Policy Iteration Algorithms for Zero-Sum Stochastic Differential Games with Long-Run Average Payoff Criteria

José Daniel López-Barrientos

Received: 7 June 2014 / Revised: 4 October 2014 / Accepted: 8 November 2014 /

Published online: 28 November 2014

© Operations Research Society of China, Periodicals Agency of Shanghai University, and Springer-Verlag Berlin Heidelberg 2014

Abstract This paper studies the policy iteration algorithm (PIA) for zero-sum stochastic differential games with the basic long-run average criterion, as well as with its more selective version, the so-called bias criterion. The system is assumed to be a nondegenerate diffusion. We use Lyapunov-like stability conditions that ensure the existence and boundedness of the solution to certain Poisson equation. We also ensure the convergence of a sequence of such solutions, of the corresponding sequence of policies, and, ultimately, of the PIA.

Keywords Ergodic payoff criterion · Zero-sum stochastic differential games · Policy iteration algorithm · Nondegenerate diffusions · Poisson equation · Schäl convergence · Bias game

Mathematics Subject Classification 91A15 · 91A10 · 91A25

1 Introduction

This paper is about two-person zero-sum stochastic differential games (SDGs) with ergodic payoffs on an infinite horizon. This sort of games has been studied, for instance in [10, 12, 26]. Our aim is to give conditions under which the policy iteration algorithm (PIA) produces convergent sequences of values and policies for an SDG with ergodic payoffs.

The motivation for our developments lies in the fact that the PIA can be an alternative theoretical tool to the classic Kakutani–Glicksberg–Fan fixed point

J. D. López-Barrientos (✉)

Faculty of Actuarial Sciences, Universidad Anáhuac, México Norte, Huixquilucan, Mexico
e-mail: actdaniel@hotmail.com; daniel.lopez@anahuac.mx

theorem for proving the existence of Nash equilibria in the context of *nonzero-sum* SDGs. Indeed, while studying the problem of the existence of saddle points in the class of nonzero-sum games, it turns out that Theorem 1 in [13] does not necessarily hold for infinite-dimensional spaces, hence it becomes necessary to impose additional conditions on the system (for instance, a separability property—cf. [11]) to ensure the existence of a Nash equilibrium. This work intends to avoid imposing such conditions, and to provide insights on the PIA for *zero-sum* SDGs under two long-run average criteria. Our results and developments can be thought of as an introduction of a useful technique for the more general problem of a nonzero-sum game.

The PIA was first introduced by Howard [21]. It was used later by Fleming [14] to study some finite horizon control problems in 1963. Bismut [8] and Puterman [35, 36] studied its convergence rate. In two-person zero-sum games, Van der Wal [41] presented a convergent version of the algorithm that works under the assumption that both, the state space and the action space are finite. More recently, Zhu et al. developed convergent versions of the PIA for the basic average payoff criterion for Markov decision processes in [43, 45]. The goal of the PIA is to generate sequences of policies and value functions for a *control problem* that converge to the optimal control and value function, respectively.

Our work can be thought of as a bridge between the theory presented in the references quoted above, and the problem of finding an equilibrium for an SDG under ergodic criteria without necessarily (i) departing from a discounted payoff problem and (ii) using the well-known vanishing discount technique. Moreover, we provide sufficient conditions for the *convergence* of the algorithm when both players choose their respective actions independently from each other. These steps are customary, at least for ergodic control problems (see, for example, [4, Chap. II, 9, Corollary 6.2, 30]).

Our set of assumptions ensure (i) the existence of the value of certain zero-sum SDG with average payoff and (ii) the convergence of the PIA to a saddle point of such game. To these ends, we will use some results from the theory of second order partial differential equations [17] and, for a given pair of strategies, we will use the concept of its bias from the game's value [32, 33] (see also [42, 44] for more insights on bias problems). We will present a standard version of the PIA and a slight modification that is proven to work in the bias game as well.

The algorithm we present resembles the one introduced in [19] and [43, 45] for controlled Markov decision processes in Borel spaces, and is inspired in the Hoffman–Karp's (HK) version presented in [41]. In the latter algorithm, the first step is to fix the action of one of the players to find the other player's *best* action, thus reducing the game in that stage to a Markov decision process. Then one finds the *current* value of the game and then moves on to the next iteration, where the other player's *best* action is fixed. There are two main differences between the HK algorithm presented in [41], and our work:

- In [41], it is supposed that both the action and the set spaces are finite; we consider them to be uncountable.

- The payoff criterion used in [41] is that of the discounted reward. We consider two more specialized criteria: long-run average and bias optimality.

It should be noted that the point of departure of our presentation is the existence of a differentiable solution of the so-called *Poisson equation*. We have given sufficient conditions and a proof of this fact in our previous works (see [24, Theorem 5.4], and [27, Theorem 3.4]).

The paper is organized as follows. Section 2 presents the basic hypotheses on the system, the type of strategies we will use, the reward rate, and the basic average payoff criterion for zero-sum SDGs. Section 3 introduces the algorithm we are interested in and quotes a result on the existence of the solution to the dynamic programming approach to the ergodic criterion problem. In Sect. 4 we present: (i) the policy convergence form we are interested in, and (ii) our main results: Lemma 4.1 and Theorem 4.3. The bias game and an extension of the PIA to this context is in Sect. 5. We present an example of the PIA for the average criterion and for the bias game in Sect. 6. Finally, we give some concluding remarks in Sect. 7.

Throughout the following, for vectors x and matrices A , we use the norms

$$|x|^2 := \sum_i x_i^2 \quad \text{and} \quad |A|^2 := \text{Tr}(AA') = \sum_{i,j} A_{ij}^2,$$

where A' and $\text{Tr}(\cdot)$ denote the transpose of $A = (A_{ij})$ and the trace of a square matrix, respectively. Moreover, if $h : \mathbb{R}^N \rightarrow \mathbb{R}$ is a smooth function, then ∇h and $\mathbb{H}h$ represent the gradient vector of h (i.e., the vector of partial derivatives of h with respect to x_i for $i = 1, \dots, N$) and the Hessian matrix of h , i.e., $\mathbb{H}h = \left(\partial_{x_i x_j}^2 h \right)$, respectively.

2 Preliminaries

This section introduces the SDG we are concerned with, as well as some important underlying concepts.

2.1 Dynamics of the System

Let U^1 and U^2 be compact subsets of some complete and separable vector normed spaces. We consider the N -dimensional process $x(\cdot)$ defined, for all $t \geq 0$ by

$$dx(t) = b(x(t), u_1(t), u_2(t))dt + \sigma(x(t))dW(t) \quad (2.1)$$

with initial condition $x(0) = x$, where $b : \mathbb{R}^N \times U^1 \times U^2 \rightarrow \mathbb{R}^N$ and $\sigma : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times m}$ are given functions, and $W(\cdot)$ is an m -dimensional Wiener process. The sets U^1 and U^2 are called control (or action) sets and, for $\ell = 1, 2$, $\{u_\ell(t) : t \geq 0\}$ is a U^ℓ -valued stochastic process representing the action that player ℓ takes at each time $t \geq 0$.

Assumption 2.1

- (a) The function b is continuous on $\mathbb{R}^N \times U^1 \times U^2$ and there exists a positive constant C_1 such that, for each x and y in $\mathbb{R}^N$,

$$\sup_{(u_1, u_2) \in U^1 \times U^2} |b(x, u_1, u_2) - b(y, u_1, u_2)| \leq C_1 |x - y|.$$

- (b) There exist positive constants C_2 and γ such that for each x and y in $\mathbb{R}^N$,

$$|\sigma(x) - \sigma(y)| \leq C_2 |x - y|,$$

and the matrix $a := \sigma\sigma'$ satisfies that

$$x'a(y)x \geq \gamma |x|^2 \text{ (uniform ellipticity)}. \quad (2.2)$$

Remark 2.2 The Lipschitz conditions on b and σ in Assumption 2.1 imply that b and σ satisfy a *linear growth* condition. That is, there exists a constant $\tilde{C} \geq C_1 + C_2$ such that for all $x \in \mathbb{R}^N$,

$$\sup_{(u_1, u_2) \in U^1 \times U^2} |b(x, u_1, u_2)| + |\sigma(x)| \leq \tilde{C}(1 + |x|).$$

Moreover, the uniform ellipticity condition (2.2) enables us to deal with the average optimality equations (2.31)–(2.33) below with the aid of the results in [17].

Definition 2.3 The space $C^\kappa(\mathbb{R}^N)$ consists of all real-valued continuous functions on $\mathbb{R}^N$ with continuous l -th partial derivative in $x_i \in \mathbb{R}$, for $i = 1, \dots, N$, $l = 0, 1, \dots, \kappa$.

For (u_1, u_2) in $U^1 \times U^2$ and h in $C^2(\mathbb{R}^N)$, let

$$\begin{aligned} \mathbb{L}^{u_1, u_2} h(x) &:= \langle \nabla h(x), b(x, u_1, u_2) \rangle + \frac{1}{2} \text{Tr}[\mathbb{H}h(x) \cdot a(x)] \\ &= \sum_{i=1}^N \partial_{x_i} h(x) b_i(x, u_1, u_2) + \frac{1}{2} \sum_{i,j=1}^n \partial_{x_i x_j}^2 h(x) a_{ij}(x), \end{aligned} \quad (2.3)$$

where $a(\cdot)$ is as in Assumption 2.1(b).

2.2 Strategies

Let U^ℓ be the set of admissible actions for player $\ell = 1, 2$, and assume that U^ℓ is compact. Denote by V^ℓ the space of probability measures on U^ℓ endowed with the topology of weak convergence. This notion allows us to think of V^ℓ as a compact metric space. (See [5, Chap. 7.4], or [7, Chap. 1] for reference.)

For our present purposes, we can restrict ourselves to consider only randomized stationary strategies. A randomized stationary strategy $\pi^1 \in V^1$ for player 1 is a stochastic kernel on U^1 given $\mathbb{R}^N$, i.e., for each Borel subset D of U^1 , $\pi^1(D|\cdot)$ is a

Borel function on $\mathbb{R}^N$, and for each $x \in \mathbb{R}^N$, $\pi^1(\cdot|x)$ is a probability measure on U^1 . The family of all randomized Markov strategies for player $\ell = 1, 2$ is denoted as Π^ℓ .

When using randomized Markov policies (π^1, π^2) in $\Pi^1 \times \Pi^2$, we will write, for $x(t) \in \mathbb{R}^N$ and $t \geq 0$, as in (2.1),

$$b(x(t), \pi^1, \pi^2) := \int_{U^2} \int_{U^1} b(x(t), u_1, u_2) \pi^1(du_1|x) \pi^2(du_2|x). \quad (2.4)$$

For $(\varphi, \psi) \in V^1 \times V^2$, we also introduce the notation

$$b(x(t), \varphi, \psi) := \int_{U^2} \int_{U^1} b(x(t), u_1, u_2) \varphi(du_1) \psi(du_2). \quad (2.5)$$

Moreover, recalling (2.3), for $h \in \mathcal{C}^2(\mathbb{R}^N)$, let

$$\mathbb{L}^{\pi^1, \pi^2} h(x(t)) := \int_{U^2} \int_{U^1} \mathbb{L}^{u_1, u_2} h(x(t)) \pi^1(du_1|x) \pi^2(du_2|x).$$

We also use

$$\mathbb{L}^{\varphi, \psi} h(x(t)) := \int_{U^2} \int_{U^1} \mathbb{L}^{u_1, u_2} h(x(t)) \varphi(du_1) \psi(du_2),$$

for $(\varphi, \psi) \in V^1 \times V^2$.

Remark 2.4 A direct calculation yields that $b(\cdot, \varphi, \psi)$ defined in (2.5), has the corresponding Lipschitz property in Assumption 2.1(a), that is, there exists a constant C_1 such that

$$\sup_{(\varphi, \psi) \in V^1 \times V^2} |b(x, \varphi, \psi) - b(y, \varphi, \psi)| \leq C_1 |x - y|$$

for all $x, y \in \mathbb{R}^N$. Moreover, the Lipschitz conditions on b and σ in Assumption 2.1(a)–(b), along with the compactness of V^1 and V^2 yield that there exists a constant $\tilde{C} \geq C_1 + C_2$ such that

$$\sup_{(\varphi, \psi) \in V^1 \times V^2} |b(x, \varphi, \psi)| + |\sigma(x)| \leq \tilde{C}(1 + |x|)$$

for all $x \in \mathbb{R}^N$.

Assumption 2.1 and Remark 2.4 ensure that, for each pair (π^1, π^2) in $\Pi^1 \times \Pi^2$, the system (2.1) admits an almost surely strong solution $x(\cdot) := \{x(t) : t \geq 0\}$, which is a Feller process whose generator coincides with the operator $\mathbb{L}^{\pi^1, \pi^2} h$ in (2.2). For more details, see [15, Theorem 2.1] and [37, Chap. III. 2]. To emphasize the dependence on $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, sometimes we write $x(\cdot)$ as $x^{\pi^1, \pi^2}(\cdot)$. Also, the corresponding transition probability is

$$\mathbb{P}_x^{\pi^1, \pi^2}(t, B) := \mathbb{P}\left(x^{\pi^1, \pi^2}(t) \in B | x^{\pi^1, \pi^2}(0) = x\right)$$

for every Borel set $B \subset \mathbb{R}^N$ and $t \geq 0$. The associated conditional expectation is written as $\mathbb{E}_x^{\pi^1, \pi^2}(\cdot)$.

2.3 Ergodicity Assumptions

The following assumption is a standard Lyapunov stability condition for continuous time (controlled and uncontrolled) Markov processes (see, for instance [16, Condition C1 in p. 1967], [29, Condition(CD0)], [45, Lemma 2.3], or [6]).

Assumption 2.5 There exists a function $w \geq 1$ in $\mathcal{C}^2(\mathbb{R}^N)$ and constants $d \geq c > 0$ such that

- (a) $\lim_{|x| \rightarrow \infty} w(x) = \infty$.
- (b) $\mathbb{L}^{u_1, u_2} w(x) \leq -cw(x) + d$ for all (u_1, u_2) in $U^1 \times U^2$ and x in $\mathbb{R}^N$.

Assumption 2.5 gives that, for each $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, the Markov process $x^{\pi^1, \pi^2}(t)$, $t \geq 0$, is Harris positive recurrent with a unique invariant probability measure μ_{π^1, π^2} for which the integral

$$\mu_{\pi^1, \pi^2}(w) := \int_{\mathbb{R}^N} w(x) \mu_{\pi^1, \pi^2}(dx) \quad (2.6)$$

is finite (see [16, Theorem 3.1, 18, Theorem 2.2, 29, Sect. 4]).

By Theorem 4.3 of [2], for each pair (π^1, π^2) in $\Pi^1 \times \Pi^2$, the probability measures μ_{π^1, π^2} and $\mathbb{P}_x^{\pi^1, \pi^2}(t, \cdot)$ are both equivalent to Lebesgue's measure λ for every $t \geq 0$ and $x \in \mathbb{R}^N$. Hence there exists a transition density function $p^{\pi^1, \pi^2}(x, t, y)$ such that

$$\mathbb{P}_x^{\pi^1, \pi^2}(t, B) = \int_B p^{\pi^1, \pi^2}(x, t, y) dy \quad (2.7)$$

for every Borel set $B \subset \mathbb{R}^N$.

Dynkin's formula (see [25, Corollary 6.5]) and, again, Assumption 2.5 ensure the *boundedness* of $\mathbb{E}_x^{\pi^1, \pi^2}[w(x(t))]$ in the sense of the following result. The proof is straightforward (see, for instance, [23, Lemma 2.10] or [29, Theorem 2.1 (iii)]).

Lemma 2.6 *The condition (b) in Assumption 2.5 implies that*

$$\mathbb{E}_x^{\pi^1, \pi^2}[w(x(t))] \leq e^{-ct} w(x) + \frac{d}{c} (1 - e^{-ct}) \quad (2.8)$$

for every (π^1, π^2) in $\Pi^1 \times \Pi^2$, $t \geq 0$, and $x \in \mathbb{R}^N$.

We now introduce the concept of the w -weighted norm, where w is the function in Assumption 2.5.

Definition 2.7 Let $\mathcal{B}_w(\mathbb{R}^N)$ denote the Banach space of real-valued measurable functions v on $\mathbb{R}^N$ with finite w -norm, which is defined as

$$\|v\|_w := \sup_{x \in \mathbb{R}^N} \frac{|v(x)|}{w(x)}.$$

Moreover, $\mathbb{M}_w(\mathbb{R}^N)$ stands for the normed linear space of finite signed measures μ on $\mathbb{R}^N$ such that

$$\|\mu\|_w := \int_{\mathbb{R}^N} w(x) d\|\mu\|_{TV} < \infty,$$

where $\|\mu\|_{TV} := \mu^+ + \mu^-$ denotes the *total variation* of μ .

By (2.6), μ_{π^1, π^2} belongs to $\mathbb{M}_w(\mathbb{R}^N)$ for every $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$. In addition, for each $v \in \mathcal{B}_w(\mathbb{R}^N)$, letting $\mu_{\pi^1, \pi^2}(v) := \int v d\mu_{\pi^1, \pi^2}$, we get

$$|\mu_{\pi^1, \pi^2}(v)| \leq \|v\|_w \int_{\mathbb{R}^N} w d\|\mu_{\pi^1, \pi^2}\|_{TV} = \|v\|_w \|\mu_{\pi^1, \pi^2}\|_w < \infty. \quad (2.9)$$

Let T be a positive constant. For (π^1, π^2) fixed, define the T -skeleton chain of $x^{\pi^1, \pi^2}(\cdot)$ by:

$$x_T^{\pi^1, \pi^2} := \{x^{\pi^1, \pi^2}(kT) : k = 0, 1, \dots\}. \quad (2.10)$$

Let $Q_m^{\pi^1, \pi^2}(x, \cdot)$ be the m -step transition probability of $x_T^{\pi^1, \pi^2}$, defined as

$$Q_m^{\pi^1, \pi^2}(x, B) := \mathbb{P}_x^{\pi^1, \pi^2}(mT, B), \quad B \subseteq \mathbb{R}^N,$$

with $\mathbb{P}_x^{\pi^1, \pi^2}$ as in (2.7).

Let us impose now the following condition on $x_T^{\pi^1, \pi^2}$ (see [43, Assumption C]).

Assumption 2.8 The skeleton chain (2.10) is uniformly w -exponentially ergodic. That is, there exist positive constants $\rho_1 < 1$ and ρ_2 such that, for all $m \geq 1$,

$$\sup_{(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2} \left\| Q_m^{\pi^1, \pi^2}(x, \cdot) - \mu_{\pi^1, \pi^2}(\cdot) \right\|_w \leq \rho_2 \rho_1^m w(x). \quad (2.11)$$

Sufficient conditions for this Assumption are given, for instance, in Assumption 4.1 and Lemma 4.8 of [28].

The proof of the following result is based on those given in Jasso-Fuentes [22] and Jasso-Fuentes and Hernández-Lerma [23, Theorem 2.7].

Theorem 2.9 Suppose that Assumptions 2.1, 2.5 and 2.8 hold. Then the process $x^{\pi^1, \pi^2}(\cdot)$ is uniformly w -exponentially ergodic, that is, there exist constants $C, \delta > 0$ such that

$$\sup_{(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2} \left| \mathbb{E}_x^{\pi^1, \pi^2} v(x(t)) - \mu_{\pi^1, \pi^2}(v) \right| \leq C e^{-\delta t} \|v\|_w w(x) \quad (2.12)$$

for all $x \in \mathbb{R}^N$, $t \geq 0$, and $v \in \mathcal{B}_w(\mathbb{R}^N)$. In (2.12), $\mu_{\pi^1, \pi^2}(v)$ is defined as in (2.6), with v in lieu of w .

Proof Fix $T > 0$ and note that any $t > 0$ can be expressed in terms of T as $t = mT + s$ for some $m = 0, 1, \dots$, and $s \in [0, T]$. Hence, for every $x \in \mathbb{R}^N$, $v \in \mathcal{B}_w(\mathbb{R}^N)$ and $t \geq 0$ we have

$$\begin{aligned} \left| \mathbb{E}_x^{\pi^1, \pi^2} v(x(t)) - \mu_{\pi^1, \pi^2}(v) \right| &= \left| \int_{\mathbb{R}^N} v(y) \left[\mathbb{P}_x^{\pi^1, \pi^2}(t, dy) - \mu_{\pi^1, \pi^2}(dy) \right] \right| \\ &\leq \|v\|_w \int_{\mathbb{R}^N} w(y) \left| \mathbb{P}_x^{\pi^1, \pi^2}(t, dy) - \mu_{\pi^1, \pi^2}(dy) \right| \\ &= \|v\|_w \int_{\mathbb{R}^N} w(y) \left| \int_{\mathbb{R}^N} \mathbb{P}_z^{\pi^1, \pi^2}(mT, dy) \mathbb{P}_x^{\pi^1, \pi^2}(s, dz) - \mu_{\pi^1, \pi^2}(dy) \right|, \end{aligned} \quad (2.13)$$

by the Chapman–Kolmogorov equation. By Fubini’s Theorem, (2.13) becomes

$$\begin{aligned} \left| \mathbb{E}_x^{\pi^1, \pi^2} v(x(t)) - \mu_{\pi^1, \pi^2}(v) \right| &\leq \|v\|_w \int_{\mathbb{R}^N} \mathbb{P}_x^{\pi^1, \pi^2}(s, dz) \left\| \mathbb{P}_z^{\pi^1, \pi^2}(mT, \cdot) - \mu_{\pi^1, \pi^2}(\cdot) \right\|_w \\ &= \|v\|_w \int_{\mathbb{R}^N} \mathbb{P}_x^{\pi^1, \pi^2}(s, dz) \left\| Q_m^{\pi^1, \pi^2}(z, \cdot) - \mu_{\pi^1, \pi^2}(\cdot) \right\|_w \\ &\leq \|v\|_w \rho_2 \rho_1^m \mathbb{E}_x^{\pi^1, \pi^2} w(x(s)) \quad \text{by (2.11)} \\ &\leq \|v\|_w \rho_2 \rho_1^m \left[e^{-cs} w(x) + \frac{d}{c} (1 - e^{-cs}) \right] \quad \text{by (2.8)} \\ &\leq \|v\|_w \rho_2 \rho_1^{-1} \left(\rho_1^{1/T} \right)^t \left(1 + \frac{d}{c} \right) w(x). \end{aligned}$$

Define $C := \rho_2 \rho_1^{-1} (1 + d/c)$ and $\delta = -(\log \rho_1)/T$, so that the result follows. $\square$

The following result is true by virtue of Dynkin’s formula and (2.12).

Lemma 2.10 Assume that (2.12) holds. Let $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, $v \in \mathcal{B}_w(\mathbb{R}^N)$, and $x \in \mathbb{R}^N$. Then

$$\lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_x^{\pi^1, \pi^2} v(x(T)) = 0. \quad (2.14)$$

Suppose in addition that $v \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ is a harmonic function, in the sense that

$$\mathbb{L}^{\pi^1, \pi^2} v(x) = 0 \quad \text{for all } x \in \mathbb{R}^N. \quad (2.15)$$

Then $v(\cdot)$ is a constant; in fact,

$$v(x) = \mu_{\pi^1, \pi^2}(v) \quad \text{for all } x \in \mathbb{R}^N. \quad (2.16)$$

Proof The limit (2.14) is straightforward from (2.12). Now, if $v \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ is a harmonic function, then for every $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, $x \in \mathbb{R}^N$ and $t \geq 0$, Dynkin's formula yields

$$\mathbb{E}_x^{\pi^1, \pi^2} v(x(t)) = v(x) + \mathbb{E}_x^{\pi^1, \pi^2} \int_0^T \mathbb{L}^{\pi^1, \pi^2} v(x(s)) ds = v(x), \quad (2.17)$$

where the last equality follows from (2.15). Letting $T \rightarrow \infty$ in (2.17) and using (2.12), we complete the proof. $\square$

Following the arguments of Lemma 2.6, it is easy to verify that the combination of Lemma 2.10, Assumption 2.5 and Dynkin's formula yields

$$\mu_{\pi^1, \pi^2}(w) \leq \frac{d}{c} \quad (2.18)$$

for every (π^1, π^2) in $\Pi^1 \times \Pi^2$.

2.4 Average Optimality

Let R be a positive real number and $\bar{B}_R$ be the closure of

$$B_R := \{x \in \mathbb{R}^N : |x| < R\}. \quad (2.19)$$

Let us now introduce the *payoff* or *reward/cost* rate function r from $\mathbb{R}^N \times U^1 \times U^2$ to $\mathbb{R}$. Let us impose some conditions on r . Recall that U^1 and U^2 are compact subsets of given complete and separable vector normed spaces.

Assumption 2.11 The function r is

- (a) continuous on $\mathbb{R}^N \times U^1 \times U^2$ and locally Lipschitz in x uniformly in $(u_1, u_2) \in U^1 \times U^2$; that is, for each $R > 0$, there exists a constant $C(R)$ such that

$$\sup_{(u_1, u_2) \in U^1 \times U^2} |r(x, u_1, u_2) - r(y, u_1, u_2)| \leq C(R)|x - y|$$

for all $x, y \in \bar{B}_R$;

- (b) in $\mathcal{B}_w(\mathbb{R}^N)$ uniformly in $(u_1, u_2) \in U^1 \times U^2$, i.e., there exists a constant M such that

$$\sup_{(u_1, u_2) \in U^1 \times U^2} |r(x, u_1, u_2)| \leq Mw(x)$$

for all $x \in \mathbb{R}^N$;

- (c) concave in U^1 , and convex in U^2 .

Analogously to (2.4) and (2.5), when using randomized Markov policies (π^1, π^2) in $\Pi^1 \times \Pi^2$, we will write, for every $x \in \mathbb{R}^N$,

$$r(x, \pi^1, \pi^2) := \int_{U^2} \int_{U^1} r(x, u_1, u_2) \pi^1(du_1|x) \pi^2(du_2|x).$$

Similarly, for $(\varphi, \psi) \in V^1 \times V^2$, we define

$$r(x, \varphi, \psi) := \int_{U^2} \int_{U^1} r(x, u_1, u_2) \varphi(du_1) \psi(du_2). \quad (2.20)$$

Remark 2.12 We can verify that, for $(\varphi, \psi) \in V^1 \times V^2$, the reward rate r satisfies Assumption 2.11(a), that is, for each $R > 0$, there exists a constant $C(R)$ such that

$$\sup_{(\varphi, \psi) \in V^1 \times V^2} |r(x, \varphi, \psi) - r(y, \varphi, \psi)| \leq C(R)|x - y|$$

for all $x, y \in \bar{B}_R$.

The following result provides important facts.

Lemma 2.13 Under Assumptions 2.1 and 2.11(a), the function $r(x, \cdot, \cdot)$ is continuous on $V^1 \times V^2$. Moreover, for a fixed h in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$, $\mathbb{L}^{\varphi, \psi} h$ is continuous in $(\varphi, \psi) \in V^1 \times V^2$.

Proof Under the given Assumptions, the functions b and r are continuous in $(u_1, u^2) \in U^1 \times U^2$, and attain their respective suprema on U^1 , and infima on U^2 . Hence, the definition of weak convergence yields the result. $\square$

Remark 2.14 [40, Theorem 4.2]. The compactness of U^ℓ (resp. V^ℓ), $\ell = 1, 2$, the linearity of $h \mapsto \mathbb{L}^{\varphi, \psi} h$, Assumption 2.11(c), and Lemma 2.13 yield Isaacs' condition:

$$\sup_{\varphi \in V^1} \inf_{\psi \in V^2} \{r(x, \varphi, \psi) + \mathbb{L}^{\varphi, \psi} h(x)\} = \inf_{\psi \in V^2} \sup_{\varphi \in V^1} \{r(x, \varphi, \psi) + \mathbb{L}^{\varphi, \psi} h(x)\}.$$

For ease of notation we will combine expressions such as (2.4) and (2.5), that is, for $(\varphi, \psi) \in V^1 \times V^2$ and $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$,

$$r(x, \varphi, \pi^2) := r(x, \varphi, \pi^2(\cdot|x)) \quad \text{and} \quad r(x, \pi^1, \psi) := r(x, \pi^1(\cdot|x), \psi).$$

Similarly, for $h \in \mathcal{C}^2(\mathbb{R}^N)$,

$$\mathbb{L}^{\varphi, \pi^2} h(x) := \mathbb{L}^{\varphi, \pi^2(\cdot|x)} h(x) \quad \text{and} \quad \mathbb{L}^{\pi^1, \psi} h(x) := \mathbb{L}^{\pi^1(\cdot|x), \psi} h(x).$$

The goal of player 1 is to maximize his rewards, whereas that of player 2 is to minimize his costs in a given time horizon with respect to some performance criterion. We shall deal with ergodic payoffs, and we study first the so-called *long run average payoff* of (2.22) below.

For $(\varphi, \psi) \in V^1 \times V^2$, we define

$$r(x(t), \varphi, \psi) := \int_{U^2} \int_{U^1} r(x(t), u_1, u_2) \varphi(du_1) \psi(du_2).$$

For each $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ and $T \geq 0$, let

$$J_T(x, \pi^1, \pi^2) := \mathbb{E}_x^{\pi^1, \pi^2} \left[\int_0^T r(x(t), \pi^1, \pi^2) dt \right] \quad (2.21)$$

the *total expected payoff* of (π^1, π^2) over the time interval $[0, T]$, when the initial state is $x \in \mathbb{R}^N$. The *ergodic payoff* (also known as long-run average payoff) given the initial state x is

$$J(x, \pi^1, \pi^2) := \limsup_{T \rightarrow \infty} \frac{1}{T} J_T(x, \pi^1, \pi^2). \quad (2.22)$$

Proposition 2.15 *Let Assumptions 2.1, 2.5, 2.8 and 2.11 hold. Then the payoff rate r is μ_{π^1, π^2} -integrable for every pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$*

Proof Given $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, define

$$J(\pi^1, \pi^2) := \mu_{\pi^1, \pi^2}(r(\cdot, \pi^1, \pi^2)) = \int_{\mathbb{R}^N} r(x, \pi^1, \pi^2) \mu_{\pi^1, \pi^2}(dx). \quad (2.23)$$

with μ_{π^1, π^2} as in (2.6).

By the definition of $J(\pi^1, \pi^2)$ in (2.23), Assumption 2.11(b), (2.6) and (2.9) yield

$$|J(\pi^1, \pi^2)| \leq \int_{\mathbb{R}^N} |r(x, \pi^1, \pi^2)| \mu_{\pi^1, \pi^2}(dx) \leq M \cdot \mu_{\pi^1, \pi^2}(w) < \infty \quad (2.24)$$

for all $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$. In fact, by (2.18),

$$|J(\pi^1, \pi^2)| \leq M \cdot \mu_{\pi^1, \pi^2}(w) \leq M \cdot \frac{d}{c}, \quad (2.25)$$

so that $J(\pi^1, \pi^2)$ is uniformly bounded on $\Pi^1 \times \Pi^2$. This yields the desired result. $\square$

It follows from (2.12) that the average payoff (2.22) coincides with the constant $J(\pi^1, \pi^2)$ in (2.23) for every $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$. Indeed, note that J_T defined in (2.21) can be expressed as

$$J_T(x, \pi^1, \pi^2) = T \cdot J(\pi^1, \pi^2) + \int_0^T \left[\mathbb{E}_x^{\pi^1, \pi^2} r(x(t), \pi^1, \pi^2) - J(\pi^1, \pi^2) \right] dt.$$

Hence, multiplying the latter equality by $\frac{1}{T}$ and letting $T \rightarrow \infty$, by (2.12), we obtain,

$$J(x, \pi^1, \pi^2) = \limsup_{T \rightarrow \infty} \frac{1}{T} J_T(x, \pi^1, \pi^2) = J(\pi^1, \pi^2) \quad \text{for all } x \in \mathbb{R}^N. \quad (2.26)$$

By virtue of this last expression, we can write (2.22) simply as $J(\pi^1, \pi^2)$.

We now define the constant values

$$\mathcal{U} := \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(\pi^1, \pi^2) \quad (2.27)$$

and

$$\mathcal{L} := \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(\pi^1, \pi^2). \quad (2.28)$$

The function $\mathcal{L}$ is called the game's *lower value*, and $\mathcal{U}$ is the game's *upper value*. Clearly, we have $\mathcal{L} \leq \mathcal{U}$. If these two numbers coincide, then the game is said to have a *value*, say $\mathcal{V}$. This number is the common value of $\mathcal{L}$ and $\mathcal{U}$, i.e.,

$$\mathcal{V} := \mathcal{L} = \mathcal{U}. \quad (2.29)$$

As a consequence of (2.26) and (2.25), $\mathcal{L}$ and $\mathcal{U}$ are finite. This implies that $\mathcal{V}$ is also finite if the second equality in (2.29) holds.

The basic problem we are concerned with is to find *average payoff equilibria* or *saddle points* of the average payoff SDG. Namely, we are interested in pairs $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ for which

$$J(\pi^1, \pi_*^2) \leq J(\pi_*^1, \pi_*^2) \leq J(\pi_*^1, \pi^2) \quad (2.30)$$

for every $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$. The set of pairs of average payoff equilibria is denoted by $(\Pi^1 \times \Pi^2)_{ae}$.

Remark 2.16 Observe that if (π_*^1, π_*^2) is an average payoff equilibrium, then the game has a value $J(\pi_*^1, \pi_*^2) =: \mathcal{V}$. As in the discounted payoff case, the converse is not necessarily true.

Definition 2.17 We say that a constant $J \in \mathbb{R}$, a function $h \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$, and a pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ verify the average payoff optimality equations if, for every $x \in \mathbb{R}^N$,

$$J = r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h(x) \quad (2.31)$$

$$= \sup_{\varphi \in V^1} \left\{ r(x, \varphi, \pi^2) + \mathbb{L}^{\varphi, \pi^2} h(x) \right\} \quad (2.32)$$

$$= \inf_{\psi \in V^2} \left\{ r(x, \pi^1, \psi) + \mathbb{L}^{\pi^1, \psi} h(x) \right\}. \quad (2.33)$$

In this case, the pair of strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ satisfying (2.31)–(2.33) is called a *canonical equilibrium*.

The following result holds by virtue of Dynkin's formula. We omit the proof because it is immediate.

Proposition 2.18 *If there is a constant J , a function h in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ and a pair (π^1, π^2) in $\Pi^1 \times \Pi^2$ such that*

$$J \geq r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h(x) \quad \text{for all } x \in \mathbb{R}^N, \quad (2.34)$$

then

$$J \geq J(\pi^1, \pi^2). \quad (2.35)$$

Similarly, if the inequality (2.34) is replaced by “ $\leq$ ”, then (2.35) should be replaced by the same inequality, i.e., if

$$J \leq r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h(x), \quad \text{then } J \leq J(\pi^1, \pi^2).$$

Therefore, if the equality holds in (2.34), then we have $J = J(\pi^1, \pi^2)$.

3 The Policy Iteration Algorithm

We now introduce the PIA, also known as policy improvement algorithm. The version we present in this section was inspired by Hernández-Lerma and Lasserre [20, Remark 2.4], Zhu et al. [45, pp. 7–8] and Van der Wal [41, Algorithm (HK)].

3.1 The PIA

- Step 1 Set $m = 0$. Select a strategy $\pi_0^2 \in \Pi^2$, and define $J(\pi_{-1}^1, \pi_{-1}^2) := -\infty$
- Step 2 Find a policy $\pi_m^1 \in \Pi^1$, a constant $J(\pi_m^1, \pi_m^2)$, and a function $h_m : \mathbb{R}^N \rightarrow \mathbb{R}$ of class $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ such that $(J(\pi_m^1, \pi_m^2), h_m)$ is a solution of (3.1)–(3.2):

$$J(\pi_m^1, \pi_m^2) = \sup_{\varphi \in V^1} \left[r(x, \varphi, \pi_m^2) + \mathbb{L}^{\varphi, \pi_m^2} h_m(x) \right] \quad (3.1)$$

$$= r(x, \pi_m^1, \pi_m^2) + \mathbb{L}^{\pi_m^1, \pi_m^2} h_m(x) \quad \text{for all } x \in \mathbb{R}^N. \quad (3.2)$$

Observe that

$$J(\pi_m^1, \pi_m^2) \geq \inf_{\psi \in V^2} \left[r(x, \pi_m^1, \psi) + \mathbb{L}^{\pi_m^1, \psi} h_m(x) \right] \quad \text{for all } x \in \mathbb{R}^N. \quad (3.3)$$

- Step 3 If $J(\pi_m^1, \pi_m^2) = J(\pi_{m-1}^1, \pi_{m-1}^2)$, then $J(\pi_m^1, \pi_m^2)$ is the value of the game and (π_m^1, π_m^2) is a saddle point. Terminate PIA. Otherwise, go to step 4.
- Step 4 Determine a strategy $\pi_{m+1}^2 \in \Pi^2$ that attains the minimum on the right hand side of (3.3), i.e., for all $x \in \mathbb{R}^N$

$$r(x, \pi_m^1, \pi_{m+1}^2) + \mathbb{L}^{\pi_m^1, \pi_{m+1}^2} h_m(x) = \inf_{\psi \in V^2} \left[r(x, \pi_m^1, \psi) + \mathbb{L}^{\pi_m^1, \psi} h_m(x) \right]. \quad (3.4)$$

Increase m in 1 and go back to step 2.

Remark 3.1 Observe that Remark 2.14 makes us indifferent between using the PIA version we have proposed, and using a modification that minimizes in (3.2) in step 2, and maximizes in (3.4) in step 4.

Definition 3.2 The PIA is said to converge if the sequence $J(\pi_m^1, \pi_m^2)$ converges to the value of the game defined in (2.30). That is,

$$J(\pi_m^1, \pi_m^2) \rightarrow \mathcal{V}.$$

To ensure the convergence of the PIA, we need to guarantee it is well-defined. To do this, it is necessary to satisfy the following conditions.

- (1) For every pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, there exists an invariant probability measure μ_{π^1, π^2} . This is the first consequence of Assumption 2.5.
- (2) For every pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, the payoff rate $r(\cdot, \pi^1, \pi^2)$ is μ_{π^1, π^2} -integrable, so that (2.23) holds, that is,

$$J(\pi^1, \pi^2) := \mu_{\pi^1, \pi^2}(r(\cdot, \pi^1, \pi^2)) = \int_{\mathbb{R}^N} r(x, \pi^1, \pi^2) \mu_{\pi^1, \pi^2}(dx).$$

This follows from Proposition 2.15.

- (3) For every pair (π^1, π^2) there is a unique solution $(J(\pi^1, \pi^2), h_{\pi^1, \pi^2})$ to the Poisson equation

$$J(\pi^1, \pi^2) = r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h_{\pi^1, \pi^2}(x) \quad \text{for all } x \in \mathbb{R}^N, \quad (3.5)$$

which is guaranteed by Proposition 3.4 below.

- (4) For each $\pi_m^2 \in \Pi^2$, there exists a strategy $\pi_m^1 \in \Pi^1$ such that (3.2) holds. This is indeed the case by virtue of Assumption 2.11, the compactness of V^1 , and Theorem 2.1 in [31].
- (5) For every function h_m in a suitable set, there exists a strategy $\pi_{m+1}^2 \in \Pi^2$ such that (3.4) holds. This statement is true by Assumption 2.11, the compactness of V^2 , and again Theorem 2.1 in [31].

As already noted above, a necessary condition for the algorithm to be well-defined is the existence of a solution $(J(\pi^1, \pi^2), h_{\pi^1, \pi^2})$ to the Poisson equation (3.5). To prove this, we introduce the concept of *bias* of (π^1, π^2) (see, for instance, [44, Sect. 2, 44, Sect. 4.1]).

Definition 3.3 Let $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$. The bias of (π^1, π^2) is the function given by

$$h_{\pi^1, \pi^2}(x) := \int_0^\infty \left[\mathbb{E}_x^{\pi^1, \pi^2} r(x(t), \pi^1, \pi^2) - J(\pi^1, \pi^2) \right] dt. \quad (3.6)$$

Observe that this function is finite-valued because (2.12) and the Assumption 2.11(b) give, for all $t \geq 0$,

$$\left| \mathbb{E}_x^{\pi^1, \pi^2} r(x(t), \pi^1, \pi^2) - J(\pi^1, \pi^2) \right| \leq C e^{-\delta t} M_W(x). \quad (3.7)$$

Hence, by (3.6) and (3.7), the bias of (π^1, π^2) is such that

$$|h_{\pi^1, \pi^2}(x)| \leq \delta^{-1} CM w(x), \quad (3.8)$$

and so

$$\|h_{\pi^1, \pi^2}\|_w \leq \delta^{-1} CM.$$

This means that the bias h_{π^1, π^2} is a finite-valued function and, in fact, is in $\mathcal{B}_w(\mathbb{R}^N)$. Actually, its w -norm is uniformly bounded on $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$.

The following result is necessary to ensure that the PIA is well-defined.

Proposition 3.4 *For each $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, the pair $(J(\pi^1, \pi^2), h_{\pi^1, \pi^2})$ is the unique solution of the Poisson equation (3.5) for which the μ_{π^1, π^2} -expectation of h_{π^1, π^2} is zero:*

$$\mu_{\pi^1, \pi^2}(h_{\pi^1, \pi^2}) = \int_{\mathbb{R}^N} h_{\pi^1, \pi^2}(x) \mu_{\pi^1, \pi^2}(dx) = 0. \quad (3.9)$$

Moreover, h_{π^1, π^2} is in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$.

Proof A slight variation of the vanishing discount technique (see, for instance the Appendix of [12], or [27, Chaps. 5, 6]) gives us that, for fixed $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, the Poisson equation (3.5) has a solution $\tilde{h}_{\pi^1, \pi^2}$, which is a member of $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$, i.e.,

$$\tilde{J}(\pi^1, \pi^2) = r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} \tilde{h}_{\pi^1, \pi^2}(x) \quad \text{for all } x \in \mathbb{R}^N. \quad (3.10)$$

To obtain (3.9) first note that, by (2.6) and (3.8), h_{π^1, π^2} is indeed μ_{π^1, π^2} -integrable for every (π^1, π^2) in $\Pi^1 \times \Pi^2$. Then, in (3.9) choose the distribution of the initial state to be μ_{π^1, π^2} and so (3.9) follows from Fubini's theorem and the invariance of μ_{π^1, π^2} . Moreover, the fact that h_{π^1, π^2} is in $\mathcal{B}_w(\mathbb{R}^N)$ follows from (3.8).

On the other hand, the fact that $\tilde{J}(\pi^1, \pi^2)$ coincides with the ergodic payoff $J(\pi^1, \pi^2)$ in (2.23) is a direct consequence of the proof of Proposition 2.18 and the part that addresses uniqueness in Theorem 3.5(i).

Next, to ensure that $\tilde{h}_{\pi^1, \pi^2}$ equals the bias h_{π^1, π^2} in (3.6) for all $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, we can use Dynkin's formula on $\tilde{h}_{\pi^1, \pi^2}(x(t))$ to obtain

$$\mathbb{E}_x^{\pi^1, \pi^2}[\tilde{h}_{\pi^1, \pi^2}(x(t))] = \tilde{h}_{\pi^1, \pi^2}(x) + \tilde{J}(\pi^1, \pi^2)t - \mathbb{E}_x^{\pi^1, \pi^2}\left[\int_0^t r(x(s), \pi^1, \pi^2)ds\right].$$

This implies

$$\tilde{h}_{\pi^1, \pi^2}(x) = \mathbb{E}_x^{\pi^1, \pi^2}\left[\int_0^t (r(x(s), \pi^1, \pi^2) - J(\pi^1, \pi^2))ds\right] + \mathbb{E}_x^{\pi^1, \pi^2}[\tilde{h}_{\pi^1, \pi^2}(x(t))]. \quad (3.11)$$

Since h_{π^1, π^2} is in $\mathcal{B}_w(\mathbb{R}^N)$ for all $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, we see that the uniform w -exponential ergodicity condition (2.11) yields that the second term of the right hand

side of (3.11) converges to $\mu_{\pi^1, \pi^2}(\tilde{h}_{\pi^1, \pi^2})$ as t goes to infinity; but, by (3.9), this last limit becomes zero. Therefore, letting $t \rightarrow \infty$ in both sides of (3.11), we obtain

$$\tilde{h}_{\pi^1, \pi^2}(x) = \mathbb{E}_x^{\pi^1, \pi^2} \left[\int_0^\infty (r(x(s), \pi^1, \pi^2) - J(\pi^1, \pi^2)) ds \right],$$

which coincides with the bias h_{π^1, π^2} defined in (3.6). These facts also yield uniqueness of solutions to equation (3.5), and Proposition 3.4 follows. $\square$

From (2.31)–(2.33) and Proposition 2.18, it is easy to see that the constant J in (2.31) is the value of the average payoff game. Moreover, the same arguments lead to the conclusion that a pair of canonical strategies is always average optimal. The following result ensures that the converse is also true. We offer a complete proof based on the vanishing discount technique for control problems (cf. [4, 30]) in [12, Theorem 4.1]. See also [9, Corollary 6.2] and [39, Theorem 12.1].

Theorem 3.5 *If Assumptions 2.1, 2.5, 2.8, and 2.11 hold, then:*

- (i) *There exist solutions to the average optimality equations (2.31)–(2.33). Moreover, the constant J equals $\mathcal{V}$, the value of the game, and the function h is unique under the extra assumption that $h(0) = 0$.*
- (ii) *A pair of strategies is average optimal if, and only if, it is canonical.*

This result guarantees the existence of the pairs (J, h) in $\mathbb{R} \times (\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N))$ and (π^1, π^2) in $\Pi^1 \times \Pi^2$ to the average optimality equations (2.31)–(2.33). However, the question now is how can we actually find (or at least approximate) values of $\mathcal{V}$, h , and (π^1, π^2) . In the next section we show that this can be done using the PIA.

4 Convergence

This section is intended to prove that the PIA converges. But first, we present the following extension of [20, Lemma 4.5]. Part (b) of Lemma 4.1, along with (3.3) gives that if $J(\pi_m^1, \pi_m^2) = J(\pi_{m+1}^1, \pi_{m+1}^2)$ in the PIA, then (π_m^1, π_m^2) is a saddle point of the average SDG.

Lemma 4.1 *Let $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ be an arbitrary pair of randomized stationary strategies. Let $\pi_*^1 \in \Pi^1$ be such that*

$$J(\pi_*^1, \pi^2) = \sup_{\varphi \in V^1} \left[r(x, \varphi, \pi^2) + \mathbb{L}^{\varphi, \pi^2} h(x) \right] \quad (4.1)$$

$$= r(x, \pi_*^1, \pi^2) + \mathbb{L}^{\pi_*^1, \pi^2} h_{\pi_*^1, \pi^2}(x) \quad \text{for all } x \in \mathbb{R}^N, \quad (4.2)$$

and let $\pi_^2 \in \Pi^2$ be such that*

$$\inf_{\psi \in V^2} \left[r(x, \pi_*^1, \psi) + \mathbb{L}^{\pi_*^1, \psi} h_{\pi_*^1, \pi_*^2}(x) \right] = r(x, \pi_*^1, \pi_*^2) + \mathbb{L}^{\pi_*^1, \pi_*^2} h_{\pi_*^1, \pi_*^2}(x) \quad (4.3)$$

for all $x \in \mathbb{R}^N$. Then

- (a) $J(\pi_*^1, \pi_*^2) \leq J(\pi_*^1, \pi^2)$, and
- (b) if $J(\pi_*^1, \pi_*^2) \leq J(\pi_*^1, \pi^2)$, then (π_*^1, π_*^2) is a saddle point of the SDG with average payoff.

Proof The relations (4.1)–(4.3) imply

$$r(x, \pi_*^1, \pi^2) + \mathbb{L}^{\pi_*^1, \pi^2} h_{\pi_*^1, \pi^2}(x) \leq r(x, \pi_*^1, \pi_*^2) + \mathbb{L}^{\pi_*^1, \pi_*^2} h_{\pi_*^1, \pi_*^2}(x) = J(\pi_*^1, \pi^2).$$

An application of Proposition 2.18 yields (a). Part (b) of the result is immediate from (a) and (2.30). $\square$

Proposition 4.2 guarantees the existence of a pair of policies (π_*^1, π_*^2) in $\Pi^1 \times \Pi^2$ that satisfies that, for every fixed $x \in \mathbb{R}^N$, there exists a subsequence $m_k \equiv m_k(x)$ of $\{m\}$ such that

$$(\pi_{m_k}^1(\cdot|x), \pi_{m_k}^2(\cdot|x)) \rightarrow (\pi_*^1(\cdot|x), \pi_*^2(\cdot|x)) \quad \text{as } k \rightarrow \infty \quad (4.4)$$

in the topology of weak convergence of $V^1 \times V^2$. This type of policy convergence was first introduced in [38, Lemma 4] for the case of nonstationary, deterministic, discrete-time policies. It can also be found in Schäl [39, Proposition 12.2] and Hernández–Lerma and Lasserre [19, Theorem D.7]. In this case we say that the sequence $\{(\pi_m^1, \pi_m^2) : m = 1, 2, \dots\}$ converges in the sense of Schäl to (π_*^1, π_*^2) .

Proposition 4.2 Let $\{(\pi_m^1, \pi_m^2) : m = 1, 2, \dots\} \subset \Pi^1 \times \Pi^2$ be the sequence generated by the PIA. If Assumptions 2.1, 2.5 and 2.11 hold, then, there exists $(\pi_*^1, \pi_*^2) \in \Pi^1 \times \Pi^2$ such that (π_*^1, π_*^2) is the limit in the sense of Schäl of $\{(\pi_m^1, \pi_m^2) : m = 1, 2, \dots\}$.

Proof Fix $x \in \mathbb{R}^N$. By the compactness of $V^1 \times V^2$, we can ensure the existence of a subsequence $m_k \equiv m_k^\ell(x)$, $\ell = 1, 2$, such that $\pi_{m_k}^\ell(\cdot|x) \rightarrow \pi_*^\ell(\cdot|x)$. Using again the compactness of V^ℓ , $\ell = 1, 2$, we easily see that $\pi_*^\ell(\cdot|x)$ is a probability measure. Furthermore, for all $B \subseteq U^\ell$, by Schäl [38, Lemma 4], $\pi_*^\ell(B|\cdot)$ is measurable on $\mathbb{R}^N$. Hence, π_*^ℓ is in Π^ℓ . This proves the result. $\square$

Theorem 4.3 Let $p > n$. Let Assumptions 2.1, 2.5, 2.8 and 2.11 hold. In addition, let (π_m^1, π_m^2) be a pair of randomized stationary policies generated by the PIA. Then $\{(\pi_m^1, \pi_m^2) : m = 1, 2, \dots\}$ converges in the sense of Schäl to a saddle point (π_*^1, π_*^2) of the average SDG. Therefore the PIA converges.

Proof For each pair (π_m^1, π_m^2) generated by the PIA, Proposition 3.4 ensures the existence of a function $h_m \in C^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ such that (3.2) holds. Now, an

invocation to Corollary 3.5 in [27] (see also [2, Lemma 3.5], [3, Lemma 3.4.18], [22, Proposition A.3]) establishes the existence of a function $h \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ such that

$$\lim_{m \rightarrow \infty} h_m = h \quad \text{for all } x \in \mathbb{R}^N.$$

On the other hand, observe that Proposition 4.2 asserts the existence of the limit (π_*^1, π_*^2) (in the sense of Schäl) of the sequence of policies $\{(\pi_m^1, \pi_m^2)\}$ generated by the PIA.

Now, fix an arbitrary state $x \in \mathbb{R}^N$ and let m_k be as in (4.4). Next, in (3.2), replace m by m_k and let $k \rightarrow \infty$ to obtain

$$J(\pi_*^1, \pi_*^2) = r(x, \pi_*^1, \pi_*^2) + \mathbb{L}^{\pi_*^1, \pi_*^2} h(x).$$

We shall use Lemma 4.1 to conclude the proof. Namely, observe that (3.1) in step 2 of the PIA ensures that (4.1) holds. In addition, (3.4) in step 4 yields (4.3). Hence, Lemma 4.1b asserts that (π_*^1, π_*^2) is a saddle point of the ergodic game and the result is thus proved. $\square$

5 Bias Optimality

Throughout the following we will suppose that Assumptions 2.1, 2.5, 2.8 and 2.11 hold.

We recall that the set of strategies that satisfy (2.30) is denoted by $(\Pi^1 \times \Pi^2)_{ae}$, that is, (π_*^1, π_*^2) is in $(\Pi^1 \times \Pi^2)_{ae}$ if and only if

$$J(\pi^1, \pi_*^2) \leq J(\pi_*^1, \pi_*^2) \leq J(\pi_*^1, \pi^2) \quad \text{for every } (\pi^1, \pi^2) \in \Pi^1 \times \Pi^2.$$

Recall as well Definition 3.3 of the bias h_{π^1, π^2} and its characterization as solution of the Poisson equation given in Proposition 3.4.

The following definition uses the concept of average payoff equilibria introduced above.

Definition 5.1 We say that an average payoff equilibrium $(\pi_*^1, \pi_*^2) \in (\Pi^1 \times \Pi^2)_{ae}$ is a *bias equilibrium* if

$$h_{\pi^1, \pi_*^2}(x) \leq h_{\pi_*^1, \pi_*^2}(x) \leq h_{\pi_*^1, \pi^2}(x) \quad (5.1)$$

for all $x \in \mathbb{R}^N$ and every pair of average payoff equilibria $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{ae}$. The function $h_{\pi_*^1, \pi_*^2}$ is called the *optimal bias function*.

The next result is an extension to SDGs of [23, Proposition 4.2]. It gives an expression for the bias function of (π^1, π^2) by using any solution h of the average optimality equations (2.31)–(2.33).

Proposition 5.2 *If $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{ae}$, then its bias h_{π^1, π^2} and any solution h of the average optimality equations (2.31)–(2.33) coincide up to an additive constant; in fact,*

$$h_{\pi^1, \pi^2}(x) = h(x) - \mu_{\pi^1, \pi^2}(h) \quad \text{for all } x \in \mathbb{R}^N. \quad (5.2)$$

Proof Let $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{ae}$ be an arbitrary average payoff equilibrium. Then (π^1, π^2) satisfies Theorem 3.5(ii) with $J = \mathcal{V}$, i.e.,

$$\mathcal{V} = r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h(x) \quad \text{for all } x \in \mathbb{R}^N. \quad (5.3)$$

In addition, the Poisson equation for (π^1, π^2) is

$$\mathcal{V} = r(x, \pi^1, \pi^2) + \mathbb{L}^{\pi^1, \pi^2} h_{\pi^1, \pi^2}(x) \quad \text{for all } x \in \mathbb{R}^N. \quad (5.4)$$

The subtraction of (5.3) from (5.4) yields that $h - h_{\pi^1, \pi^2}$ is a harmonic function. Consequently, (5.2) follows from Dynkin's formula, Lemma 2.10 and (3.9). $\square$

If the optimal bias function $h_{\pi^1_*, \pi^2_*}$ exists, then, by Proposition 3.4, it is in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ for any bias equilibrium (π^1_*, π^2_*) .

Let $(\Pi^1 \times \Pi^2)_{bias}$ be the family of bias equilibria. By Definition 5.1,

$$(\Pi^1 \times \Pi^2)_{bias} \subset (\Pi^1 \times \Pi^2)_{ae}.$$

Let $(J, h) \in \mathbb{R} \times (\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N))$ be a solution of the average optimality equations (2.31)–(2.33). We define for every $x \in \mathbb{R}^N$ the sets

$$\Pi^1(x) := \left\{ \varphi \in V^1 : J = \inf_{\psi \in V^2} [r(x, \varphi, \psi) + \mathbb{L}^{\varphi, \psi} h(x)] \right\}, \quad (5.5)$$

$$\Pi^2(x) := \left\{ \psi \in V^2 : J = \sup_{\varphi \in V^1} [r(x, \varphi, \psi) + \mathbb{L}^{\varphi, \psi} h(x)] \right\}. \quad (5.6)$$

We now present an extension of Prieto–Rumeau and Hernández–Lerma [34, Lemma 4.6].

Lemma 5.3 *For every $x \in \mathbb{R}^N$, $\Pi^1(x)$ and $\Pi^2(x)$ are convex compact sets.*

Proof Recall from Sect. 2.2 that the sets V^1 and V^2 (endowed with the topology of weak convergence) are compact. Thus, we only need to show that $\Pi^\ell(x)$ is a closed set ($\ell = 1, 2$). But this is a consequence of Lemma 4.4 in [34] and Lemma 2.13. The proof that $\Pi^1(x)$ and $\Pi^2(x)$ are convex mimics that of Lemma 4.6 in [34]. $\square$

Remark 5.4 *By Theorem 3.5(ii), (π^1, π^2) is in $(\Pi^1 \times \Pi^2)_{ae}$ if and only if $\pi^1(\cdot|x)$ is in $\Pi^1(x)$ and $\pi^2(\cdot|x)$ is in $\Pi^2(x)$ for all $x \in \mathbb{R}^N$.*

Theorem 5.5 *The set $(\Pi^1 \times \Pi^2)_{bias}$ is nonempty.*

Proof Let $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{ae}$ be an average payoff equilibrium. Using the expression (5.2) for the bias function h_{π^2, π^2} , we obtain that finding bias equilibria is equivalent to solving a new SDG with ergodic payoff. Let us call this problem *bias game*. The components of this game are:

- The dynamic system (2.1)
- The action sets $\Pi^1(x)$ and $\Pi^2(x)$ for each $x \in \mathbb{R}^N$
- The reward rate

$$r'(x, \pi^1, \pi^2) := -h(x).$$

Observe that the bias game satisfies Assumptions 2.1, 2.5, 2.8 and 2.11. Hence, Theorem 3.5 ensures the existence of a constant $\tilde{J}$; a function $\tilde{h} \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$; and a pair (π^1, π^2) such that $(\pi^1(\cdot|x), \pi^2(\cdot|x))$ is in $\Pi^1(x) \times \Pi^2(x)$ for every $x \in \mathbb{R}^N$. Moreover, $\tilde{h}$ and (π^1, π^2) satisfy the average payoff optimality equations (2.31)–(2.33); that is,

$$\tilde{J} = -h(x) + \mathbb{L}^{\pi^1, \pi^2} \tilde{h}(x) \quad (5.7)$$

$$= \sup_{\varphi \in \Pi^1(x)} \left[-h(x) + \mathbb{L}^{\varphi, \pi^2} \tilde{h}(x) \right] \quad (5.8)$$

$$= \inf_{\psi \in \Pi^2(x)} \left[-h(x) + \mathbb{L}^{\pi^1, \psi} \tilde{h}(x) \right]. \quad (5.9)$$

Hence, by virtue of Theorem 3.5(ii) and (5.2), (π^1, π^2) is a bias equilibrium. i.e., $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{bias}$. By (5.7)–(5.9) the value of the bias game is

$$\tilde{J} = \mu_{\pi^1, \pi^2}(-h) := \mathcal{V}_h^*.$$

□

As an abuse of terminology, we take the supremum and the infimum of equations (5.8) and (5.9), respectively, over the sets $\Pi^1(x)$ and $\Pi^2(x)$, when, actually we should take them over the sets of probability measures defined on $\Pi^1(x)$ and $\Pi^2(x)$. However, note that this is not necessary because, by Lemma 5.3 these sets are readily convex and compact.

The set-valued mapping $x \mapsto \Pi^1(x)$ (resp. $x \mapsto \Pi^2(x)$) is upper (resp. lower) semicontinuous. See the proof of [23, Lemma 5.2] for details. Let $(\pi^1, \pi^2) \in (\Pi^1 \times \Pi^2)_{bias}$. Using (5.2), we define

$$\mathcal{H}(x) := h_{\pi^1, \pi^2}(x) = h(x) + \mathcal{V}_h^*, \quad (5.10)$$

where $\mathcal{V}_h^*$ is the value of the bias game.

5.1 Bias Optimality Equations

We give a characterization of bias equilibria by means of the bias optimality equations defined as follows.

Definition 5.6 We say that the constant $J \in \mathbb{R}$, the functions $\mathcal{H}, \tilde{h} \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ and the pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ verify the bias optimality equations if J and $\mathcal{H}$ satisfy the average optimality equations (2.31)–(2.33) and, in addition for every $x \in \mathbb{R}^N$, $\tilde{h}$ satisfies

$$\mathcal{H}(x) = \mathbb{L}^{\pi^1, \pi^2} \tilde{h}(x) \quad (5.11)$$

$$= \sup_{\varphi \in \Pi^1(x)} \mathbb{L}^{\varphi, \pi^2} \tilde{h}(x) \quad (5.12)$$

$$= \inf_{\psi \in \Pi^2(x)} \mathbb{L}^{\pi^1, \psi} \tilde{h}(x). \quad (5.13)$$

Theorem 5.7 Under our hypotheses, the following assertions are true.

- (i) A solution of the bias optimality equations (2.31)–(2.33) and (5.11)–(5.13), with $\mathcal{H}(0) = \mathcal{V}_h^*$, exists, is unique, and, further, $J = \mathcal{V}$, with $\mathcal{V}$ as in (2.29).
- (ii) A pair of stationary strategies $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ is a bias equilibrium if and only if it verifies the bias optimality equations.

Proof By Theorem 3.5 we know that the equations (2.31)–(2.33) have a unique solution (V, h) . Now, since $\mathbb{L}^{\pi^1, \pi^2}$ is a differential operator, it follows that, if h satisfies (2.31)–(2.33), then, so does $\mathcal{H}$ in (5.10). On the other hand, the same arguments in the proof of Theorem 5.5 ensure the existence of a function, say $\tilde{h} \in \mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ such that $(\mathcal{V}_h^*, \tilde{h})$ is the unique solution to the average optimality equations for the bias game with reward rate $-h(\cdot)$, i.e., $(\mathcal{V}_h^*, \tilde{h})$ satisfies (5.7)–(5.9). Hence, from (5.10) we can see that $\mathcal{H}(x)$ satisfies (5.11)–(5.13).

Part (ii) follows from Theorem 3.5(ii) applied to the bias game. $\square$

5.2 The PIA for the Bias Game

By the proof of Theorem 5.5, the bias game can be expressed as an SDG with a particular average payoff. We will use this and a modification of the PIA presented in Sect. 3 to find another characterization of bias equilibria.

We assume that the original SDG with average payoff of Sect. 2.4 has been solved, i.e., J is the game value, (π_0^1, π_0^2) belongs to $(\Pi^1 \times \Pi^2)_{ae}$, and $h(x) = h_{\pi_0^1, \pi_0^2}(x) + \mu_{\pi_0^1, \pi_0^2}(h)$ for all $x \in \mathbb{R}^N$.

Step 1 Set $m = 0$. Fix $\pi_0^2 \in \Pi^2(x)$ and define $\tilde{J}_0 := -\infty$.

- Step 2 Find a policy $\pi_m^1(\cdot|x) \in \Pi^1(x)$, a constant $\tilde{J}_m$, and a function $\tilde{h}_m : \mathbb{R}^N \rightarrow \mathbb{R}$ such that $(\tilde{J}_m, \tilde{h}_m)$ is a solution of (5.8).
- Step 3 If $\tilde{J}_m = \tilde{J}_{m-1}$, then $(\pi_m^1, \pi_m^2) \in (\Pi^1 \times \Pi^2)_{bias}$. Terminate PIA. Otherwise, go to step 4.
- Step 4 Determine an average optimal strategy $\pi_{m+1}^2(\cdot|x) \in \Pi^2(x)$ that attains the minimum on (5.9).

Increase m in 1 and go to step 2.

Analogously to the end of Sect. 3, there are some critical parts we must verify to ensure that this version of the PIA for the bias game is well-defined and yields a pair of bias equilibria.

- (1) In step 2, $\pi_m^1(\cdot|x)$ is such that

$$\begin{aligned}\tilde{J}_m &= \sup_{\varphi \in \Pi^1(x)} \left[-h(x) + \mathbb{L}^{\varphi, \pi_m^2} \tilde{h}_m(x) \right] \\ &= -h(x) + \mathbb{L}^{\pi_m^1, \pi_m^2} \tilde{h}_m(x),\end{aligned}$$

which is consistent with Theorem 5.7. Similarly, the strategy $\pi_{m+1}^2(\cdot|x)$ of step 3 is such that

$$\begin{aligned}\tilde{J}_m &= \inf_{\psi \in \Pi^2(x)} \left[-h(x) + \mathbb{L}^{\pi_m^1, \psi} \tilde{h}_m(x) \right] \\ &= -h(x) + \mathbb{L}^{\pi_m^1, \pi_{m+1}^2} \tilde{h}_m(x).\end{aligned}$$

- (2) Proposition 2.15 gives that $-h$ is μ_{π^1, π^2} -integrable.
- (3) Lemma 5.3 can be invoked to ensure the compactness of $\Pi^1(x)$. Thus Theorem 2.1 in [38] (with $V^1 \times V^2$ replaced by $\Pi^1(x) \times \Pi^2(x)$) allows us to extend [23, Theorem 3.2] to the context of randomized strategies. These steps enable us to guarantee the existence of $\tilde{J}_m$, a function $\tilde{h}_m$ in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$, and φ in $\Pi^1(x)$, that maximizes (5.8).
- (4) Assumption 2.11, Lemma 5.3 and Theorem 2.1 in [38] (with $\Pi^2(x)$ in lieu of V^2) allow the extension of [23, Theorem 3.2] that ensures that, for h_m given, there exists $\pi^2(\cdot|x) \in \Pi^2(x)$ that minimizes (5.9).

These remarks, together with Lemma 4.1 and Theorem 4.3 give that the PIA for the bias game is well-defined.

6 An Example

Now we will give an example to illustrate our results. This is an extension of the example presented in [12] which in turn was motivated by the manufacturing system in [1].

Consider the scalar diffusion

$$dx(t) = [ax(t) + bu_1(t) + cu_2(t)]dt + \sigma dW(t), \quad x(0) = x \quad (6.1)$$

where $a, b, c, \sigma > 0$, $b > c$, and the players' choices are given by $u_1, u_2 \in [0, d]$. Let the reward rate be given by $r(x, u_1, u_2) = -(u_1x + u_2x)$. It is very easy to see that (6.1) meets Assumptions 2.1 and [28, Assumption 4.1] hold (by Lemma 4.8 in [28], this last assertion yields that Assumption 2.8 holds as well), while the reward rate satisfies Assumption 2.11. A Lyapunov function such that Assumption 2.5 holds is $w(x) := x^2 + 1$.

Recall (2.30). If we are intended to find Nash equilibria for a zero sum game under the long-run average payoff criterion (2.26), the average payoff optimality equation (2.32) is given by

$$J = -x\pi^2 + axh'(x) + ch'(x)\pi^2 + \frac{1}{2}h''(x)\sigma^2 + \sup_{\phi \in V^1} \{(-x + bh'(x))\phi\}.$$

By Theorem 4.8 in [11], it is possible to simplify this expression by writing

$$J = axh'(x) - (x - ch'(x))u_2 + \frac{1}{2}h''(x)\sigma^2 + \sup_{u \in U_1} \{(-x + bh'(x))u\}, \quad (6.2)$$

that is, to consider only deterministic strategies. Likewise, equation (2.33) reduces to

$$J = axh'(x) - (x - bh'(x))u_1 + \frac{1}{2}h''(x)\sigma^2 + \inf_{u \in U_2} \{(-x + ch'(x))u\}. \quad (6.3)$$

Then, we apply the PIA for the expected long-run average criterion presented in Sect. 3.

Step 1 Set $m = 0$, $J_m = -\infty$, and fix $u_2^m = 0$.

Step 2 Note that (6.2) becomes

$$J_1 = axh' + \frac{1}{2}h''\sigma^2 + \sup_{u \in [0, d]} \{(-x + bh')u\}. \quad (6.4)$$

Then,

$$u_1^1 = \begin{cases} 0 & \text{when } h' < \frac{x}{b}, \\ u \in [0, d] & \text{when } h' = \frac{x}{b}, \\ d & \text{when } h' > \frac{x}{b}. \end{cases} \quad (6.5)$$

The function h and the constant J_1 can be found by solving the ordinary differential equation of second order (6.4).

Step 3 Clearly $J_1 \neq J_0 = -\infty$. Then we go to Step 4.

Step 4 We write (3.4) by means of (6.3):

$$(-x + ch')u_2^1 = \inf_{u \in [0, d]} \{(-x + ch')u\},$$

and determine

$$u_2^1 = \begin{cases} 0 & \text{when } h' > \frac{x}{c}, \\ u \in [0, d] & \text{when } h' = \frac{x}{c}, \\ d & \text{when } h' < \frac{x}{c}. \end{cases} \quad (6.6)$$

We increase m by 1, and observe that $J_2 = J_1$, which by Theorem 4.3 implies, along with (6.5) and (6.6), that

$$(u_1, u_2) := \begin{cases} (d, 0) & \text{when } h' > \frac{x}{c} \text{ and } x > 0; \text{ or } h' > \frac{x}{b} \text{ and } x < 0, \\ (d, d) & \text{when } \frac{x}{b} < h' < \frac{x}{c}, x > 0, \\ (0, d) & \text{when } h' < \frac{x}{b} \text{ and } x > 0; \text{ or } h' \leq \frac{x}{c} \text{ and } x < 0, \\ (0, 0) & \text{when } \frac{x}{c} < h' < \frac{x}{b}, x < 0, \\ (d, u) & \text{when } h' = \frac{x}{c}, x > 0, u \in [0, d], \\ (u, d) & \text{when } h' = \frac{x}{b}, x > 0, u \in [0, d], \\ (u, 0) & \text{when } h' = \frac{x}{b}, x < 0, u \in [0, d]. \end{cases} \quad (6.7)$$

is a Nash equilibrium and that J_2 is the optimal value of the game.

Now we present the PIA for the bias game that arises from this example (see Sect. 5.1).

Step 1 Set $m = 0$, fix $u_2^m = 0$ and define $\tilde{J}_0 = -\infty$.

Step 2 Note that (5.8) becomes

$$\tilde{J}_1 = -h + \frac{\sigma^2}{2} \tilde{h}'' + \sup_{u \in [0, d]} \{b\tilde{h}'u\}, \quad (6.8)$$

where h is the function referred to in (6.4). Then

$$u_1^1 = \begin{cases} d & \text{if } \tilde{h}' > 0 \text{ and } h' > \frac{x}{b}, \\ 0 & \text{if } \tilde{h}' < 0 \text{ and } h' < \frac{x}{b}, \\ u \in [0, d] & \text{if } \tilde{h}' = 0 \text{ and } h' = \frac{x}{b}. \end{cases} \quad (6.9)$$

The function $\tilde{h}$ and the constant $\tilde{J}_1$ can be found by solving the ordinary differential equation of second order (6.8).

Step 3 Clearly $\tilde{J}_1 \neq \tilde{J}_0$. Then go to Step 4.

Step 4 We use (6.8) to determine the minimum referred to in (5.9). That is

$$u_2^1 = \begin{cases} d & \text{if } \tilde{h}' < 0 \text{ and } h' < \frac{x}{c}, \\ 0 & \text{if } \tilde{h}' > 0 \text{ and } h' > \frac{x}{c}, \\ u \in [0, d] & \text{if } \tilde{h}' = 0 \text{ and } h' = \frac{x}{c}. \end{cases} \quad (6.10)$$

We increase m by 1, and observe that $\tilde{J}_2 = \tilde{J}_1$, which by Theorem 4.3 implies, along with (6.7), (6.9)–(6.10), that

$$(u_1, u_2) := \begin{cases} (d, 0) & \text{when } \tilde{h} > 0, h' > \frac{x}{c} \text{ and } x > 0; \text{ or } \tilde{h}' > 0, h' > \frac{x}{b} \text{ and } x < 0, \\ (d, d) & \text{when } \tilde{h}' = 0, \frac{x}{b} < h' < \frac{x}{c}, x > 0, \\ (0, d) & \text{when } \tilde{h}' < 0, h' < \frac{x}{b} \text{ and } x > 0; \text{ or } \tilde{h}' < 0, h' \leq \frac{x}{c} \text{ and } x < 0, \\ (0, 0) & \text{when } \tilde{h}' = 0, \frac{x}{c} < h' < \frac{x}{b}, x < 0, \\ (d, u) & \text{when } \tilde{h}' = 0, h' = \frac{x}{c}, x > 0, u \in [0, d], \\ (u, d) & \text{when } \tilde{h}' = 0, h' = \frac{x}{b}, x > 0, u \in [0, d], \\ (u, 0) & \text{when } \tilde{h}' = 0, h' = \frac{x}{b}, x < 0, u \in [0, d]. \end{cases}$$

is a Nash equilibrium for the bias game, and that $\tilde{J}_2$ is the optimal value of the game.

7 Concluding Remarks

This work is intended to give sufficient conditions under which the PIA converges in a certain class of games. This represents a breakthrough with respect to the current literature because, nor the space state, neither the action sets are finite, or even denumerable.

The main contributions of our work are the PIA versions for SDGs presented in Sects. 3 and 5, Lemma 4.1 and Theorem 4.3. Our versions of the PIA are suitable extensions to the continuous-time scheme of that presented in [19, 45] and of algorithm (HK) in [41]. As for Lemma 4.1 and Theorem 4.3, we acknowledge that they are inspired in [19, Lemma 4.5] (see also [45, Proposition 4.1]) and [19, Theorem 4.3] (see also Theorem 4.2 in [45]), respectively. This work represents as well a continuation of our study of bias games (presented in [12]), in the sense that it provides a way of finding bias equilibria in terms of an ergodic game.

One of the main tools we used to assert the convergence of randomized stationary policies for each of the players was Theorem 4.3. Through this result we saw that the sequence of saddle points generated by the PIA, converges, in the sense of Schäl, to an equilibrium of the SDG with ergodic payoff.

We have made a rather implicit, but extensive use of the uniform ellipticity condition (2.2) on the diffusion (2.1). This particular hypothesis, along with the

other assumptions in our work allowed us to assert the existence of each of the members of the sequence $\{h_m\}$ referred to in step 2 of the PIA. This condition also enables our use of Theorem 3.4 in [27] to ensure the existence of a function h in $\mathcal{C}^2(\mathbb{R}^N) \cap \mathcal{B}_w(\mathbb{R}^N)$ such that $h_m \rightarrow h$ as $m \rightarrow \infty$.

The last key to our results is the w -exponential ergodicity referred to in (2.12). This relation gives the existence and finiteness of the bias function h_{π^1, π^2} in (3.5). Therefore (2.12) represents the *sine qua non* condition to extend the controlled version of the PIA presented in [19] to the SDGs' context.

References

- [1] Akella, R., Kumar, P.: Optimal control of production rate in a failure prone manufacturing system. *IEEE Trans. Autom. Control* **31**, 116–126 (1985)
- [2] Arapostathis, A., Borkar, V.S.: Uniform recurrence properties of controlled diffusions and applications to optimal control. *SIAM J. Control Optim.* **41**, 4181–4223 (2010)
- [3] Arapostathis, A., Ghosh, M.K., Borkar, V.S.: *Ergodic Control of Diffusion Processes*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge (2011)
- [4] Bensoussan, A.: *Perturbation Methods in Optimal Control*. Wiley, New York (1998)
- [5] Bertsekas, D.P., Shreve, S.E.: *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, Belmont (1996)
- [6] Bhattacharya, R.N.: Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.* **6**, 541–553 (1978)
- [7] Billingsley, P.: *Convergence of Probability Measures*. Wiley Series in Probability and Statistics, 2nd edn. Wiley, New York (1999)
- [8] Bismut, J.: An approximation method in optimal stochastic control. *SIAM J. Control Optim.* **16**, 122–130 (1978)
- [9] Borkar, V.S., Ghosh, M.K.: Ergodic control of multidimensional diffusions II: adaptive control. *Appl. Math. Optim.* **21**, 191–220 (1990)
- [10] Borkar, V.S., Ghosh, M.K.: Stochastic differential games; occupation measure based approach. *J. Optim. Theory Appl.* **73**, 359–385 (1992) (Correction **88**, 251–252 (1996))
- [11] Escobedo-Trujillo, B.A., López-Barrientos, J.D.: Nonzero-sum stochastic differential games with additive structure and average payoffs. *J. Dyn. Games* (2014) (to appear)
- [12] Escobedo-Trujillo, B.A., López-Barrientos, J.D., Hernández-Lerma, O.: Bias and Overtaking equilibria for zero-sum stochastic differential games. *J. Optim. Theory Appl.* **153**, 662–687 (2012)
- [13] Fan, K.: Fixed point and minimax theorems in locally convex topological linear spaces. *Proc. Natl. Acad. Sci. USA* **38**, 121–126 (1952)
- [14] Fleming, W.H.: Some Markovian optimization problems. *J. Math. Mech.* **12**, 131–140 (1963)
- [15] Ghosh, M.K., Arapostathis, A., Marcus, S.I.: Optimal control of switching diffusions to flexible manufacturing systems. *SIAM J. Control Optim.* **31**, 1183–1204 (1993)
- [16] Ghosh, M.K., Arapostathis, A., Marcus, S.I.: Ergodic control of switching diffusions. *SIAM J. Control Optim.* **35**, 1962–1988 (1997)
- [17] Gilbarg, D., Trudinger, N.S.: *Elliptic Partial Differential Equations of Second Order*. Springer, Heidelberg (1998). (Reprinted version)
- [18] Glynn, P.W., Meyn, S.P.: A Lyapunov bound for solutions of the Poisson equation. *Ann. Probab.* **24**, 916–931 (1996)
- [19] Hernández-Lerma, O., Lasserre, J.B.: Policy iteration for Markov control processes on Borel spaces. *Acta Appl. Math.* **47**, 125–154 (1991)
- [20] Hernández-Lerma, O., Lasserre, J.B.: *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York (1996)
- [21] Howard, R.A.: *Dynamic Programming and Markov Processes*. MIT Press, Cambridge (1960)
- [22] Jasso-Fuentes, H.: Infinite-horizon optimal control problems for Markov diffusion processes. Ph.D. dissertation. Mathematics Department, CINVESTAV-IPN, México (2007)

- [23] Jasso-Fuentes, H., Hernández-Lerma, O.: Characterizations of overtaking optimality for controlled diffusion processes. *Appl. Math. Optim.* **57**, 349–369 (2008)
- [24] Jasso-Fuentes, H., López-Barrientos, J.D.: On the use of stochastic differential games against nature to ergodic control problems with unknown parameters. *Int. J. Control* (2014). doi:[10.1080/00207179.2014.984764](https://doi.org/10.1080/00207179.2014.984764)
- [25] Klebaner, F.C.: Introduction to Stochastic Calculus with Applications, 2nd edn. Imperial College Press, London (2005)
- [26] Kushner, H.J.: Numerical approximations for stochastic differential games. *SIAM J. Control Optim.* **41**, 457–486 (2002)
- [27] López-Barrientos, J.D.: Basic and advanced optimality criteria for zero-sum stochastic differential games. Ph. D. dissertation. Department of Mathematics, CINVESTAV-IPN, México. <http://www.math.cinvestav.mx/sites/default/files/tesis-daniel-2012> (2012)
- [28] Mendoza-Pérez, A.F., Hernández-Lerma, O.: Markov control processes with pathwise constraints. *Math. Methods Oper. Res.* **71**, 477–502 (2010)
- [29] Meyn, S.P., Tweedie, R.L.: Stability of Markovian processes. III. Foster–Lyapunov criteria for continuous-time processes. *Adv. Appl. Probab.* **25**, 518–548 (1993)
- [30] Morimoto, H., Okada, M.: Some results on the Bellman equation of ergodic control. *SIAM J. Control Optim.* **38**, 159–174 (1999)
- [31] Nowak, A.S.: Measurable selection theorems for minimax stochastic optimization problems. *SIAM J. Control Optim.* **23**, 466–476 (1985)
- [32] Nowak, A.S.: Sensitive equilibria for ergodic stochastic games with countable state spaces. *Math. Methods Oper. Res.* **50**, 65–76 (1999)
- [33] Nowak, A.S.: Optimal strategies in a class of zero-sum ergodic stochastic games. *Math. Methods Oper. Res.* **50**, 399–419 (1999)
- [34] Prieto-Rumeau, T., Hernández-Lerma, O.: Bias and overtaking equilibria for zero-sum continuous-time Markov games. *Math. Methods Oper. Res.* **61**, 437–454 (2005)
- [35] Puterman, M.L.: Optimal control of diffusion processes with reflection. *J. Optim. Theory Appl.* **22**, 103–116 (1977)
- [36] Puterman, M.L.: On the convergence of policy iteration for controlled diffusions. *J. Optim. Theory Appl.* **33**, 137–144 (1981)
- [37] Rogers, L.C.G., Williams, D.: Diffusions, Markov Processes and Martingales, vol. 1. Wiley, Chichester (1994)
- [38] Schäl, M.: A selection theorem for optimization problems. *Arch. Math.* **XXV**, 219–224 (1974)
- [39] Schäl, M.: Conditions for optimality in dynamic programming and for the limit of n -stage optimal policies to be optimal. *Z. Wahrs. Verw. Geb* **32**, 179–196 (1975)
- [40] Sion, M.: On general minimax theorems. *Pac. J. Math.* **8**, 171–176 (1958)
- [41] Van der Wal, J.: Discounted Markov games: generalized policy iteration method. *J. Optim. Theory Appl.* **25**, 125–138 (1978)
- [42] Zhu, Q.: Bias optimality and strong n ($n = -1, 0$) discount optimality for Markov decision processes. *J. Math. Anal. Appl.* **334**, 576–592 (2007)
- [43] Zhu, Q.: Average optimality for continuous-time Markov decision processes with a policy iteration approach. *J. Math. Anal. Appl.* **339**, 691–704 (2008)
- [44] Zhu, Q., Prieto-Rumeau, T.: Bias and overtaking optimality for continuous-time jump Markov decision processes in Polish spaces. *J. Appl. Probab.* **45**, 417–429 (2008)
- [45] Zhu, Q., Yang, X., Huang, C.: Policy iteration for continuous-time average reward Markov decision processes in Polish spaces. *Abstr. Appl. Anal.* **2009**, 1–17 (2009). doi:[10.1155/2009/103723](https://doi.org/10.1155/2009/103723)