Kybernetika

Fernando Luque-Vásquez; J. Adolfo Minjárez-Sosa

# Empirical approximation in Markov games under unbounded payoff: discounted and average criteria

*Kybernetika*, Vol. 53 (2017), No. 4, 694–716

Persistent URL: `http://dml.cz/dmlcz/146951`

KYBERNETIKA — VOLUME *53* (2017), NUMBER 4, PAGES 694–716

# EMPIRICAL APPROXIMATION IN MARKOV GAMES UNDER UNBOUNDED PAYOFF: DISCOUNTED AND AVERAGE CRITERIA

Fernando Luque-Vásquez and J. Adolfo Minjárez-Sosa

This work deals with a class of discrete-time zero-sum Markov games whose state process $\{x_t\}$ evolves according to the equation $x_{t+1} = F(x_t, a_t, b_t, \xi_t)$, where $a_t$ and $b_t$ represent the actions of player 1 and 2, respectively, and $\{\xi_t\}$ is a sequence of independent and identically distributed random variables with unknown distribution $\theta$. Assuming possibly unbounded payoff, and using the empirical distribution to estimate $\theta$, we introduce approximation schemes for the value of the game as well as for optimal strategies considering both, discounted and average criteria.

*Keywords:* Markov games, empirical estimation, discounted and average criteria

*Classification:* 91A15, 62G07

## 1. INTRODUCTION

In most studies where a stochastic game is analyzed, it is assumed that all the components that define its behavior are completely known by players. However, the environment itself where it evolves makes this assumption unrealistic or too strong. Hence, it is important to have available approximation and estimation algorithms to provide the players some insights on the evolution of the game, in order to more accurately select their actions.

This paper proposes an empirical approximation-estimation algorithm for a class of discrete-time two person zero-sum Markov games evolving according to the difference equation

$$x_{t+1} = F(x_t, a_t, b_t, \xi_t), \quad t = 0, 1, \ldots, \tag{1}$$

where $\{x_t\}$ is the state process, $(a_t, b_t)$ represents the actions chosen by players 1 and 2, respectively, at time $t$, and $\{\xi_t\}$ is the disturbance process which is an observable sequence of independent and identically distributed random variables in a Borel space with arbitrary unknown distribution $\theta$.

Specifically, assuming possibly unbounded payoffs, we use an empirical procedure to estimate $\theta$ which in turn defines an algorithm to approximate the value of the game and

DOI: 10.14736/kyb-2017-4-0694

optimal pairs of strategies. This is done under both, discounted and average criteria, by applying the following general ideas.

As is well known (see, e. g. [5,12,16]), the study of Markov games in discounted case is analyzed via Shapley's equation, which can be represented as $T_\theta^\alpha V^\alpha = V^\alpha$, where $V^\alpha$ is the value of the game and $T_\theta^\alpha$ is a minimax (maximin) operator depending on the distribution $\theta$ and the discount factor $\alpha$. Then, in the scenario of $\theta$ completely known, a stationary optimal pair of strategies $\left(\varphi_*^1, \varphi_*^2\right)$ could be computed. Now, under unknown $\theta$, given a sample $\bar{\xi}_n = (\xi_0, \xi_1, \ldots, \xi_{n-1})$ , the corresponding empirical measure $\theta_n(\cdot) = \theta_n\left(\cdot; \bar{\xi}_n\right)$ defines a random operator $T_{\theta_n}^\alpha$ and an empirical value $V_n^\alpha$ satisfying $T_{\theta_n}^\alpha V_n^\alpha = V_n^\alpha$. So, it is possible to get an optimal pair $\left(\varphi_n^1, \varphi_n^2\right)$ for the $n-$empirical game. Under suitable conditions, we prove the convergence $V_n^\alpha \to V^\alpha$ and the existence of a limit point $\left(\varphi_\infty^1, \varphi_\infty^2\right)$ of $\left\{\left(\varphi_n^1, \varphi_n^2\right)\right\}$ that is an optimal pair for the original game. It is worth observing that by the randomness of operator $T_{\theta_n}^\alpha$ as well as of function $V_n^\alpha$, the pair $\left(\varphi_\infty^1, \varphi_\infty^2\right)$ is a random variable. Hence, additionally we prove that its expectation determines an optimal (non random) pair of strategies $\left(\hat{\varphi}_\infty^1, \hat{\varphi}_\infty^2\right)$ .

On the other hand, the average criterion is studied as a limit of the discounted case. That is, as in [12], we analyze the relation between the average game and the limit behavior of the discounted game, as the discount factor converges to 1. In particular, the limit behavior is obtained by choosing an appropriate sequence $\{\alpha_n\}$ of discount factors converging to one, then we combine this with the process of empirical measures $\{\theta_n\}$ to get the value functions $V_n^{\alpha_n}$ for the $\alpha_n-$discounted empirical game. In this sense, the average optimality is obtained letting $n \to \infty$. For the nature of the average criterion, in contrast to the discounted case, the pair of strategies $\left(\pi^1, \pi^2\right)$, $\pi^1 = \left\{\varphi_n^1\right\}$ and $\pi^2 = \left\{\varphi_n^2\right\}$, computed as the game evolves over time, turns out to be average optimal for the original game.

Similar problems have been studied previously in [16,17] for the discounted and average games, respectively, but under the assumption that $\theta$ has a density. In this particular case, because unbounded payoff is assumed, a complicated density estimation method is proposed. Thus, our present results, in addition to providing a more general method for estimating value functions and for construction of optimal strategies, can be seen as approximation methods in cases where $\theta$ can be known but difficult to handle, i. e., $\theta$ is replaced by a simpler distribution, namely, the empirical distribution $\theta_n$. Further, strategies in [16] are constructed as the game evolves. This means that in distant stages players have more information about the unknown density, and therefore their decisions might be better. However, since the discounted optimality criterion depends heavily on the early stages where the information about the density is poor, this procedure does not guarantee optimality of the resulting strategies. It is for this reason that in [16] the optimality is studied in an asymptotic sense, unlike this work where we obtain an approximation method of optimal strategies.

Approximation algorithms for stochastic games, as well as games with partial information have been studied from several points of view (see, e. g., [1,5,6,13,14,16,17, 18,19,20,21,22], and reference therein). However, for statistical estimation and control procedures for stochastic games, literature remains scarce; we can cite, in addition to [16,17], for instance [18,21,22]. In particular, [18] deals with semi-Markov zero-sum games with unknown sojourn time distribution. The works [21,22] study repeated games assuming that the transition law depends on an unknown parameter, which is estimated by maximum likelihood method.

The paper is organized as follows. In Section 2 is presented the game model, while Section 3 contains the assumptions and some preliminary results about standard games. Next, in Section 4 is introduced the discounted empirical game, and the empirical average game is analyzed in Section 5. We conclude in Section 6 with some remarks about our assumptions.

**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 $\gamma(dx \mid y)$ 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 $Y$. Finally, throughout the paper we assume that a probability space $(\Omega, \mathcal{F}, P)$ is given, and *a.s.* means *almost surely* with respect to $P$.

## 2. THE GAME MODEL

A two person zero-sum Markov game model on Borel spaces is generally described by the following objects. The state space $X$, the actions spaces $A$ and $B$ for player 1 and 2, respectively, and the corresponding constraint sets $\mathbb{K}_A \subset X \times A$ and $\mathbb{K}_B \subset X \times B$. It is assumed that all these sets are Borel spaces. For each $x \in X$, the $x-$sections

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

and

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

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

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

of admissible state-actions triplets is a Borel subset of the Cartesian product $X \times A \times B$. Moreover, the dynamic of the game is represented by a transition law $Q(\cdot|x, a, b)$ which is the distribution of the state variable at time $t+1$, given that the state and actions of players at time $t$ are $x$, $a$, and $b$, respectively. Finally, the one-stage payoff $r(\cdot, \cdot, \cdot)$ is a measurable function on $\mathbb{K}$.

Let $\{\xi_t\}$ be a sequence of observable independent and identically distributed (i.i.d.) random variables defined on the probability space $(\Omega, \mathcal{F}, P)$, taking values in a Borel space $S$, with common distribution $\theta \in \mathbb{P}(S)$. Consider a Markov game evolving according to the difference equation

$$x_{t+1} = F(x_t, a_t, b_t, \xi_t), \quad t \in \mathbb{N}_0. \tag{2}$$

Then the transition law $Q$ is determined by the function $F : \mathbb{K} \times S \to X$ and the distribution $\theta$ as

$$\begin{aligned} Q(D|x,a,b) \quad : \quad &= P\left[x_{t+1} \in D | x_t = x, a_t = a, b_t = b\right] \\ = \quad &\int_S 1_D[F(x,a,b,s)]\theta(\mathrm{d}s), \qquad D \in \mathcal{B}(X), (x,a,b) \in \mathbb{K}. \end{aligned} \tag{3}$$

Hence, this paper is concerned with a zero-sum Markov game, modeled by

$$\mathcal{GM} := (X, A, B, \mathbb{K}_A, \mathbb{K}_B, S, F, \theta, r), \tag{4}$$

which, in a standard sense, is played as follows. At each time $t \in \mathbb{N}_0$, the players observe the state of the game $x_t = x \in X$. Next, players 1 and 2 select, independently, actions $a_t = a \in A(x)$ and $b_t = b \in B(x)$ respectively. Then, player 1 receives a payoff $r(x,a,b)$ from player 2, and the game jumps to a new state $x_{t+1} = y \in X$ according to the transition law (3). Once the game is in the new state, the process is repeated. Therefore, according to the optimality criterion, the goal of player 1 (player 2) is to maximize (minimize) either a discounted or average payoff.

In our empirical approximation settings, we assume that the distribution $\theta \in \mathbb{P}(S)$ is unknown by players, and it is estimated by the corresponding empirical distribution $\theta_t$. Thus, before choosing the actions, on the record of $\xi_0, \xi_1, \ldots, \xi_t$, players 1 and 2 get an estimated $\theta_t(\xi_0, \xi_1, \ldots, \xi_t) = \theta_t$ to select the actions $a = a_t(\theta_t)$ and $b = b_t(\theta_t)$ respectively.

The actions are selected by means of strategies defined as follows. Let $H_0 := X$ and $H_t := \mathbb{K} \times S \times H_{t-1}$ for $t \in \mathbb{N}$. Then, a generic element of $H_t$ is denoted as

$$h_t := (x_0, a_0, b_0, s_0, \ldots, x_{t-1}, a_{t-1}, b_{t-1}, s_{t-1}, x_t)$$

which represents the history of the game up to time $t$. On the other hand, for each $x \in X$, we denote $\mathbb{A}(x) := \mathbb{P}(A(x))$ and $\mathbb{B}(x) := \mathbb{P}(B(x))$, as well as the sets of stochastic kernels

$$\Phi^1 \quad : \quad = \left\{\varphi^1 \in \mathbb{P}(A|X) : \varphi^1(\cdot|x) \in \mathbb{A}(x)\ \forall x \in X\right\}$$

$$\Phi^2 \quad : \quad = \left\{\varphi^2 \in \mathbb{P}(B|X) : \varphi^2(\cdot|x) \in \mathbb{B}(x)\ \forall x \in X\right\}.$$

A *strategy* for player 1 is a sequence $\pi^1 = \{\pi^1_t\}$ of stochastic kernels $\pi^1_t \in \mathbb{P}(A|H_t)$ such that $\pi^1_t(A(x_t)|h_t) = 1\ \forall h_t \in H_t, t \in \mathbb{N}_0$. We denote by $\Pi^1$ the family of all strategies for player 1. A strategy $\pi^1 = \{\pi^1_t\} \in \Pi^1$ is called a *Markov strategy* if $\pi^1_t \in \Phi^1\ \forall t \in \mathbb{N}_0$, and it is called stationary if $\pi^1_n(\cdot|h_n) = \varphi^1(\cdot|x_n)\ \forall h_n \in H_n, n \in \mathbb{N}_0$, for some stochastic kernel $\varphi^1$ in $\Phi^1$, so that $\pi^1$ is of the form $\pi^1 = \left\{\varphi^1, \varphi^1, \ldots\right\} := \left\{\varphi^1\right\}$. We denote by $\Pi^1_s$ the class of stationary strategies for player 1. The sets $\Pi^2$ and $\Pi^2_s$ of all strategies and all stationary strategies for player 2 are defined similarly.

Wherever appropriate, we shall use the following notation related with the probability measures in the sets $\mathbb{A}(x)$ and $\mathbb{B}(x)$. For probability measures $\varphi^1(\cdot|x) \in \mathbb{A}(x)$ and

$\varphi^2(\cdot|x) \in \mathbb{B}(x)$, $x \in X$, we write $\varphi^i(x) = \varphi^i(\cdot|x)$, $i = 1, 2$, and, in addition, for a measurable function $u : \mathbb{K} \to \Re$,

$$u(x, \varphi^1, \varphi^2) = u(x, \varphi^1(x), \varphi^2(x)) := \int_{B(x)} \int_{A(x)} u(x, a, b) \varphi^1(\mathrm{d}a|x) \varphi^2(\mathrm{d}b|x). \tag{5}$$

For instance, for $x \in X$ and $s \in S$, we have

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

and

$$v(F(x, \varphi^1, \varphi^2, s)) := \int_{B(x)} \int_{A(x)} v((F(x, a, b, s)) \varphi^1(\mathrm{d}a|x) \varphi^2(\mathrm{d}b|x),$$

for a measurable function $v : X \to \Re$.

**Optimality criteria.** For each pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and initial state $x_0 = x \in X$, we define the *total expected $\alpha-$discounted payoff* as

$$V_\alpha^\theta(x, \pi^1, \pi^2) := E_x^{\pi^1, \pi^2} \left[ \sum_{t=0}^{\infty} \alpha^t r(x_t, a_t, b_t) \right], \tag{6}$$

where $\alpha \in (0, 1)$ represents the discount factor, and $E_x^{\pi^1, \pi^2}$ denotes the expectation operator with respect to the probability measure $P_x^{\pi^1, \pi^2}$ induced by the pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and $x_0 = x$ (see, e. g., [3]). We also define the *long-run expected average payoff* as

$$J(x, \pi^1, \pi^2) := \liminf_{n \to \infty} \frac{1}{n} E_x^{\pi^1, \pi^2} \sum_{t=0}^{n-1} r(x_t, a_t, b_t). \tag{7}$$

The lower and the upper value of the discounted game are given as:

$$L_\alpha(x) := \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} V_\alpha^\theta(x, \pi^1, \pi^2), \quad x \in X,$$

and

$$U_\alpha(x) := \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} V_\alpha^\theta(x, \pi^1, \pi^2), \quad x \in X,$$

respectively. Observe that, in general, $U_\alpha(\cdot) \geq L_\alpha(\cdot)$, but if it holds that $U_\alpha(\cdot) = L_\alpha(\cdot)$, the common function is called *the $\alpha-$value of the game* and is denoted by $V_\alpha^\theta(\cdot)$. Now, if the discounted game has a value $V_\alpha^\theta(\cdot)$, a strategy $\pi_*^1 \in \Pi^1$ is said to be *$\alpha-$optimal for player 1* if

$$V_\alpha^\theta(x) = \inf_{\pi^2 \in \Pi^2} V_\alpha^\theta(x, \pi_*^1, \pi^2), \quad x \in X.$$

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

$$V_\alpha^\theta(x) = \sup_{\pi^1 \in \Pi^1} V_\alpha^\theta(x, \pi^1, \pi_*^2), \quad x \in X.$$

In this case, $(\pi^1_*, \pi^2_*)$ is an $\alpha-$optimal pair of strategies.

The lower value $L(\cdot)$ and upper value $U(\cdot)$, for the average payoff criterion, are defined similarly, and the *average value of the game* is denoted by $J(\cdot)$. Then, if the average game has a value $J(\cdot)$, a strategy $\pi^1_* \in \Pi^1$ is said to be *average optimal for player 1* if

$$\inf_{\pi^2 \in \Pi^2} J(x, \pi^1_*, \pi^2) = J(x), \quad x \in X;$$

and a strategy $\pi^2_* \in \Pi^2$ is said to be *average optimal for the player 2* if

$$\sup_{\pi^1 \in \Pi^1} J(x, \pi^1, \pi^2_*) = J(x), \quad x \in X.$$

The pair $(\pi^1_*, \pi^2_*)$ is called *average optimal pair* of strategies.

## 3. ASSUMPTIONS AND PRELIMINARY RESULTS

In order to guarantee the existence of values of the discounted and average games, we impose the following sets of assumptions. The first one contains standard continuity and compactness conditions, while Assumption 3.2 is an ergodicity condition, needed to analyze the average criterion (see, e. g., [12, 17] and references therein, and [8, 9] for Markov control processes).

**Assumption 3.1.** (a) The multifunctions $x \longmapsto A(x)$ and $x \longmapsto B(x)$ are compact-valued and continuous.

(b) The payoff function $r$ is continuous on $\mathbb{K}$, and there exist a continuous function $W : X \to [1, \infty)$ and a constant $M > 0$ such that $|r(x, a, b)| \leq MW(x)$ for all $(x, a, b) \in \mathbb{K}$. Moreover, the function

$$(x, a, b) \longmapsto \int_S W\left[F(x, a, b, s)\right] \theta(\mathrm{d}s)$$

is continuous on $\mathbb{K}$.

(c) For each $s \in S$, the function $F(x, a, b, s)$ is continuous in $(x, a, b) \in \mathbb{K}$.

**Assumption 3.2.** There exist a measurable function $\psi : \mathbb{K} \to [0, 1]$, a probability measure $m^*$ on $X$ and a constant $\beta \in (0, 1)$ such that:

**(a)** $\int_S W[F(x, a, b, s)]\theta(\mathrm{d}s) \leq \beta W(x) + \psi(x, a, b)\,\mathrm{d}$, $\forall (x, a, b) \in \mathbb{K}$, where

$$d := \int_X W(x) m^*(\mathrm{d}x) < \infty;$$

**(b)** $Q(D|x, a, b) \geq \psi(x, a, b) m^*(D), \quad \forall D \in \mathcal{B}(X), (x, a, b) \in \mathbb{K};$

**(c)** $\int_X \bar{\Psi}(x) m^*(\mathrm{d}x) > 0$, where $\bar{\Psi}(x) := \inf_{a \in A(x)} \inf_{b \in B(x)} \psi(x, a, b)$ is assumed to be a measurable function.

**Assumption 3.3.** For the constants $\beta$ and $d$ in Assumption 3.2, the function $W$ satisfies

$$W[F(x,a,b,s)] \leq \beta W(x)+d, \ \ \forall (x,a,b,s)\in \mathbb{K}\times S.$$

**Remark 3.4.** (a) If the payoff function $r$ is bounded, say by the constant $M$, then Assumption 3.3 holds by taking $W\equiv 1$ and $d=1$.

(b) Observe that Assumption 3.1(c) implies that the mapping

$$(x,a,b)\longmapsto \int_S v\left[F(x,a,b,s)\right]\mu(\mathrm{d}s)$$

is continuous on $\mathbb{K}$ for every bounded and continuous function $v$ on $X$ and $\mu\in\mathbb{P}(S)$.

(c) We consider the following class of probability measures

$$\mathcal{M}(S):=\left\{\mu\in\mathbb{P}(S):\int_S W[F(x,a,b,s)]\mu(\mathrm{d}s)\leq \beta W(x)+d,\ (x,a,b)\in\mathbb{K}\right\}.$$

Observe that Assumption 3.2(a) implies that $\theta\in\mathcal{M}(S)$, that is

$$\int_S W[F(x,a,b,s)]\theta(\mathrm{d}s)\leq \beta W(x)+d,\ (x,a,b)\in\mathbb{K}. \tag{8}$$

On the other hand, under Assumption 3.3, any probability measure $\mu\in\mathbb{P}(S)$ belongs to $\mathcal{M}(S)$, that is $\mathcal{M}(S)=\mathbb{P}(S)$.

Let $\mathbb{B}_W$ be the family of measurable functions $v:X\to\Re$ with finite $W$-norm

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

We denote by $\mathbb{C}_W\subset\mathbb{B}_W$ the class of continuous functions $v\in\mathbb{B}_W$.

For each $\mu\in\mathbb{P}(S)$ and $\alpha\in(0,1)$, we define the operator

$$T_\mu^\alpha v(x):=\inf_{\varphi^2\in\mathbb{B}(x)}\sup_{\varphi^1\in\mathbb{A}(x)}\left[r(x,\varphi^1,\varphi^2)+\alpha\int_S v(F(x,\varphi^1,\varphi^2,s))\mu(\mathrm{d}s)\right],\ \ v\in\mathbb{B}_W, x\in X. \tag{9}$$

Provided that Assumption 3.1 holds, standard results on game theory (see, e. g., [12]) ensure that, for each $\mu\in\mathcal{M}(S)$, $T_\mu^\alpha$ maps $\mathbb{C}_W$ into itself, and furthermore the interchange of inf and sup in (9) holds:

$$T_\mu^\alpha v(x)=\sup_{\varphi^1\in\mathbb{A}(x)}\inf_{\varphi^2\in\mathbb{B}(x)}\left[r(x,\varphi^1,\varphi^2)+\alpha\int_S v(F(x,\varphi^1,\varphi^2,s))\mu(\mathrm{d}s)\right]. \tag{10}$$

**Remark 3.5.** (Contraction property of $T_\mu^\alpha$) For each discount factor $\alpha\in(0,1)$, we fix an arbitrary number $\gamma_\alpha\in(\alpha,1)$ and define the function $\bar{W}(x):=W(x)+e$, $x\in X$, where $e:=d\left(\gamma_\alpha/\alpha-1\right)^{-1}$. Consider the space $\mathbb{B}_{\bar{W}}$ of measurable functions $v:X\to\Re$ with finite $\bar{W}-$norm, that is

$$\|v\|_{\bar{W}}:=\sup_{x\in\mathbf{X}}\frac{|v(x)|}{\bar{W}(x)}<\infty.$$

Observe that $\mathbb{B}_W = \mathbb{B}_{\bar{W}}$ and the norms $\|\cdot\|_W$ and $\|\cdot\|_{\bar{W}}$ are equivalent since

$$\|v\|_{\bar{W}} \leq \|v\|_W \leq l_\alpha \|v\|_{\bar{W}}, \quad v \in \mathbb{B}_W \tag{11}$$

where

$$l_\alpha := 1 + e = 1 + \frac{\alpha d}{\gamma_\alpha - \alpha}. \tag{12}$$

Then, from [25, Lemma 1], for any $\mu \in \mathcal{M}(S)$, the function $\bar{W}$ satisfies the inequality

$$\alpha \int_S \bar{W}[F(x,a,b,s)]\mu(\mathrm{d}s) \leq \gamma_\alpha \bar{W}(x), \ \ \forall (x,a,b) \in \mathbb{K}. \tag{13}$$

Thus, following straightforward calculations, it is easy to see that, for each $\alpha \in (0,1)$ and $\mu \in \mathcal{M}(S)$, inequality (13) implies that operator $T_\mu^\alpha$ is a contraction with respect to the $\bar{W}-$norm with modulus $\gamma_\alpha$. That is, for all $v, u \in B_W$,

$$\left\|T_\mu^\alpha v - T_\mu^\alpha u\right\|_{\bar{W}} \leq \gamma_\alpha \|v - u\|_{\bar{W}}. \tag{14}$$

From [12, Theorem 4.2], we have the following result.

**Theorem 3.6.** Suppose that Assumption 3.1 holds and $\theta \in \mathcal{M}(S)$. Then, for each $\alpha \in (0,1)$:

**(a)** The discounted game has a value $V_\alpha^\theta \in \mathbb{C}_W$ and there exists a constant $L := L(\alpha, \beta)$ such that

$$\left\|V_\alpha^\theta\right\|_W \leq L.$$

**(b)** The value $V_\alpha^\theta$ satisfies $T_\theta^\alpha V_\alpha^\theta = V_\alpha^\theta$, and there exists $(\varphi_*^1, \varphi_*^2) \in \Phi^1 \times \Phi^2$, such that $\varphi_*^1(x) \in \mathbb{A}(x)$ and $\varphi_*^2 \in \mathbb{B}(x)$ satisfy

$$\begin{aligned} V_\alpha^\theta(x) &= r(x,\varphi_*^1,\varphi_*^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi_*^1,\varphi_*^2,s)]\theta(\mathrm{d}s) \\ &= \max_{\varphi^1 \in \mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_*^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi^1,\varphi_*^2,s)]\theta(\mathrm{d}s) \right] \qquad (15) \\ &= \min_{\varphi^2 \in \mathbb{B}(x)} \left[ r(x,\varphi_*^1,\varphi^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi_*^1,\varphi^2,s)]\theta(\mathrm{d}s) \right], \quad \forall x \in X. \quad (16) \end{aligned}$$

In addition, $\pi_*^1 = \left\{\varphi_*^1\right\} \in \Pi_s^1$ and $\pi_*^2 = \left\{\varphi_*^2\right\} \in \Pi_s^2$ form an optimal pair of strategies.

Furthermore, from [12, Theorem 4.3], we have the following result related with the average game.

**Theorem 3.7.** Under Assumptions 3.1 and 3.2, the average game has a value $J(\cdot) = j^*$, that is

$$j^* = \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(x,\pi^1,\pi^2) = \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(x,\pi^1,\pi^2), \quad \forall x \in \mathbf{X}.$$

In addition, both players has optimal strategies.

**Remark 3.8.** (Vanishing discount factor approach) The relation between the discounted and average criteria is given as follows. Let $z \in X$ be a fixed state, and define, for $\alpha \in (0,1)$

$$j_\alpha^\theta := (1-\alpha)V_\alpha^\theta(z), \qquad \phi_\alpha^\theta(x) := V_\alpha^\theta(x) - V_\alpha^\theta(z), \quad x \in X. \tag{17}$$

Observe that from Theorem 3.6(b),

$$\begin{aligned} j_\alpha^\theta + \phi_\alpha^\theta(x) &= T_\theta^\alpha \phi_\alpha^\theta(x) = r(x,\varphi_*^1,\varphi_*^2) + \alpha \int_S \phi_\alpha^\theta[F(x,\varphi_*^1,\varphi_*^2,s)]\theta(\mathrm{d}s) \\ &= \max_{\varphi^1 \in \mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_*^2) + \alpha \int_S \phi_\alpha^\theta[F(x,\varphi^1,\varphi_*^2,s)]\theta(\mathrm{d}s) \right] \\ &= \min_{\varphi^2 \in \mathbb{B}(x)} \left[ r(x,\varphi_*^1,\varphi^2) + \alpha \int_S \phi_\alpha^\theta[F(x,\varphi_*^1,\varphi^2,s)]\theta(\mathrm{d}s) \right], \quad \forall x \in X. \end{aligned} \tag{18}$$

Then, from [12, Theorem 4.3], under Assumptions 3.1 and 3.2

$$\lim_{t\to\infty} j_{\alpha_t}^\theta = j^*, \tag{19}$$

for any sequence $\{\alpha_t\}$ of discount factors, such that $\alpha_t \nearrow 1$. Moreover,

$$\sup_{\alpha \in (0,1)} \left\| \phi_\alpha^\theta \right\|_W < \infty. \tag{20}$$

## 4. THE DISCOUNTED EMPIRICAL GAME

We consider the approximated zero-sum Markov game model of the form:

$$\mathcal{GM}_t^\alpha := (X, A, B, \mathbb{K}_A, \mathbb{K}_B, S, F, \theta_t, r), \tag{21}$$

where $\theta_t \in \mathbb{P}(S)$, $t \in \mathbb{N}_0$, is the empirical distribution of the disturbance process $\{\xi_t\}$ defined as follows. Let $\nu \in \mathbb{P}(S)$ be a given probability measure. Then

$$\begin{aligned} &\theta_0 := \nu, \\ &\theta_t(D) = \theta_t(D)(\omega) := \frac{1}{t}\sum_{i=0}^{t-1} 1_D(\xi_i(\omega)), \quad \text{for all } t \in \mathbb{N},\ D \in \mathcal{B}(S),\ \omega \in \Omega. \end{aligned}$$

Note that for each $D \in B(S)$, $\theta_t(D)(\cdot)$ is a random variable, and for each $\omega \in \Omega$, $\theta_t(\cdot)(\omega)$ is the uniform distribution on the set $\{\xi_0(\omega), \ldots, \xi_{t-1}(\omega)\} \subset S$.

The empirical approximation scheme consists in solving the approximate game $\mathcal{GM}_t$, for each $t \in \mathbb{N}$. That is, the discounted game is analyzed when both players use the empirical distribution $\theta_t$ instead of the original distribution $\theta$. This procedure leads up to obtain an optimal pair of strategies $(\pi_t^1, \pi_t^2) \in \Pi_s^1 \times \Pi_s^2$ for the game $\mathcal{GM}_t^\alpha$, $t \in \mathbb{N}$, provided, of course, that the corresponding value of the game $V_\alpha^{\theta_t}$ exists. Under this settings, our hope is that the optimal pair $(\pi_t^1, \pi_t^2)$ has a good performance in the game $\mathcal{GM}$, whenever the sequence of empirical values $\{V_\alpha^{\theta_t}\}$ gives a good approximation to the value $V_\alpha^\theta$. We introduce these ideas in precise terms as follows.

For each $t \in \mathbb{N}_0$, let

$$V_\alpha^{\theta_t}(x, \pi^1, \pi^2) := E_t^{x,\pi^1,\pi^2}\left[\sum_{i=0}^{\infty} \alpha^i r(x_i, a_i, b_i)\right], \tag{22}$$

be the $\alpha$-discounted expected payoff function in which all random variables $\xi_0^t, \xi_1^t, \ldots,$ have the same distribution $\theta_t$.

Observe that, under Assumption 3.3, $\theta_t \in \mathcal{M}(S)$, for every $t \in \mathbb{N}$, that is

$$\int_S W[F(x, a, b, s)]\theta_t(\mathrm{d}s)(\omega) \leq \beta W(x) + d, \quad \forall (x, a, b) \in \mathbb{K}.$$

Then Theorem 3.6 yields the following result.

**Theorem 4.1.** Suppose that Assumptions 3.1 and 3.3 hold. Then for each $t \in \mathbb{N}$ and $\omega \in \Omega$,

**(a)** the game $\mathcal{GM}_t$ has a value $V_\alpha^{\theta_t} = V_\alpha^{\theta_t(\omega)} \in \mathbb{C}_W$ such that

$$\left\|V_\alpha^{\theta_t}\right\|_W \leq L \quad \text{and} \quad T_{\theta_t}^\alpha V_\alpha^{\theta_t} = V_\alpha^{\theta_t};$$

**(b)** there exists $(\varphi_t^1, \varphi_t^2) = (\varphi_t^1(\omega), \varphi_t^2(\omega)) \in \Phi^1 \times \Phi^2$ such that, $\varphi_t^1(x, \omega) := \varphi_t^1(\cdot|x, \omega) \in \mathbb{A}(x)$ and $\varphi_t^2(x, \omega) := \varphi_t^2(\cdot|x, \omega) \in \mathbb{B}(x)$ satisfy

$$\begin{aligned} V_\alpha^{\theta_t}(x) &= r(x, \varphi_t^1, \varphi_t^2) + \alpha \int_S V_\alpha^{\theta_t}[F(x, \varphi_t^1, \varphi_t^2, s)]\theta_t(\mathrm{d}s) \\ &= \max_{\varphi^1 \in \mathbb{A}(x)} \left[r(x, \varphi^1, \varphi_t^2) + \alpha \int_S V_\alpha^{\theta_t}[F(x, \varphi^1, \varphi_t^2, s)]\theta_t(\mathrm{d}s)\right] \qquad (23) \\ &= \min_{\varphi^2 \in \mathbb{B}(x)} \left[r(x, \varphi_t^1, \varphi^2) + \alpha \int_S V_\alpha^{\theta_t}[F(x, \varphi_t^1, \varphi^2, s)]\theta_t(\mathrm{d}s)\right], \forall x \in X. \end{aligned} \tag{24}$$

**Remark 4.2.** (a) Observe that, for each $t \in \mathbb{N}_0$, the value function $V_\alpha^{\theta_t}$ is a random function, and $\varphi_t^i(x, \omega)$, $i = 1, 2$, define a random optimal pair of strategies $(\pi_t^1, \pi_t^2) := \left(\{\varphi_t^1\}, \{\varphi_t^2\}\right) \in \Pi_s^1 \times \Pi_s^2$ for the game $\mathcal{GM}_t^\alpha$.

(b) Let us fix $(x, \omega) \in X \times \Omega$, and consider the multifunction given by $(x, \omega) \longmapsto \mathbb{A}(x)$. Since $A(x)$ is a compact subset of $A$, $\mathbb{A}(x)$ is a compact subset of $\mathbb{P}(A)$ (with the weak topology). Then, from [10, Proposition D.5] (see also [23]), there exists $\varphi_\infty^1 \in \Phi^1$ such that $\varphi_\infty^1(x, \omega) = \varphi_\infty^1(\cdot|x, \omega) \in \mathbb{A}(x)$ is an accumulation point of $\{\varphi_t^1(x, \omega)\}$. Similarly, there exists $\varphi_\infty^2 \in \Phi^2$ such that $\varphi_\infty^2(x, \omega) = \varphi_\infty^2(\cdot|x, \omega) \in \mathbb{B}(x)$ is an accumulation point of $\{\varphi_t^2(x, \omega)\}$.

**The estimation process.** The key points to obtain the approximation of the empirical values $V_\alpha^{\theta_t}$ to the value $V_\alpha^\theta$ are the convergence properties of the empirical distribution. At first glance, it is well-known that $\theta_t$ converges weakly to $\theta$ a.s., that is, for each $(x,a,b)\in\mathbb{K}$,

$$\int_S u(F(x,a,b,s))\theta_t(\mathrm{d}s)\to\int_S u(F(x,a,b,s))\theta(\mathrm{d}s)\ \text{ a.s., as } t\to\infty,$$

for every continuous and bounded function $u$ on $X$. However, in the scenario of possibly unbounded payoff, this kind of convergence is not sufficient for our objectives. In fact, we need uniform convergence on the set $\mathbb{K}$. In order to state our estimation process, we impose the following assumption.

**Assumption 4.3.** The family of functions

$$\mathcal{V}_W:=\left\{\frac{V_\alpha^\theta(F(x,a,b,\cdot))}{W(x)}:(x,a,b)\in\mathbb{K}\right\}\tag{25}$$

is equicontinuous on $S$.

**Remark 4.4.** Observe that from Theorem 3.6, the family of functions $\mathcal{V}_W$ is uniformly bounded. Then, under Assumption 4.3 and using [24, Theorem 3.1] we have

$$\Delta_t\to 0\ \ a.s.,\ \text{ as }\ t\to\infty,\tag{26}$$

where

$$\Delta_t:=\sup_{(x,a,b)\in\mathbb{K}}\left|\int_S\frac{V_\alpha^\theta(F(x,a,b,s))}{W(x)}\theta_t(\mathrm{d}s)-\int_S\frac{V_\alpha^\theta(F(x,a,b,s))}{W(x)}\theta(\mathrm{d}s)\right|.$$

Hence, we can state our main results related with the discounted empirical approximation as follows.

**Theorem 4.5.** Under Assumptions 3.1, 3.3, and 4.3, $P-a.s.$

(a) $\left\|V_\alpha^{\theta_t}-V_\alpha^\theta\right\|_W\to 0$ as $t\to\infty$;

(b) the random pair of strategies $(\pi_\infty^1,\pi_\infty^2)\in\Pi_s^1\times\Pi_s^2$ defined as $\pi_\infty^1=\left\{\varphi_\infty^1\right\}$ and $\pi_\infty^2=\left\{\varphi_\infty^2\right\}$ is optimal for the game $\mathcal{GM}$ (see Remark 4.2 (b)).

(c) Furthermore, there exists an optimal (non random) pair of strategies $(\hat\varphi_\infty^1,\hat\varphi_\infty^2)\in\Pi_s^1\times\Pi_s^2$ for the game $\mathcal{GM}$ given as

$$\hat\varphi_\infty^i(\cdot|x)=\int_\Omega\varphi_t^i(\cdot|x,\omega)P(\mathrm{d}\omega),\quad i=1,2.$$

Proof. (a) Since $\theta_t\in\mathcal{M}(S)$, $t\in\mathbb{N}_0$, from (14) we have that the operator $T_{\theta_t}^\alpha$ is a contraction on $\mathbb{B}_{\bar W}$. Hence, from Theorems 3.6 and 4.1, for each $t\in\mathbb{N}_0$,

$$\begin{aligned}\left\|V_\alpha^\theta-V_\alpha^{\theta_t}\right\|_{\bar W}&\le\left\|T_\theta^\alpha V_\alpha^\theta-T_{\theta_t}^\alpha V_\alpha^\theta\right\|_{\bar W}+\left\|T_{\theta_t}^\alpha V_\alpha^\theta-T_{\theta_t}^\alpha V_\alpha^{\theta_t}\right\|_{\bar W}\\&\le\left\|T_\theta^\alpha V_\alpha^\theta-T_{\theta_t}^\alpha V_\alpha^\theta\right\|_{\bar W}+\gamma_\alpha\left\|V_\alpha^\theta-V_\alpha^{\theta_t}\right\|_{\bar W}\quad\text{a.s.}\end{aligned}$$

Thus

$$\left\|V_\alpha^{\theta_t}-V_\alpha^\theta\right\|_{\bar{W}} \leq \frac{1}{1-\gamma_\alpha}\left\|T_\theta^\alpha V_\alpha^\theta-T_{\theta_t}^\alpha V_\alpha^\theta\right\|_{\bar{W}}. \tag{27}$$

On the other hand, using the fact that $\bar{W}(\cdot)>W(\cdot)$, for each $x\in X$ and $t\in\mathbb{N}_0$,

$$\begin{aligned}&\left\|T_\theta^\alpha V_\alpha^\theta-T_{\theta_t}^\alpha V_\alpha^\theta\right\|_{\bar{W}}\\ \leq\;& \sup_{x\in X}\sup_{\varphi^1\in\mathbb{A}(x),\varphi^2\in\mathbb{B}(x)}\left|\int_S\frac{V_\alpha^\theta[F(x,\varphi^1,\varphi^2,s)]}{W(x)}\theta(\mathrm{d}s)-\int_S\frac{V_\alpha^\theta[F(x,\varphi^1,\varphi^2,s)]}{W(x)}\theta_t(\mathrm{d}s)\right|\\ \leq\;& \Delta_t. \end{aligned}\tag{28}$$

Combining (27) and (28) we get

$$\left\|V_\alpha^{\theta_t}-V_\alpha^\theta\right\|_{\bar{W}} \leq \frac{1}{1-\gamma_\alpha}\Delta_t,$$

and from (11)

$$\left\|V_\alpha^{\theta_t}-V_\alpha^\theta\right\|_{W} \leq \frac{l_\alpha}{1-\gamma_\alpha}\Delta_t. \tag{29}$$

Finally, (26) proves the part (a).

(b) Since for each $(x,\omega)\in X\times\Omega$, $\varphi_\infty^1(x,\omega)=\varphi_\infty^1(\cdot|x,\omega)\in\mathbb{A}(x)$ is an accumulation point of $\{\varphi_t^1(x,\omega)\}$, there exists a subsequence $\{\varphi_{t_k}^1(x,\omega)\}$ of $\{\varphi_t^1(x,\omega)\}$ such that $\varphi_\infty^1(x,\omega)=\lim_{k\to\infty}\varphi_{t_k}^1(x,\omega)$. Under the same arguments, there exists a subsequence $\{\varphi_{t_k}^2(x,\omega)\}$ of $\{\varphi_t^2(\cdot|x,\omega)\}$ such that $\varphi_\infty^2(x,\omega)=\lim_{k\to\infty}\varphi_{t_k}^2(x,\omega)$. Observe that we can use the same subsequence $\{t_k\}$ for both cases. In the remainder of the proof, to ease notation, we let $t_k=k$.

We shall now proceed to prove the optimality of the pair $(\pi_\infty^1,\pi_\infty^2)\in\Pi_s^1\times\Pi_s^2$.

Firstly, observe that, for each $x\in X$,

$$\sup_{(a,b)\in A(x)\times B(x)}\left|\int_S V_\alpha^{\theta_t}\left(F(x,a,b,s)\right)\theta_t(\mathrm{d}s)-\int_S V_\alpha^{\theta}\left(F(x,a,b,s)\right)\theta(\mathrm{d}s)\right|\to 0 \text{ a.s., as } t\to\infty. \tag{30}$$

Indeed,

$$\begin{aligned}&\left|\int_S V_\alpha^{\theta_t}\left(F(x,a,b,s)\right)\theta_t(\mathrm{d}s)-\int_S V_\alpha^{\theta}\left(F(x,a,b,s)\right)\theta(\mathrm{d}s)\right|\\ &\leq \int_S\left|V_\alpha^{\theta_t}\left(F(x,a,b,s)\right)-V_\alpha^{\theta}\left(F(x,a,b,s)\right)\right|\theta_t(\mathrm{d}s)\\ &+\left|\int_S V_\alpha^{\theta}(F(x,a,b,s))\theta_t(\mathrm{d}s)-\int_S V_\alpha^{\theta}(F(x,a,b,s))\theta(\mathrm{d}s)\right|\\ &\leq \left\|V_\alpha^{\theta_t}-V_\alpha^\theta\right\|_W(\beta W(x)+d)+\Delta_t W(x).\end{aligned}$$

Thus, (30) follows from part (a) and (26).

Now, from (23),

$$V_\alpha^{\theta_k}(x) = \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_k^2) + \alpha \int_S V_\alpha^{\theta_k}[F(x,\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) \right]. \tag{31}$$

In addition, for a fixed $\bar\varphi^1 \in \mathbb{A}(x)$

$$\begin{aligned} &\liminf_k \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_k^2) + \alpha \int_S V_\alpha^{\theta_k}[F(x,\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) \right] \\ \geq\ &\liminf_k \left[ r(x,\bar\varphi^1,\varphi_k^2) + \alpha \int_S V_\alpha^{\theta_k}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) \right]. \end{aligned} \tag{32}$$

On the other hand,

$$\begin{aligned} \int_S V_\alpha^{\theta_k}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) &= \int_S V_\alpha^{\theta_k}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) \\ -\int_S V_\alpha^{\theta}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta(\mathrm{d}s) &+ \int_S V_\alpha^{\theta}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta(\mathrm{d}s). \end{aligned}$$

Then, from (30), Fatou´s Lemma, and using the continuity of the functions $V_\alpha^\theta$ and $F$,

$$\begin{aligned} \liminf_k \int_S V_\alpha^{\theta_k}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta_k(\mathrm{d}s) &= \liminf_k \int_S V_\alpha^{\theta}[F(x,\bar\varphi^1,\varphi_k^2,s)]\theta(\mathrm{d}s) \\ &\geq \int_S V_\alpha^{\theta}[F(x,\bar\varphi^1,\varphi_\infty^2,s)]\theta(\mathrm{d}s) \quad a.s. \end{aligned} \tag{33}$$

Therefore, taking liminf as $k \to \infty$ in (31), relations (32) and (33) together part (a) yield

$$V_\alpha^\theta(x) \geq r(x,\bar\varphi^1,\varphi_\infty^2) + \alpha \int_S V_\alpha^\theta[F(x,\bar\varphi^1,\varphi_\infty^2,s)]\theta(\mathrm{d}s).$$

Since $\bar\varphi^1 \in \mathbb{A}(x)$ is arbitrary, we have

$$V_\alpha^\theta(x) \geq \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_\infty^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi^1,\varphi_\infty^2,s)]\theta(\mathrm{d}s) \right],$$

which implies

$$V_\alpha^\theta(x) = \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi_\infty^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi^1,\varphi_\infty^2,s)]\theta(\mathrm{d}s) \right], \tag{34}$$

because (see (15))

$$\begin{aligned} V_\alpha^\theta(x) &= \min_{\varphi^2\in\mathbb{B}(x)} \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi^1,\varphi^2,s)]\theta(\mathrm{d}s) \right] \\ &\leq \max_{\varphi^1\in\mathbb{A}(x)} \left[ r(x,\varphi^1,\varphi^2) + \alpha \int_S V_\alpha^\theta[F(x,\varphi^1,\varphi^2,s)]\theta(\mathrm{d}s) \right], \quad \forall\varphi^2 \in \mathbb{B}(x). \end{aligned}$$

Similarly, from (24),

$$V_\alpha^{\theta_k}(x)=\min_{\varphi^2\in\mathbb{B}(x)}\left[r(x,\varphi_k^1,\varphi^2)+\alpha\int_S V_\alpha^{\theta_k}[F(x,\varphi_k^1,\varphi^2,s)]\theta_k(\mathrm{d}s)\right],$$

and for an arbitrary and fixed $\bar{\varphi}^2\in\mathbb{B}(x)$

$$\begin{aligned}&\limsup_k \min_{\varphi^2\in\mathbb{B}(x)}\left[r(x,\varphi_k^1,\varphi^2)+\alpha\int_S V_\alpha^{\theta_k}[F(x,\varphi_k^1,\varphi^2,s)]\theta_k(\mathrm{d}s)\right]\\ \leq\ &\limsup_k\left[r(x,\varphi_k^1,\bar{\varphi}^2)+\alpha\int_S V_\alpha^{\theta_k}[F(x,\varphi_k^1,\bar{\varphi}^2,s)]\theta_k(\mathrm{d}s)\right].\end{aligned}$$

Thus, applying Fatou's Lemma with limsup, we obtain

$$V_\alpha^{\theta}(x)\leq r(x,\varphi_\infty^1,\bar{\varphi}^2)+\alpha\int_S V_\alpha^{\theta}[F(x,\varphi_\infty^1,\bar{\varphi}^2,s)]\theta(\mathrm{d}s),$$

which, in turns, implies

$$V_\alpha^{\theta}(x)=\min_{\varphi^2\in\mathbb{B}(x)}\left[r(x,\varphi_\infty^1,\varphi^2)+\alpha\int_S V_\alpha^{\theta}[F(x,\varphi_\infty^1,\varphi^2,s)]\theta(\mathrm{d}s)\right]. \tag{35}$$

Finally, combining (34) and (35), and applying standard procedures in game theory, we prove that $(\pi_\infty^1,\pi_\infty^2)\in\Pi_s^1\times\Pi_s^2$ is a random optimal pair of strategies for the game $\mathcal{GM}$.

(c) For $i=1,2$, let $\hat{\pi}_\infty^i$ be the strategies determined by

$$\hat{\varphi}_\infty^i(x)=\hat{\varphi}_\infty^i(\cdot|x):=\int_\Omega \varphi_\infty^i(\cdot|x,\omega)P(\mathrm{d}\omega)\in\Phi^i.$$

We define

$$H(x,a,b):=r(x,a,b)+\alpha\int_S V_\alpha^{\theta}[F(x,a,b,s)]\theta(\mathrm{d}s),\quad (x,a,b)\in\mathbb{K}, \tag{36}$$

Observe that, from (35) and (5),

$$\begin{aligned}V_\alpha^{\theta}(x)&=\min_{\varphi^2\in\mathbb{B}(x)}H(x,\varphi_\infty^1(\omega),\varphi^2)\\&=\min_{\varphi^2\in\mathbb{B}(x)}\int_{A(x)}H(x,a,\varphi^2)\varphi_\infty^1(\mathrm{d}a|x,\omega)\quad a.s.,\quad x\in X.\end{aligned}$$

Hence,

$$\begin{aligned}V_\alpha^{\theta}(x)&=\int_\Omega\min_{\varphi^2\in\mathbb{B}(x)}\int_{A(x)}H(x,a,\varphi^2)\varphi_\infty^1(\mathrm{d}a|x,\omega)P(\mathrm{d}\omega)\\&\leq\min_{\varphi^2\in\mathbb{B}(x)}\int_{A(x)}H(x,a,\varphi^2)\int_\Omega\varphi_\infty^1(\mathrm{d}a|x,\omega)P(\mathrm{d}\omega)\\&=\min_{\varphi^2\in\mathbb{B}(x)}\int_{A(x)}H(x,a,\varphi^2)\hat{\varphi}_\infty^1(\mathrm{d}a|x)\\&=\min_{\varphi^2\in\mathbb{B}(x)}H(x,\hat{\varphi}_\infty^1,\varphi^2),\quad x\in X.\end{aligned}$$

Therefore, from (36) and Theorem 3.6,

$$V_\alpha^\theta(x) = \min_{\varphi^2 \in \mathbb{B}(x)} \left[ r(x, \hat{\varphi}_\infty^1, \varphi^2) + \alpha \int_S V_\alpha^\theta[F(x, \hat{\varphi}_\infty^1, \varphi^2, s)]\theta(\mathrm{d}s) \right], \quad \forall x \in X. \tag{37}$$

Similarly, we can prove that, for each $x \in X$,

$$V_\alpha^\theta(x) = \max_{\varphi^1 \in \mathbb{A}(x)} \left[ r(x, \varphi^1, \hat{\varphi}_\infty^2) + \alpha \int_S V_\alpha^\theta[F(x, \varphi^1, \hat{\varphi}_\infty^2, s)]\theta(\mathrm{d}s) \right], \quad \forall x \in X,$$

which, combined with (37), yields the optimality of the pair $(\hat{\pi}_\infty^1, \hat{\pi}_\infty^2) \in \Pi_s^1 \times \Pi_s^2$ for the game $\mathcal{GM}$. □

## 5. EMPIRICAL APPROXIMATION UNDER AVERAGE CRITERION

The empirical approximation scheme for the average criterion is obtained by combining the vanishing discount factor approach (see Remark 3.8) and a suitable convergence property of the empirical process. Therefore, we will take advantage of the results introduced in previous sections for the discounted criterion. However, due to the additional difficulties in the asymptotic analysis of the average payoff, the following stronger condition is needed.

**Assumption 5.1.** a) The disturbance space $S$ is the $k-$dimensional Euclidean space $\Re^k$.

b) Let $m > \max\{2, k\}$ be an arbitrary real number and $\bar{m} := km/[(m-k)(m-2)]$. Then $E\left|\xi_0\right|^{\bar{m}} < \infty$.

c) The family of functions (see (17) and (25))

$$\bar{\mathcal{V}}_W := \left\{ \frac{\phi_\alpha^\theta\left(F(x,a,b,.)\right)}{W(x)} : (x,a,b) \in \mathbb{K}, \alpha \in (0,1) \right\},$$

or equivalently (see (17))

$$\hat{\mathcal{V}}_W := \left\{ \frac{V_\alpha^\theta\left(F(x,a,b,.)\right)}{W(x)} : (x,a,b) \in \mathbb{K}, \alpha \in (0,1) \right\},$$

is equi-Lipschitzian on $\Re^k$. That is, there exists a constant $L_\phi > 0$ such that

$$\left| \frac{\phi_\alpha^\theta\left(F(x,a,b,s)\right)}{W(x)} - \frac{\phi_\alpha^\theta\left(F(x,a,b,s')\right)}{W(x)} \right| \le L_\phi \left|s - s'\right|, \quad \forall s, s' \in \Re^k, (x,a,b) \in \mathbb{K},$$

where $|\cdot|$ is the Euclidean distance in $\Re^k$.

**Remark 5.2.** (Equicontinuity and equi-Lipschitz conditions) Clearly, in the case $S = \Re^k$, equi-Lipschitz Assumption 5.1 (c) implies equicontinuity Assumption 4.3. Now, equi-Lipschitz assumption is satisfied under several set of conditions, for instance (see [4]) when

$$s \longmapsto \frac{\phi_\alpha^\theta\left(F(x,a,b,s)\right)}{W(x)}$$

is a convex or concave function for all $(x, a, b) \in \mathbb{K}$, $\alpha \in (0, 1)$. Another set of conditions can be obtained from [8,9] by imposing Lipschitz-like conditions on the payoff functions $r$ and on the transition kernel (3).

Under Assumptions 3.1, 3.2, and 5.1 (c), from (20) and Theorem 3.6, the family of functions $\bar{\mathcal{V}}_W$ is uniformly bounded. Furthermore, [2, Proposition 3.4] yields the existence of a constant $\bar{M}$ such that

$$E\left[\bar{\Delta}_t\right] \leq \bar{M} t^{-1/m}, \tag{38}$$

where

$$\bar{\Delta}_t := \sup_{(x,a,b)\in\mathbb{K},\alpha\in(0,1)} \left| \int_{\Re^k} \frac{\phi_\alpha(F(x,a,b,s))}{W(x)} \theta_t(\mathrm{d}s) - \int_{\Re^k} \frac{\phi_\alpha(F(x,a,b,s))}{W(x)} \theta(\mathrm{d}s) \right|. \tag{39}$$

The empirical vanishing discount factor approach consists in the following. Let $\nu \in (0, 1/2m)$ be an arbitrary real number where $m$ is the constant introduced in Assumption 5.1(b). By borrowing the ideas in [7, 15], we fix an arbitrary nondecreasing sequence of discount factors $\{\bar{\alpha}_t\}$ such that

**D.1** $(1 - \bar{\alpha}_t)^{-1} = \mathbf{O}(t^\nu)$ as $t \to \infty$;

**D.2** $\displaystyle\lim_{n\to\infty} \frac{\kappa(n)}{n} = 0$,

where $\kappa(n)$ is the number of changes of value of $\{\bar{\alpha}_t\}$ among the first $n$ terms.

An example of a sequence $\{\bar{\alpha}_t\}$ satisfying Conditions D.1 and D.2 is the following. Let $m = 3$ (see Assumption 5.1), $\nu = 1/10$ and $\{\alpha_t\}$ be the sequence defined as

$$\alpha_t := 1 - \frac{1}{t^\nu}.$$

Now, define the sequence $\{\bar{\alpha}_t\}$ by

$$\bar{\alpha}_t = \alpha_k \quad \text{if } \frac{(k-1)k}{2} \leq t < \frac{k(k+1)}{2}, \quad t \in \mathbb{N},\ k = 2, 3, \ldots$$

Then,

$$(1 - \bar{\alpha}_t)^{-1} = (1 - \alpha_k)^{-1} = k^\nu = \mathbf{O}(t^\nu)$$

since $k \leq t$. Moreover, for $n \geq 1$, $\kappa(n) = (k-2)$ if $(k-1)k/2 \leq n < k(k+1)/2$, therefore $\kappa(n) < \sqrt{2n}$ and

$$\frac{\kappa(n)}{n} \to 0.$$

For a fixed $t \in \mathbb{N}_0$, let $V_{\bar{\alpha}_t}^{\theta_t}(\cdot, \cdot, \cdot)$ be the $\bar{\alpha}_t-$discounted payoff function under the empirical distribution $\theta_t$ (see (22)), and we denote by $V_{\bar{\alpha}_t}^{\theta_t}(\cdot)$ the corresponding value of the game $\mathcal{GM}_t^{\bar{\alpha}_t}$ (see (21), Theorems 3.6 and 4.1). The functions $\phi_{\bar{\alpha}_t}^{\theta_t}(\cdot)$ and $j_{\bar{\alpha}_t}^{\theta_t}$ are defined accordingly (see (17)). Hence, from Theorem 3.6(b) (see (18)), there exists a random pair $(\bar{\varphi}_t^1, \bar{\varphi}_t^2) \in \Phi^1 \times \Phi^2$ such that, for every $x \in X$,

$$\begin{aligned} j_{\bar{\alpha}_t}^{\theta_t} + \phi_{\bar{\alpha}_t}^{\theta_t}(x) &= T_{\theta_t}^{\bar{\alpha}_t}\phi_{\bar{\alpha}_t}^{\theta_t}(x) = r(x, \bar{\varphi}_t^1, \bar{\varphi}_t^2) + \bar{\alpha}_t \int_S \phi_{\bar{\alpha}_t}^{\theta_t}[F(x, \bar{\varphi}_t^1, \bar{\varphi}_t^2, s)]\theta_t(\mathrm{d}s) \\ &= \max_{\varphi^1 \in \mathbb{A}(x)} \left[ r(x, \varphi^1, \bar{\varphi}_t^2) + \bar{\alpha}_t \int_S \phi_{\bar{\alpha}_t}^{\theta_t}[F(x, \varphi^1, \bar{\varphi}_t^2, s)]\theta_t(\mathrm{d}s) \right] \\ &= \min_{\varphi^2 \in \mathbb{B}(x)} \left[ r(x, \bar{\varphi}_t^1, \varphi^2) + \bar{\alpha}_t \int_S \phi_{\bar{\alpha}_t}^{\theta_t}[F(x, \bar{\varphi}_t^1, \varphi^2, s)]\theta_t(\mathrm{d}s) \right]. \end{aligned} \tag{40}$$

Let $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ be the pair of strategies determined by $(\bar{\varphi}_t^1, \bar{\varphi}_t^2) \in \Phi^1 \times \Phi^2$. That is, $\pi_*^i = \left\{ \bar{\varphi}_t^i \right\} = \left\{ \bar{\varphi}_t^i(\cdot|x, \omega) \right\}$ for $i = 1, 2$. Then, our main result is stated as follows.

**Theorem 5.3.** Under Assumption 3.1, 3.2, and 5.1, $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ is a random pair of average optimal strategies for the game $\mathcal{GM}$, that is

$$j^* = \inf_{\pi^2 \in \Pi^2} J(x, \pi_*^1, \pi^2) = \sup_{\pi^1 \in \Pi^1} J(x, \pi^1, \pi_*^2), \quad \forall x \in X. \tag{41}$$

Furthermore, the strategies $\bar{\pi}_*^i = \left\{ \varphi_t^i \right\}$, $i = 1, 2$, where

$$\varphi_t^i(\cdot|x) = \int_\Omega \bar{\varphi}_t^i(\cdot|x, \omega) P(\mathrm{d}\omega),$$

form an average optimal (non random) pair of strategies.

The proof of Theorem 5.3 is based in the following facts. For each $t \in \mathbb{N}_0$ (see Remark 3.5), we define

$$\begin{aligned} \gamma_t &\equiv \gamma_{\bar{\alpha}_t} := \frac{1 + \bar{\alpha}_t}{2} \in (\bar{\alpha}_t, 1), \\ e_t &: = d\left( \frac{\gamma_t}{\bar{\alpha}_t} - 1 \right)^{-1} = d\left( \frac{2\bar{\alpha}_t}{1 - \bar{\alpha}_t} \right), \end{aligned}$$

and

$$l_t \equiv l_{\bar{\alpha}_t} := 1 + e_t = 1 + \frac{2d\bar{\alpha}_t}{1 - \bar{\alpha}_t}.$$

It is easy to see that

$$\frac{l_t}{1 - \gamma_t} \leq 2(1 + d)(1 - \bar{\alpha}_t)^{-2},$$

which, from Condition D.1, yields

$$\frac{l_t}{1 - \gamma_t} = \mathbf{O}(t^{2\nu}), \text{ as } t \to \infty. \tag{42}$$

Moreover, applying similar arguments as the proof of Theorem 4.5 (see (29)) and from definition of the function $\phi_\alpha^\theta$ (see (17)), we can obtain

$$\left\| V_{\bar{\alpha}_t}^{\theta_t} - V_{\bar{\alpha}_t}^{\theta} \right\|_W \leq \frac{l_t}{1 - \gamma_t} \bar{\Delta}_t.$$

Hence, for all $(\pi^1,\pi^2)\in\Pi^1\times\Pi^2$ and $x\in X$, from (38) and (42),

$$E_x^{\pi^1,\pi^2}\left\|V_{\bar{\alpha}_t}^{\theta_t}-V_{\bar{\alpha}_t}^{\theta}\right\|_W=\mathbf{O}(t^{2\nu})\mathbf{O}(t^{-1/m}),\quad\text{as}\quad t\to\infty. \tag{43}$$

Then, because $2\nu<1/m$, we get

$$\lim_{t\to\infty}E_x^{\pi^1,\pi^2}\left\|V_{\bar{\alpha}_t}^{\theta_t}-V_{\bar{\alpha}_t}^{\theta}\right\|_W=0.$$

Again, from definition of the functions $\phi_\alpha^\theta(x)$ and $j_\alpha^\theta$ (see (17)), we have

$$\lim_{t\to\infty}E_x^{\pi^1,\pi^2}\left\|\phi_{\bar{\alpha}_t}^{\theta_t}-\phi_{\bar{\alpha}_t}^{\theta}\right\|_W=0 \tag{44}$$

and

$$\lim_{t\to\infty}E_x^{\pi^1,\pi^2}\left\|j_{\bar{\alpha}_t}^{\theta_t}-j_{\bar{\alpha}_t}^{\theta}\right\|_W=0. \tag{45}$$

On the other hand, following similar ideas as the proof of [11, relation (35)] and once the necessary changes have been made, we obtain

$$\lim_{t\to\infty}E_x^{\pi^1,\pi^2}\left\|\phi_{\bar{\alpha}_t}^{\theta_t}-\phi_{\bar{\alpha}_t}^{\theta}\right\|_W W(x_t)=0 \tag{46}$$

and

$$\lim_{t\to\infty}E_x^{\pi^1,\pi^2}\bar{\Delta}_tW(x_t)=0. \tag{47}$$

## 5.1. Proof of the Theorem 5.3

We first prove the optimality of $\pi_*^2=\left\{\bar{\varphi}_t^2(\cdot|x,\omega)\right\}=\left\{\bar{\varphi}_t^2\right\}$, for which we will show

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

Let $\pi^1=\left\{\pi_t^1\right\}\in\Pi^1$ be an arbitrary strategy for player 1. Then

$$\begin{aligned}\mathcal{L}_t&:=r(x_t,\pi_t^1,\bar{\varphi}_t^2)+\bar{\alpha}_t\int_{\Re^k}\phi_{\bar{\alpha}_t}^\theta[F(x_t,\pi_t^1,\bar{\varphi}_t^2,s)]\theta(\mathrm{d}s)-j_{\bar{\alpha}_t}^\theta-\phi_{\bar{\alpha}_t}^\theta(x_t)\\&=r(x_t,\pi_t^1,\bar{\varphi}_t^2)+\bar{\alpha}_tE_x^{\pi^1,\pi_*^2}\left[\phi_{\bar{\alpha}_t}^\theta(x_{t+1})\mid h_t\right]-j_{\bar{\alpha}_t}^\theta-\phi_{\bar{\alpha}_t}^\theta(x_t),\end{aligned} \tag{48}$$

which implies

$$\begin{aligned}n^{-1}E_x^{\pi^1,\pi_*^2}\left[\sum_{t=0}^{n-1}\left(r(x_t,a_t,b_t)-j_{\bar{\alpha}_t}^\theta\right)\right]&=n^{-1}E_x^{\pi^1,\pi_*^2}\left[\sum_{t=0}^{n-1}\left(\phi_{\bar{\alpha}_t}^\theta(x_t)-\bar{\alpha}_t\phi_{\bar{\alpha}_t}^\theta(x_{t+1})\right)\right]\\&\quad+n^{-1}E_x^{\pi^1,\pi_*^2}\left[\sum_{t=0}^{n-1}\mathcal{L}_t\right].\end{aligned}$$

Hence, from (7) and (19)

$$\begin{aligned} J(x,\pi^1,\pi^2_*)-j^* &= \liminf_{n\to\infty}\left\{n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right]\right. \\ &\qquad \left.+n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\mathcal{L}_t\right]\right\}. \end{aligned} \tag{49}$$

Therefore, the remainder of the proof consists in to prove

$$\liminf_{n\to\infty}\left\{n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right]+n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\mathcal{L}_t\right]\right\}\le 0. \tag{50}$$

Observe that Condition D.2 implies that $\{\bar{\alpha}_t\}$ remains constant for long time periods. Then, for $n\ge l\ge 1$,

$$\begin{aligned} &n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right] \\ &= n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{l-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right] \\ &+ n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=l}^{n-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right] \\ &\le n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{l-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right] \\ &+ (1-\bar{\alpha}_l)M_0+n^{-1}M_0\sum_{i=1}^{\kappa(n)}\alpha_i^* \\ &\le (1-\bar{\alpha}_l)M_0+M_0\kappa(n)n^{-1}, \end{aligned} \tag{51}$$

where $\alpha_1^*,\alpha_2^*,\ldots,\alpha^*_{\kappa(n)}$ are the different values of $\bar{\alpha}_t$ for $t\le n$, and $M_0$ is a constant such that $E_x^{\pi^1,\pi^2_*}\left|\phi^\theta_\alpha(x_{t+1})\right|<M_0\ \forall\alpha\in(0,1)$ (see (20)). Then, because $l$ is arbitrary and $\bar{\alpha}_t\nearrow 1$, from (51) and Condition D.2 we get

$$\lim_{n\to\infty}n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\left(\phi^\theta_{\bar{\alpha}_t}(x_t)-\bar{\alpha}_t\phi^\theta_{\bar{\alpha}_t}(x_{t+1})\right)\right]=0. \tag{52}$$

Now, we will proceed to prove

$$\lim_{n\to\infty}n^{-1}E_x^{\pi^1,\pi^2_*}\left[\sum_{t=0}^{n-1}\mathcal{L}_t\right]\le 0. \tag{53}$$

To this end, we will prove

$$\limsup_{t\to\infty}E_x^{\pi^1,\pi^2_*}\left[\mathcal{L}_t\right]\le 0.$$

Observe that from (39), for each $t \in \mathbb{N}_0$,

$$\left| \int_{\Re^k} \phi^{\theta}_{\bar{\alpha}_t}(F(x,a,b,s))\theta_t(\mathrm{d}s) - \int_{\Re^k} \phi^{\theta}_{\bar{\alpha}_t}(F(x,a,b,s))\theta(\mathrm{d}s) \right| \leq \bar{\Delta}_t W(x). \tag{54}$$

Hence, adding and subtracting the terms

$$\bar{\alpha}_t \int_{\Re^k} \phi^{\theta}_{\bar{\alpha}_t}[F(x_t,\pi^1_t,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s) \quad \text{and} \quad \bar{\alpha}_t \int_{\Re^k} \phi^{\theta_t}_{\bar{\alpha}_t}[F(x_t,\pi^1_t,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s)$$

we get

$$\mathcal{L}_t \leq \bar{\Delta}_t W(x_t) + \mathcal{L}^0_t + \mathcal{L}^1_t, \tag{55}$$

where

$$\begin{aligned} \mathcal{L}^0_t \quad &: \; = \left| \int_{\Re^k} \phi^{\theta}_{\bar{\alpha}_t}[F(x_t,\pi^1_t,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s) - \int_{\Re^k} \phi^{\theta_t}_{\bar{\alpha}_t}[F(x_t,\pi^1_t,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s) \right|, \\ \mathcal{L}^1_t \quad &: \; = r(x_t,\pi^1_t,\bar{\varphi}^2_t) + \bar{\alpha}_t \int_{\Re^k} \phi^{\theta_t}_{\bar{\alpha}_t}[F(x_t,\pi^1_t,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s) - j^{\theta}_{\bar{\alpha}_t} - \phi^{\theta}_{\bar{\alpha}_t}(x_t). \end{aligned}$$

Note that $\mathcal{L}^0_t \leq \left\| \phi^{\theta_t}_{\bar{\alpha}_t} - \phi^{\theta}_{\bar{\alpha}_t} \right\|_W$, and therefore, from (44),

$$\lim_{t\to\infty} E^{\pi^1,\pi^2_*}_x \mathcal{L}^0_t = 0. \tag{56}$$

For $\mathcal{L}^1_t$, adding and subtracting $j^{\theta_t}_{\bar{\alpha}_t}$ and $\phi^{\theta_t}_{\bar{\alpha}_t}(x_t)$, from the definition of $\bar{\varphi}^2_t$ (see (40))

$$\begin{aligned} \mathcal{L}^1_t \quad &\leq \quad \max_{\varphi^1 \in \mathbb{A}(x)} \left[ r(x_t,\varphi^1,\bar{\varphi}^2_t) + \bar{\alpha}_t \int_S \phi^{\theta_t}_{\bar{\alpha}_t}[F(x_t,\varphi^1,\bar{\varphi}^2_t,s)]\theta_t(\mathrm{d}s) \right] - j^{\theta_t}_{\bar{\alpha}_t} - \phi^{\theta_t}_{\bar{\alpha}_t}(x_t) \\ &\quad\; + j^{\theta_t}_{\bar{\alpha}_t} - j^{\theta}_{\bar{\alpha}_t} + \phi^{\theta_t}_{\bar{\alpha}_t}(x_t) - \phi^{\theta}_{\bar{\alpha}_t}(x_t) \\ &\leq \quad \left| j^{\theta_t}_{\bar{\alpha}_t} - j^{\theta}_{\bar{\alpha}_t} \right| + \left\| \phi^{\theta_t}_{\bar{\alpha}_t} - \phi^{\theta}_{\bar{\alpha}_t} \right\|_W W(x_t). \end{aligned} \tag{57}$$

Thus, (45) and (46) implies

$$\limsup_{t\to\infty} E^{\pi^1,\pi^2_*}_x \mathcal{L}^1_t \leq 0. \tag{58}$$

Combining (47), (55), (56), and (58) we get (53), which, together with (52), yields (50). Thus, from (49)

$$J(x,\pi^1,\pi^2_*) \leq j^*, \qquad x \in X.$$

Finally, since $\pi^1 \in \Pi^1$ is arbitrary, from Theorem 3.7,

$$j^* = \sup_{\pi^1 \in \Pi^1} J(x,\pi^1,\pi^2_*), \quad \forall x \in X.$$

The optimality of $\pi^1_*$ is proved similarly.

Finally, the average optimality of the pair $(\bar{\pi}^1_*,\bar{\pi}^2_*) \in \Pi^1 \times \Pi^2$ is proved following similar arguments as part (c) of Theorem 4.5. □

## 6. CONCLUDING REMARKS

Our results are based on two class of conditions. The first one, composed by Assumptions 3.1 and 3.2, contains standard mild requirements ensuring the existence of values as well as optimal pairs of strategies for the discounted and average games. These assumptions are an adaptation from those widely used to study Markov control processes (MCPs) with possibly unbounded costs (see, e. g., [8, 9, 11, 15]). In particular, observe that the continuity of the function $F$ required in Assumption 3.1(c) implies that the stochastic kernel defined in (3) is weakly continuous.

The another class of conditions is the related with the empirical procedures. Indeed, the empirical approximation-estimation processes introduced in this paper are strongly based on the equicontinuity and equi-Lipschitz conditions for the discounted and average criteria respectively. Such conditions have been used in several contexts within the field of MCPs. For instance (see [10] and references therein), under equicontinuity conditions it is possible to show the existence of solutions of optimality equations, as limit of a sequence of functions, by using Ascoli's theorem. In our case, equicontinuity is applied in order to obtain the convergence of empirical procedures given in Remark 4.4 (see Assumption 4.3). Clearly, if the disturbance space $S$ is countable, i. e., if the disturbance process $\{\xi_t\}$ is formed by discrete random variables, the equicontinuity with respect to the discrete topology is trivially satisfied .

On the other hand, taking into account that function $V_\alpha^\theta/W$ is uniformly bounded (see Theorem 3.6(a)), ask for the convexity of function $s \longmapsto \phi_\alpha^\theta(F(x,a,b,s))/W(x)$ is a sufficient condition for the equi-Lipschitz Assumption 5.1(c), which in turn implies the equicontinuity Assumption 4.3. Thus, in the case of dealing with real random variables, that is by taking $S = \Re$, the convexity could be more easily handled, namely, by means of its derivative. Moreover, by imposing convexity on components of game model is possible to obtain convexity properties of the value of the game. This issue is part of a future work of the authors.

## ACKNOWLEDGEMENT

This work was partially supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) under grant CB2015/254306.

(Received October 14, 2016)

## REFERENCES

[1] H. S. Chang: Perfect information two-person zero-sum Markov games with imprecise transition probabilities. Math. Meth. Oper. Res. *64* (2006), 235–351. DOI:10.1007/s00186-006-0081-5

[2] R. M. Dudley: The speed of mean Glivenko-Cantelli convergence. Ann. Math. Stat. *40* (1969), 40–50. DOI:10.1214/aoms/1177697802

[3] E. B. Dynkin and A. A. Yushkevich: Controlled Markov Processes. Springer–Verlag, New York 1979. DOI:10.1007/978-1-4615-6746-2

[4] E. Fernández-Gaucherand : A note on the Ross–Taylor Theorem. Appl. Math. Comp. *64* (1994), 207–212. DOI:10.1016/0096-3003(94)90064-7

[5] J. Filar and K. Vrieze: Competitive Markov Decision Processes. Springer–Verlag, New York 1997. DOI:10.1007/978-1-4612-4054-9

[6] M. K. Ghosh, D. McDonald, and S. Sinha: Zero-sum stochastic games with partial information. J. Optim. Theory Appl. *121* (2004), 99–118. DOI:10.1023/b:jota.0000026133.56615.cf

[7] E. I. Gordienko: Adaptive strategies for certain classes of controlled Markov processes. Theory Probab. Appl. *29* (1985), 504–518. DOI:10.1137/1129064

[8] E. I. Gordienko and O. Hernández-Lerma: Average cost Markov control processes with weighted norms: existence of canonical policies. Appl. Math. *23* (1995), 199–218.

[9] E. I. Gordienko and O. Hernández-Lerma: Average cost Markov control processes with weighted norms: value iteration. Appl. Math. *23* (1995), 219–237.

[10] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer–Verlag, New York 1996. DOI:10.1007/978-1-4612-0729-0

[11] N. Hilgert and J. A. Minjárez-Sosa: Adaptive control of stochastic systems with unknown disturbance distribution: discounted criterion. Math. Meth. Oper. Res. *63* (2006), 443–460. DOI:10.1007/s00186-005-0024-6

[12] A. Jaśkiewicz and A. Nowak: Zero-sum ergodic stochastic games with Feller transition probabilities. SIAM J. Control Optim. *45* (2006), 773–789. DOI:10.1137/s0363012904443257

[13] A. Jaśkiewicz and A. Nowak: Approximation of noncooperative semi-Markov games. J. Optim. Theory Appl. *131* (2006), 115–134. DOI:10.1007/s10957-006-9128-2

[14] A. Krausz and U. Rieder: Markov games with incomplete information. Math. Meth. Oper. Res. *46* (1997), 263–279. DOI:10.1007/bf01217695

[15] J. A. Minjárez-Sosa: Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under average criterion. Appl. Math. (Warsaw) *26* (1999), 267–280.

[16] J. A. Minjárez-Sosa and O. Vega-Amaya: Asymptotically optimal strategies for adaptive zero-sum discounted Markov games. SIAM J. Control Optim. *48* (2009), 1405–1421. DOI:10.1137/060651458

[17] J. A. Minjárez-Sosa and O. Vega-Amaya: Optimal strategies for adaptive zero-sum average Markov games. J. Math. Analysis Appl. *402* (2013), 44–56. DOI:10.1016/j.jmaa.2012.12.011

[18] J. A. Minjárez-Sosa and F. Luque-Vásquez: Two person zero-sum semi-Markov games with unknown holding times distribution on one side: discounted payoff criterion. Appl. Math. Optim. *57* (2008), 289–305. DOI:10.1007/s00245-007-9016-7

[19] A. Neyman and S. Sorin: Stochastic Games and Applications. Kluwer, 2003. DOI:10.1007/978-94-010-0189-2

[20] T. Prieto-Rumeau and J. M. Lorenzo: Approximation of zero-sum continuous-time Markov games under the discounted payoff criterion. TOP *23* (2015), 799–836. DOI:10.1007/s11750-014-0354-8

[21] N. Shimkin and A. Shwartz: Asymptotically efficient adaptive strategies in repeated games. Part I: Certainty equivalence strategies. Math. Oper. Res. *20* (1995), 743–767. DOI:10.1287/moor.20.3.743

[22] N. Shimkin and A. Shwartz: Asymptotically efficient adaptive strategies in repeated games. Part II: Asymptotic optimality. Math. Oper. Res. *21* (1996), 487–512. DOI:10.1287/moor.21.2.487

[23] M. Schäl: Conditions for optimality and for the limit of $n$-stage optimal policies to be optimal. Z. Wahrs. Verw. Gerb. *32* (1975), 179–196. DOI:10.1007/bf00532612

[24] R. Ranga Rao: Relations between weak and uniform convergence of measures with applications. Ann. Math. Statist. *33* (1962), 659–680. DOI:10.1214/aoms/1177704588

[25] J. A. E. E. Van Nunen and J. Wessels: A note on dynamic programming with unbounded rewards. Manag. Sci. *24* (1978), 576–580. DOI:10.1287/mnsc.24.5.576

*Fernando Luque-Vásquez, Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Col. Centro, 83000 Hermosillo, Sonora. Mexico.*
*e-mail: fluque@gauss.mat.uson.mx*

*J. Adolfo Minjárez-Sosa, Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Col. Centro, 83000 Hermosillo, Sonora. Mexico.*
*e-mail: aminjare@gauss.mat.uson.mx*