a, b) \in \mathbb{K}$, $\hat{T}(u, x, \cdot, b)$ is usc on $A(x)$ and $\hat{T}(u, x, a, \cdot)$ is lsc on $B(x)$. Hence, by applying well-known properties of weak convergence of measures on the sets $\mathbb{A}(x)$ and $\mathbb{B}(x)$ (see, e.g., Theorem 2.8.1 in [1]), we can prove that the function $\hat{T}(u, x, \cdot, \varphi^2)$ is usc on $\mathbb{A}(x)$ while $\hat{T}(u, x, \varphi^1, \cdot)$ is lsc on $\mathbb{B}(x)$. In addition, since $\hat{T}(u, x, \varphi^1, \varphi^2)$ is concave in φ^1 and convex in φ^2, the well-known Fan's Minimax Theorem implies that we can interchange inf and sup in (23), i.e.,

$$Tu(x) = \sup_{\varphi^1 \in \mathbb{A}(x)} \inf_{\varphi^2 \in \mathbb{B}(x)} \hat{T}(u, x, \varphi^1(x), \varphi^2(x)), \quad x \in X. \tag{51}$$

(b) Moreover, suitable measurable selection theorems yield the existence of $\varphi^1_* \in \mathbb{A}(x)$ and $\varphi^2_* \in \mathbb{B}(x)$ such that (see, e.g., Lemma 4.3 in [29])

$$\begin{aligned} Tu(x) &= \hat{T}(u, x, \varphi^1_*(x), \varphi^2_*(x)) \\ &= \max_{\varphi^1 \in \mathbb{A}(x)} \hat{T}(u, x, \varphi^1, \varphi^2_*) = \min_{\varphi^2 \in \mathbb{B}(x)} \hat{T}(u, x, \varphi^1_*, \varphi^2), \quad x \in \mathbf{X}. \end{aligned}$$

(c) For $u, v \in \mathbb{B}_W$, (21)–(23) and properties of the W-norm imply

$$\begin{aligned} |Tu(x) - Tv(x)| &\le \sup_{a \in A(x)} \sup_{b \in B(x)} \alpha_\theta(x, a, b) \int_{\mathbf{X}} |u(y) - v(y)| \, Q(\mathrm{d}y \mid x, a, b) \\ &\le \alpha^* \|u - v\|_W \sup_{a \in A(x)} \sup_{b \in B(x)} \int_{\mathbb{X}} W(y) Q(\mathrm{d}y \mid x, a, b) \\ &\le \alpha^* \beta \|u - v\|_W W(x), \end{aligned}$$

which in turn yields

$$\|Tu - Tv\|_W \le \alpha^* \beta \|u - v\|_W .$$

Hence, T is a contraction operator on $\mathbb{B}_W$ with modulus $\alpha^* \beta < 1$. Similarly, the operator

$$T_{\varphi^1 \varphi^2} u(x) := \hat{T}(u, x, \varphi^1(x), \varphi^2(x)), \quad x \in \mathbf{X}, \tag{52}$$

defined for a pair of stationary strategies $(\varphi^1, \varphi^2) \in \Pi^1_S \times \Pi^2_S$, is a contraction operator on $\mathbb{B}_W$ with modulus $\alpha^* \beta$.

(d) Thus, there exist unique fixed points v and $v_{\varphi^1 \varphi^2}$ in $\mathbb{B}_W$ of operators T and $T_{\varphi^1 \varphi^2}$, respectively, that is

$$Tv(x) = v(x) \quad \text{and} \quad T_{\varphi^1 \varphi^2} v_{\varphi^1 \varphi^2}(x) = v_{\varphi^1 \varphi^2}(x), \quad x \in \mathbf{X}. \tag{53}$$

(e) Finally, we also apply the following properties of the weighted function W. From (22), for each $x \in \mathbf{X}$, $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, and $n \in \mathbb{N}_0$,

$$E_x^{\pi^1, \pi^2} \left[W(x_{n+1}) \right] \le \beta E_x^{\pi^1, \pi^2} \left[W(x_n) \right].$$

Iteration of this inequality yields

$$E_x^{\pi^1,\pi^2}\left[W\left(x_{n+1}\right)\right] \leq \beta^{n+1} W\left(x\right), \quad x \in \mathbf{X},\ n \in \mathbb{N}_0. \tag{54}$$

Furthermore, from (54) and (15), for each $u \in \mathbb{B}_W$, $x \in \mathbf{X}$, $\left(\pi^1, \pi^2\right) \in \Pi^1 \times \Pi^2$, and $n \in \mathbb{N}_0$,

$$\begin{aligned}\left|E_x^{\pi^1,\pi^2} \Gamma_n u(x_n)\right| &\leq \left(\alpha^*\right)^n \|u\|_W\, E_x^{\pi^1,\pi^2}\left[W\left(x_n\right)\right] \\ &\leq \left(\beta\alpha^*\right)^n \|u\|_W\, W\left(x\right).\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} E_x^{\pi^1,\pi^2} \Gamma_n u(x_n) = 0, \quad x \in \mathbf{X},\ \left(\pi^1, \pi^2\right) \in \Pi^1 \times \Pi^2. \tag{55}$$

Proof of Theorem 1 From (23) and (51)

$$\begin{aligned}v(x) = Tv(x) &= \sup_{\varphi^1 \in \mathbb{A}(x)} \inf_{\varphi^2 \in \mathbb{B}(x)} \hat{T}(v, x, \varphi^1(x), \varphi^2(x)) \\ &= \inf_{\varphi^2 \in \mathbb{B}(x)} \sup_{\varphi^1 \in \mathbb{A}(x)} \hat{T}(v, x, \varphi^1(x), \varphi^2(x)), \quad x \in \mathbf{X},\end{aligned}$$

where v is the fixed point of T (see Remark 2 (d)). In addition, from Remark 2 (b), there exists a pair of stationary strategies $(\varphi_*^1, \varphi_*^2) \in \Pi_S^1 \times \Pi_S^2$ such that

$$v(x) = \hat{T}(v, x, \varphi_*^1(x), \varphi_*^2(x)) = T_{\varphi_*^1\varphi_*^2} v(x) \tag{56}$$

$$= \max_{\varphi^1 \in \mathbb{A}(x)} \hat{T}(v, x, \varphi^1(x), \varphi_*^2(x)) \tag{57}$$

$$= \min_{\varphi^2 \in \mathbb{B}(x)} \hat{T}(u, x, \varphi_*^1(x), \varphi^2(x)), \quad x \in \mathbf{X}. \tag{58}$$

On the other hand, $V(\cdot, \varphi_*^1, \varphi_*^2)$ is the unique fixed point of $T_{\varphi_*^1\varphi_*^2}$ belonging to $\mathbb{B}_W$, i.e.,

$$v_{\varphi_*^1\varphi_*^2}(\cdot) = V(\cdot, \varphi_*^1, \varphi_*^2). \tag{59}$$

Indeed, from (53), (24), and (52)

$$v_{\varphi_*^1\varphi_*^2}(x) = \int_B \int_A \left[r(x, a, b) + \alpha_\theta(x, a, b) \int_{\mathbf{X}} v_{\varphi_*^1\varphi_*^2}(y) Q(\mathrm{d}y|x, a, b)\right] \varphi_*^1(da|x)\varphi_*^2(db|x),$$

for every x in $\mathbf{X}$. Iterating this equation, we obtain

$$v_{\varphi_*^1\varphi_*^2}(x) = E_x^{\varphi_*^1,\varphi_*^2} \sum_{n=0}^{m-1} \Gamma_n r(x_n, a_n, b_n) + E_x^{\varphi_*^1,\varphi_*^2} \Gamma_m v_{\varphi_*^1\varphi_*^2}(x_m).$$

Now, letting $m \to \infty$, from (55) and (16), we obtain (59).

Since $V(\cdot, \varphi_*^1, \varphi_*^2)$ is the unique fixed point of $T_{\varphi_*^1\varphi_*^2}$, (56) implies that $v(x) = V(x, \varphi_*^1, \varphi_*^2)$, $x \in \mathbf{X}$. Therefore, considering (57) and (58), Theorem 1 will be proved if we show that

$$V(x, \varphi^1, \varphi_*^2) \leq V(x, \varphi_*^1, \varphi_*^2) \leq V(x, \varphi_*^1, \varphi^2), \quad \forall (\varphi^1, \varphi^2) \in \Pi_M^1 \times \Pi_M^2,\ x \in \mathbf{X}. \tag{60}$$

To prove the first inequality in (60), let $\varphi^1 \in \Pi_M^1$ be an arbitrary Markov strategy for player 1. Then, for all $n \in \mathbb{N}$,

$$
\begin{aligned}
&E_x^{\varphi^1,\varphi_*^2}\left[\Gamma_{n+1}V(x_{n+1},\varphi_*^1,\varphi_*^2)|h_n,a_n,b_n\right] = \Gamma_{n+1}E_x^{\varphi^1,\varphi_*^2}\left[V(x_{n+1},\varphi_*^1,\varphi_*^2)|h_n,a_n,b_n\right]\\
&\quad = \Gamma_{n+1}\int_{\mathbf{X}} V(y,\varphi_*^1,\varphi_*^2)Q(\mathrm{d}y|x_n,\varphi_n^1,\varphi_*^2)\\
&\quad = \Gamma_n\left\{\alpha_\theta(x_n,\varphi_n^1,\varphi_*^2)\int_{\mathbf{X}} V(y,\varphi_*^1,\varphi_*^2)Q(\mathrm{d}y|x_n,\varphi_n^1,\varphi_*^2)\right.\\
&\qquad \left. + r(x_n,\varphi_n^1,\varphi_*^2) - r(x_n,\varphi_n^1,\varphi_*^2)\right\}\\
&\quad \le \Gamma_n\left\{\sup_{\varphi^1\in\mathbb{B}(x)}\hat{T}(v,x_n,\varphi^1(x_n),\varphi_*^2(x_n)) - r(x_n,\varphi_n^1,\varphi_*^2)\right\}\\
&\quad = \Gamma_n\left\{v(x_n) - r(x_n,\varphi_n^1,\varphi_*^2)\right\}\\
&\quad = \Gamma_n\left\{V(x_n,\varphi_*^1,\varphi_*^2) - r(x_n,\varphi_n^1,\varphi_*^2)\right\},
\end{aligned}
\tag{61}
$$

where the last two equalities come from (56) and (57). Now, from (61), for all $n \in \mathbb{N}$,

$$
\Gamma_n V(x_n,\varphi_*^1,\varphi_*^2) - E_x^{\varphi^1,\varphi_*^2}\left[\Gamma_{n+1}V(x_{n+1},\varphi_*^1,\varphi_*^2)|h_n,a_n,b_n\right] \ge \Gamma_n r(x_n,\varphi_n^1,\varphi_*^2),
$$

which, by taking expectation $E_x^{\varphi^1,\varphi_*^2}$ and adding over $n = 0, 1, \ldots, m-1$, $m > 0$, implies

$$
V(x,\varphi_*^1,\varphi_*^2) - E_x^{\varphi^1,\varphi_*^2}\left[\Gamma_{m+1}V(x_{m+1},\varphi_*^1,\varphi_*^2)\right] \ge E_x^{\varphi^1,\varphi_*^2}\sum_{n=0}^{m-1}\Gamma_n r(x_n,a_n,b_n).
$$

Letting $m \to \infty$, from (16) and (55), we get

$$
V(x,\varphi_*^1,\varphi_*^2) \ge V(x,\varphi^1,\varphi_*^2), \quad x \in \mathbf{X},
$$

that is, the first inequality in (60) holds. The second inequality is proved similarly. Hence, the proof of Theorem 1 is completed. □

References

1. Ash RB, Doléans-Dade C (2000) Probability and measure theory, 2nd edn. Academic Press, London
2. Bäuerle N, Rieder U (2011) Markov decision processes with applications to finance. Springer, Berlin
3. Carmon Y, Shwartz A (2009) Markov decision processes with exponentially representable discounting. Oper Res Lett 37(1):51–55
4. Cruz-Suárez H, Ilhuicatzi-Roldán R, Montes-de Oca R (2014) Markov decision processes on Borel spaces with total cost and random horizon. J Optim Theory Appl 162(1):329–346
5. Dutta PK, Sundaram R (1992) Markovian equilibrium in a class of stochastic games: existence theorems for discounted and undiscounted models. Econ Theory 2(2):197–214
6. Dynkin EB, Yushkevich AA (1979) Controlled Markov processes. Springer, Berlin
7. Engwerda J (2005) LQ dynamic optimization and differential games. Wiley, New York
8. Feinberg EA, Shwartz A (1999) Constrained dynamic programming with two discount factors: applications and an algorithm. IEEE Trans Autom Control 44(3):628–631
9. Filar J, Vrieze K (2012) Competitive Markov decision processes. Springer, Berlin

10. González-Hernández J, López-Martínez RR, Minjárez-Sosa JA (2008) Adaptive policies for stochastic systems under a randomized discounted cost criterion. Bol Soc Mat Mexicana 3(14):149–163
11. 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
12. González-Hernández J, López-Martínez RR, Minjárez-Sosa JA, Gabriel-Arguelles JR (2013) Constrained Markov control processes with randomized discounted cost criteria: occupation measures and extremal points. Risk Decis Anal 4(3):163–176
13. He W, Sun Y (2017) Stationary Markov perfect equilibria in discounted stochastic games. J Econ Theory 169:35–61
14. Hernández-Lerma O, Lasserre JB (1996) Discrete-time Markov control processes: basic optimality criteria, vol 30. Springer, Berlin
15. Huang Y, Guo X (2012) Constrained optimality for first passage criteria in semi-Markov decision processes. In: Hernández-Hernández D, Minjárez-Sosa JA (eds) Optimization, control, and applications of stochastic systems, systems & control: foundations & applications. Birkhauser, Boston, pp 181–202 chap. 11
16. Huang Y, Wei Q, Guo X (2013) Constrained Markov decision processes with first passage criteria. Ann Oper Res 206(1):197–219
17. Jaśkiewicz A, Nowak AS (2006) Approximation of noncooperative semi-Markov games. J Optim Theory Appl 131(1):115–134
18. Jaśkiewicz A, Nowak AS (2006) Zero-sum ergodic stochastic games with Feller transition probabilities. SIAM J Control Optim 45(3):773–789
19. Krausz A, Rieder U (1997) Markov games with incomplete information. Math Methods Oper Res 46(2):263–279
20. Luque-Vásquez F (2002) Zero-sum semi-Markov games in Borel spaces: discounted and average payoff. Bol Soc Mat Mexicana 8:227–241
21. Maitra A, Parthasarathy T (1970) On stochastic games. J Optim Theory Appl 5(4):289–300
22. Maschler M, Solan E, Zamir S (2013) Game theory. Cambridge University Press, Cambridge
23. Minjárez-Sosa JA (2015) Markov control models with unknown random state-action-dependent discount factors. Top 23(3):743–772
24. 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
25. Minjárez-Sosa JA, Vega-Amaya O (2009) Asymptotically optimal strategies for adaptive zero-sum discounted Markov games. SIAM J Control Optim 48(3):1405–1421
26. Minjárez-Sosa JA, Vega-Amaya O (2013) Optimal strategies for adaptive zero-sum average Markov games. J Math Anal Appl 402(1):44–56
27. Neyman A, Sorin S (2003) Stochastic games and applications, vol 570. Kluwer, Dordrecht
28. Nowak AS (1984) On zero-sum stochastic games with general state space. I. Prob Math Stat 4(1):13–32
29. Nowak AS (1985) Measurable selection theorems for minimax stochastic optimization problems. SIAM J Control Optim 23(3):466–476
30. Nowak AS (1987) Nonrandomized strategy equilibria in noncooperative stochastic games with additive transition and reward structure. J Optim Theory Appl 52(3):429–441
31. Nowak AS, Szajowski K (1999) Nonzero-sum stochastic games. In: Stochastic and differential games. Annals of the international society of dynamic games, vol 4, chap 7. Springer, Berlin, pp 297–342
32. Osborne MJ, Rubinstein A (1994) A course in game theory. MIT Press, Cambridge
33. Rieder U (1991) Non-cooperative dynamic games with general utility functions. In: Raghavan TES, Ferguson TS, Parthasarathy T, Vrieze OJ (eds) Stochastic games and related topics, theory and decision library, vol 7. Springer, Berlin, pp 161–174
34. Schäl M (1975) Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Probab Theory Rel Fields 32(3):179–196
35. Shapley LS (1953) Stochastic games. Proc Natl Acad Sci USA 39(10):1095–1100
36. Shimkin N, Shwartz A (1995) Asymptotically efficient adaptive strategies in repeated games. Part I: certainty equivalence strategies. Math Oper Res 20(3):743–767
37. Shimkin N, Shwartz A (1996) Asymptotically efficient adaptive strategies in repeated games. Part II: asymptotic optimality. Math Oper Res 21(2):487–512
38. Wei Q, Guo X (2011) Markov decision processes with state-dependent discount factors and unbounded rewards/costs. Oper Res Lett 39(5):369–374