Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games

Víctor Bucarey López, Eugenio Della Vecchia, Alain Jean-Marie, Fernando Ordoñez

Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games

Víctor Bucarey López, Eugenio Della Vecchia, Alain Jean-Marie, Fernando Ordoñez

***Abstract*—In this work we study Stackelberg equilibria for discounted stochastic games. We consider two solution concepts for these games: Stationary Strong Stackelberg Equlibrium (SSSE) and Fixed Point Equilibrium (FPE) solutions. The SSSE solution is obtained by explicitly solving the Stackelberg equilibrium conditions, while the FPE can be computed efficiently using value or policy iteration algorithms. However, previous work has overlooked the relationship between these two different solution concepts. Here we investigate the conditions for existence and equivalence of these solution concepts. Our theoretical results prove that the FPE and SSSE exist and coincide for important classes of games, including Myopic Follower Strategy and Team games. This however does not hold in general and we provide numerical examples where one of SSSE or FPE does not exist, or when they both exist, they differ. Our computational results compare the solutions obtained by value iteration, policy iteration and a mathematical programming formulations for this problem. Finally, we present a discounted stochastic Stackelberg game for a security application to illustrate the solution concepts and the efficiency of the algorithms studied.**

***Index Terms*—Stochastic Games, Stackelberg Equilibrium, Optimal Control**

I. Introduction

Stackelberg games model interactions between strategic agents, where one agent, the leader, can enforce a commitment to a strategy and the remaining agents, referred to as followers, take that decision into account when selecting their own strategies. This Stackelberg game interaction has been extended to a multistage setting where leader and followers repeatedly make strategic decisions [1].

Such dynamic Stackelberg models have been considered in applications in economics [2], marketing and supply chain management [3], dynamic congestion pricing [4], and security [5]. For example, in a dynamic security application, a defender could decide on a strategy to patrol a number of targets over multiple periods and the attackers would take this defender patrol into consideration when deciding where to attack in each period. Stochastic games are dynamic games with a state transition probability function. Here we consider that players select stationary policies for a stochastic game. In such policies the strategies selected depend only on the state and not the period. Such stationary policies are natural for games that are played over a long period (even optimal in certain cases [6]) and are useful in practice as they are easy to communicate and implement, something important in sensitive applications such as security. Therefore, we assume a Leader commits to an optimal stationary policy. Then, a Follower observes the Leader's policy and the Markov chain it induces and responds optimally with its own stationary strategy obtained by solving the observed Markov Decision Process (MDP).

The equilibrium solutions of Stackelberg games and Stackelberg stochastic games, are characterized with a system of optimality conditions that are non-linear and non convex in general. Previous work has computed such equilibrium solutions for specific problems with specialized methods. In the specific case of Stackelberg security games, bi-level mathematical optimization formulations have been used to compute the strong Stackelberg equilibrium solution [7] and, in the case of stochastic games, the Strong Stackelberg Equilibrium in *Stationary strategies* (SSSE) [5]. An alternative method determines the SSSE by solving a non-linear potential game formulation [8]. These solution methods are either tailored for specific problems or only capable of solving small instances.

The motivation for this work is the development of efficient algorithms to compute the SSSE. In particular, we consider the relation of SSSE to fixed points of suitably defined *dynamic programming operators*. The formalism of operators is used in the solution of MDPs and zero-sum stochastic games. It provides two advantages: 1) it can be used for proofs of the existence of solutions and 2) it leads to computationally efficient iterative algorithms, such as value iteration (VI) and policy iteration (PI), and proofs of their convergence. By applying the formalism of operators, we obtain new results including: a) proofs of existence of SSSE; b) examples where SSSE does not exist, which have not been reported in the literature; c) proofs of convergence to an SSSE of efficient algorithms (VI and PI) for certain classes of games.

In the remainder of this introduction we present in detail the problem under consideration (Section I-A), present related literature (Section I-B), describe the contributions of our work (Section I-C) and introduce the notation used (Section I-D).

A. Problem Statement

To describe the concept of Strong Stackelberg Equilibria in Stationary policies (SSSE) for stochastic games, consider a dynamic system evolving in discrete time on a finite set of states and actions, where two players control the evolution. Players have a perfect information on the state of the system. One of them, called *Leader* or Player A, observes the current

Víctor Bucarey is with the Institute of Engineering Sciences at the Universidad de O'Higgins, Chile. Eugenio Della Vecchia is with Departmento de Matemática at the Universidad Nacional de Rosario, Argentina. Alain Jean-Marie is with Inria, France. Fernando Ordoñez is with Departamento de Ingeniería Industrial at the Universidad de Chile, Chile.

email: victor.bucarey@uoh.cl; eugenio@fceia.unr.edu.ar; alain.jean-marie@inria.fr; fordon@dii.uchile.cl.

Author version, submitted to *IEEE Trans. Automatic Control* on October 27, 2021.

state s and commits to a, possibly mixed, stationary strategy f that depends solely on the state s. Then the other player, called *Follower* or Player B, observes the state and *strategy* of Player A and plays his best response denoted by g. Given the selected strategies at every state s there is a one-step reward for each player ($r_A^{fg}(s)$ and $r_B^{fg}(s)$ for player A and B, respectively) and a random transition probability $Q^{fg}(s'|s)$ to state s'.

Aggregated payoffs for both players are evaluated with the expected total discounted revenue over the infinite horizon, each player having their own discount factor. The aim for the leader is to find a policy that, in each state, maximizes her revenue taking into account that the follower will observe this policy and will respond by optimizing his own payoff. This general "Stackelberg" approach to the solution of the game is complemented with the rule that if the follower is indifferent between strategies, he chooses the one that benefits the leader: this refinement is the *strong* Stackelberg solution.

B. Related bibliography

The study of Strong Stackelberg Equilibria (SSE) has received much attention in the recent literature due to its relevance in security applications [7]. In static games, the need to generalize the standard Stackelberg equilibrium has been pointed out by Leitman [9] who introduced a conservative version of it. This generalization is formalized in [10] as the weak Stackelberg equilibrium, together with the definition of the optimistic generalization, the strong Stackelberg equilibrium. The relationship between SSE and bi-level optimization appears in [10], solution methods for static Stackelberg Games are discussed in [11].

Stackelberg equilibria in multi-stage and dynamic games have been studied in [12], [13]. In particular, authors propose in [13] to focus on *feedback* strategies that can be obtained via dynamic programming. The idea is reused in [10] which introduces strong *sequential* Stackelberg equilibria, in a setting similar to ours. The notable difference is that, in the problem they consider, the follower gets to observe the *action* of the leader, not just its strategy. In their analysis, the formalism of operators linked to dynamic programming, introduced by Denardo [14] and developed in [15], is essential. Our analysis uses this formalism as well.

The stochastic game model we study in this paper is also the topic of [16]. Although they do not provide any formal definition of SSSE, the authors show that SSSE always exist in stochastic games for *team* games where both players have the same rewards by reducing the game to an MDP. They also propose mathematical programming formulations to find the SSSE, extending to stochastic Stackelberg games the analysis used for MDPs (see [17, ch. 6.]) and Nash equilibrium in stochastic games [18]. Similar mathematical programming formulations are established in [5] and [19] for problems in security applications. However, no prior work has provided a proof of the relationship between the solutions of these mathematical programming formulations and the SSSE of the stochastic game being considered. In this paper we present conditions that guarantee the existence of SSSE for stochastic games in diverse classes of problems, including team games. Furthermore we present numerical examples that suggest that the mathematical programming formulation computes the SSSE solution when it exists.

The complexity of computing an SSE is studied in [20] showing that it is NP-hard to determine an SSE for a stochastic Stackelberg game with any discount factor $\beta > 0$ common for both players. The possibility that a stochastic Stackelberg game does not have an SSE is not considered.

As mentioned above, security applications are an important motivation for research on dynamic games. An attacker-defender Stackelberg security game is also considered in [8] for a particular stochastic Markov chain game. The setting is less general than ours, as it is restricted to zero-sum games and ergodic Markov chains. In addition the computational results presented show solutions only for instances with few states. Another related work is the stochastic game model described in [18, Chapter 6.3], where the authors are interested in the average reward and the solution concept used is the Nash Equilibrium, a choice different from ours. However, the model has the feature that only one player, the defender, controls the transitions between states. This feature is one of the properties that guarantees the existence of SSSE, as we will show later.

C. Contribution

While previous work has formulated Stackelberg equilibrium for stochastic games and considered different solution methods to compute the SSSE, to the best of our knowledge the general question of the existence of SSSE is still largely open, in the sense that, so far: a) no case of non-existence is reported, b) few sufficient conditions for existence have been established. Here we address this gap using the formalism of operators for the mathematical analysis of the problem. Our analysis gives proofs of the existence of SSSE for important classes of games and provides efficient solution methods.

The rest of this paper is structured as follows: First, we give a formal definition of the Strong Stationary Stackelberg Equilibrium (SSSE) in stochastic games (Section I-D). In Section II we develop the operator-based analysis of such games by introducing an operator acting on the space of value functions. The operator introduced is related to the one-step evaluation of each player's payoff. We then define *Fixed Point Equilibria* (FPE) as the fixed points of this operator. Next, we introduce the class of games with *Myopic Follower Strategy* (MFS), for which specific operators are relevant. We prove that these operators are contractive. Finally, we introduce the algorithms for computing FPE, for general games and for games with MFS. We prove the convergence of both value iteration and policy iteration to the FPE of games with MFS. We also recall the mathematical programming formulation for SSSE.

Next, we focus on the general question of existence of SSSE and FPE, and how they are related. We prove that games with MFS and Team Games have both SSSE and FPE and that they coincide. The operator formalism is instrumental in this proof. We also address the classes of Zero-Sum Games and Acyclic Games. This analysis is developed in Section III.

In Section IV we provide examples to illustrate different situations. In a first case, an FPE and an SSSE exist and coin-

cide, although the game does not have MFS (the assumption of our main existence result). In a second case, depending on the parameters: either no SSSE exists, or no FPE exists, or both an SSSE and an FPE exist but they do not coincide. In a third case, an FPE exists but value iteration does not necessarily converge to it. Finally, in Section V, we take advantage of the convergence properties we have shown, to propose a solution methodology in a dynamic security game, representing the problem of security patrols in a network.

D. Notation and definitions

Let $\mathcal{S}$ represent the finite set of states of the game. Let $\mathcal{A}, \mathcal{B}$ denote the finite set of actions available to players A and B respectively, and we denote by $\mathcal{A}_s \subset \mathcal{A}$ and $\mathcal{B}_s \subset \mathcal{B}$ the actions available in state $s \in \mathcal{S}$. For a given state $s \in \mathcal{S}$ and actions $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$, $Q^{ab}(s'|s)$ represents the transition probabilities of reaching the state $s' \in \mathcal{S}$. We denote with Q the family of these probability distributions. The reward received by each player in state s when selecting actions $a \in \mathcal{A}_s$ and $b \in \mathcal{B}_s$ is referred to as the one-step reward functions and are given by $r_A = r_A^{ab}(s)$ and $r_B = r_B^{ab}(s)$. The constants $\beta_A, \beta_B \in [0, 1)$ are discount factors for Player A and B respectively. In our setting time increases discretely and the time horizon is infinite. Therefore we represent a two-person stochastic discrete game $\mathcal{G}$ by

$$\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, Q, r_A, r_B, \beta_A, \beta_B) \ .$$

a) Strategies: We denote by $\mathbb{P}(\mathcal{A}_s)$ and $\mathbb{P}(\mathcal{B}_s)$ the sets of distribution functions over $\mathcal{A}_s$ and $\mathcal{B}_s$, respectively. The sets of stationary strategies are defined by:

$$W_A = \{f : \mathcal{S} \to \mathbb{P}(\mathcal{A}) | \, f(s) \in \mathbb{P}(\mathcal{A}_s)\}$$
$$W_B = \{g : \mathcal{S} \to \mathbb{P}(\mathcal{B}) | \, g(s) \in \mathbb{P}(\mathcal{B}_s)\} \ .$$

For $f \in W_A$, $f(s)$ is a probability measure on $\mathcal{A}_s$. In order to simplify the notation, we represent with $f(s, a) = (f(s))(\{a\}) = f(a|s)$ the probability that Player A chooses action a when in state s. Likewise, for $g \in W_B$, we denote with $g(s, b) = g(b|s)$ the probability that Player B chooses b when in state s. In the case that $g \in W_B$ is a deterministic policy, we will denote directly with $g(s)$ the element of $\mathcal{B}_s$ that has probability one. The notation will be clear from context. The set W_B is assumed to be equipped with a total order $\prec_B$, which will be used for determining a unique element in case Player B is indifferent between several policies.

In order to simplify notation, given stationary strategies f and g, we define the reward for player $i(= A, B)$ by:

$$r_i^{fg}(s) = \sum_{a \in \mathcal{A}_s} \sum_{b \in \mathcal{B}_s} f(s, a) g(s, b) r_i^{ab}(s). \tag{1}$$

b) Values: Given a pair $(f, g) \in W_A \times W_B$, the evolution of the states is that of a Markov chain on $\mathcal{S}$ with transition probabilities

$$Q^{fg}(s'|s) = \sum_{a \in \mathcal{A}_s} \sum_{b \in \mathcal{B}_s} f(s, a) g(s, b) Q^{ab}(s'|s).$$

Denote with $\{S_n\}_n$ the (random) sequence of states of this Markov chain and $\mathbb{E}_s^{fg}$ the expectation corresponding to the distribution of this sequence, conditioned on the initial state being $S_0 = s$. Then the value of this pair of strategies for Player i, from state s, is:

$$V_i^{fg}(s) = \mathbb{E}_s^{fg}\left[\sum_{k=0}^{\infty} \beta_i^k r_i^{f(S_k), g(S_k)}(S_k)\right] \tag{2}$$
$$= \mathbb{E}_s^{fg}\left[\sum_{k=0}^{\infty} \beta_i^k \sum_{a \in \mathcal{A}_{s_k}} \sum_{b \in \mathcal{B}_{s_k}} f(S_k, a) g(S_k, b) r_i^{ab}(S_k)\right] \ .$$

c) Reaction sets: We proceed with the definition of the player's reaction sets. These definitions rely heavily on the fact that when the leader selects a stationary strategy, the follower faces a finite-state, finite-action, discounted MDP. It is then well-known that there exists optimal stationary and deterministic policies which maximize simultaneously the follower's values starting from any state. Moreover, the set of optimal policies is the cartesian product of the set of optimal decisions in each state. This fact results from *e.g.* Corollary 6.2.8, p. 153 in [17].

Accordingly, let:

$$R_B(f) := \{g \in W_B \mid V_B^{fg}(s) \geq V_B^{fh}(s), \forall s \in \mathcal{S}, h \in W_B\} \cap \prod_{s \in \mathcal{S}} \{0, 1\}^{|\mathcal{B}_s|} \tag{3}$$

$$SR_B(f) := \{g \in R_B(f) \mid V_A^{fg}(s) \geq V_A^{fh}(s), \forall s \in \mathcal{S}, h \in R_B(f)\} \tag{4}$$

$$\gamma_B(f) := \max_{\prec_B} SR_B(f)$$

$$R_A(s) := \{f \in W_A \mid V_A^{f\gamma_B(f)}(s) \geq V_A^{h\gamma_B(h)}(s), \forall h \in W_A\} \ . \tag{5}$$

Given that Player A selects strategy f, $R_B(f)$ represents the set of deterministic best-response strategies of Player B. As argued above, this set is nonempty. The set $SR_B(f)$ is that of *strong* best-responses, which break ties in favor of Player A. It is possible to break ties simultaneously in all states s, because optimal policies of the MDP form a cartesian product. We denote by $\gamma_B(f)$ the deterministic policy that is the actual best response of Player B to Player A's f. Finally, $R_A(s)$ is the set of Player A's best strategies when starting from state s. Observe that the inequality involved in (5) compares $V_A^{f\gamma_B(f)}(s)$, the value Player A can get by choosing strategy f, with $V_A^{h\gamma_B(h)}(s)$, the value of another strategy h. This is a way of specifying that Player A wants to maximize her expected gain. In contrast, (4) refers to the comparison of $V_A^{fg}(s)$, the value A gets when playing f if B plays g, with $V_A^{fh}(s)$, where A still plays f but B plays another strategy h. This is a way of specifying that B maximizes A's value knowing that she plays f (the 'strong' requirement).

d) Equilibria: With these notations, we can now define Strong Stackelberg Equilibria of the dynamic game, called here Stationary SSE, as the SSE for the static game where players use stationary strategies in $W_A \times W_B$.

Definition I.1 (SSSE)**.** A strategy pair $(f, g) \in W_A \times W_B$ is a Stationary Strong Stackelberg Equilibrium if

i/ $g = \gamma_B(f)$;

ii/ $f \in R_A(s)$ for all $s \in \mathcal{S}$.

In an SSSE, the strategy f maximizes *simultaneously* the leader's reward in every state. In contrast with MDP where this is always possible, there is no guarantee that this will happen in a Stackelberg stochastic game. Indeed, in Section IV-B we provide an example where $\cap_s R_A(s)$ is empty, and consequently there is no SSSE.

To the best of our knowledge, the literature does not provide general statements about the existence of an SSSE. We address in Section III this issue in special cases.

II. Operators, Fixed Points and Algorithms

In this section, we develop the formalism of operators, as commonly found in texts on MDPs [17], and also for games in [10], [14], [15]. We focus on fixed points of these operators, as a means to discuss existence of equilibria, and also as a computational procedure. Accordingly, we study the monotonicity and contractivity of these operators. This allows us to prove the convergence of value iteration, policy iteration and mathematical programming-based algorithms, in certain situations. Most of the proofs in this section are quite standard in the literature of stochastic processes. We will omit them providing a reference for interested readers.

A. Definition of operators

We start with the definition of one-step (or "return function" [14]) operators. The set of value functions, i.e. mappings from $\mathcal{S}$ to $\mathbb{R}$, will be denoted with $\mathcal{F}(\mathcal{S})$. Given $(f,g) \in W_A \times W_B$ we define $T_i^{fg} : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$, such that

$$\left(T_i^{fg} v\right)(s) = \sum_{a \in \mathcal{A}_s} f(s,a) \sum_{b \in \mathcal{B}_s} g(s,b) \left[r_i^{ab}(s) + \beta_i \sum_{z \in \mathcal{S}} Q^{ab}(z|s) v(z) \right] .$$

It is important to note that the value $(T_i^{fg} v)(s)$ depends only on $f(s)$ and $g(s)$, and not on the rest of the strategies f and g. In the following, with a slight abuse of notation, we will use this quantity for values of f and g specified only at state s.

The set of pairs of value functions is $\mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$. A typical element of it will be denoted as $v = (v_A, v_B)$. Using T_A^{fg} and T_B^{fg} we define the operator T^{fg} on $\mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$ as:

$$(T^{fg} v)_i = T_i^{fg} v_i$$

for $i = A, B$.

It will be recalled in Lemma II.1 that T_i^{fg} is a contraction for $i = A, B$. It follows that T^{fg} is contractive as well. As a consequence of Banach's theorem, it admits a unique fixed point on the complete space $\mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$ with the supremum norm, that turns out to be $V^{fg} = (V_A^{fg}, V_B^{fg})$, these functions being defined in (2).

a) Extended reaction sets: We now extend the definitions of reaction sets to involve value functions, in a way similar as in [10], [14], [15]. They correspond to a dynamic game with only one step and a "scrap value" $v = (v_A, v_B)$. In contrast to the sets introduced in Section I-D for SSSE, the sets we discuss here are relative to *local* strategies depending on each state, rather than *global* strategies in W_A and W_B.

For $s \in \mathcal{S}$, $f \in \mathbb{P}(\mathcal{A}_s)$, $v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$, and $v_B \in \mathcal{F}(\mathcal{S})$, let:

$$R_B(s,f,v_B) := \{g \in \mathcal{B}_s \mid (T_B^{fg} v_B)(s) \geq (T_B^{fh} v_B)(s), \forall h \in \mathcal{B}_s\} \quad (6)$$

$$SR_B(s,f,v) := \{g \in R_B(s,f,v_B) \mid (T_A^{fg} v_A)(s) \geq (T_A^{fh} v_A)(s), \forall h \in R_B(s,f,v_B)\} \quad (7)$$

$$\gamma_B(s,f,v) := \max_{\prec_B} SR_B(s,f,v) \quad (8)$$

$$R_A(s,v) := \{f(s) \in \mathbb{P}(\mathcal{A}_s) \mid (T_A^{f\gamma_B(s,f,v)} v_A)(s) \geq (T_A^{h\gamma_B(s,h,v)} v_A)(s), \forall h \in \mathbb{P}(\mathcal{A}_s)\} . \quad (9)\ (10)$$

The definition of Player B's response in (8) is such that one unique, non-ambiguous policy is defined as a solution. Any $f(s) \in R_A(s,v)$ is considered as a solution of the problem.

b) The dynamic programming operator: The one-step Strong Stackelberg problem naturally leads to a mapping in the space of value functions, which is formalized as follows.

Definition II.1 (Dynamic programming operator T)**.** Let T: $\mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$ be defined as:

$$(Tv)_i(s) = \left(T_i^{R_A(s,v), \gamma_B(s, R_A(s,v), v)} v_i\right)(s) \quad (11)$$

for $i = A, B$.

Observe that the definition depends on the ordering $\prec_B$. By changing the ordering, many operators can be defined for the same problem.

c) Fixed points: We are now in position to define the fixed-point equilibria.

Definition II.2 (Fixed Point Equilibrium FPE)**.** A strategy pair $(f,g) \in W_A \times W_B$ is an FPE if $v^* = V^{fg}$, the unique fixed point of T^{fg}, is such that $Tv^* = v^*$. Equivalently, if

i/ $g(s) = \gamma_B(s,f,v^*)$ for all $s \in \mathcal{S}$;
ii/ $f(s) \in R_A(s,v^*)$ for all $s \in \mathcal{S}$.

B. Properties of operators

The following property is well-known (e.g. [17]) for finite-state, finite-action discounted Markov Reward Processes:

Lemma II.1. For $i = A, B$, the operator T_i^{fg} is linear, monotone, contractive and V_i^{fg} defined in (2) is its unique fixed point. This fixed point has the expression

$$V_i^{fg} = (I - \beta_i Q^{fg})^{-1} r_i^{fg} , \quad (12)$$

where r_i^{fg} is defined in (1), where the probability transition matrix Q^{fg} is defined similarly in Section I-D, and I is the identity matrix of appropriate dimension.

We now introduce a particular class of games, and the particular properties of operators for these games.

Definition II.3 (Myopic Follower Strategy, MFS)**.** A stochastic game $\mathcal{G}$ is said to be with *Myopic Follower Strategy* if $R_B(s, f, v_B)$ does not depend on v_B. In this case, we denote the extended reaction set as $\overline{R}_B(s, f)$.

Games with *Myopic Follower Strategy* are games where the response of the follower is independent of the expected future values: for her, only immediate rewards are relevant. As it is stated in Lemma III.2, this setting happens whenever $\beta_B = 0$ (as it is used in [5]) or the leader controls the transitions of the games, see for e.g. single-controller games in [18]. When a game is with MFS, the reaction of the follower depends only on the leader's value v_A. This can be expressed as follows:

$$\gamma_B(s, f, v) = \overline{\gamma}_B(s, f, v_A) \quad \forall f \in W_A . \forall v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S}), \forall s \in \mathcal{S}, \quad (13)$$

where $\overline{\gamma}_B$ is the best response operator γ_B restricted only to the values of v_A in its third component. Equation (13) derives from the fact that neither SR_B nor γ_B will depend on v_B. Then the following lemma is relevant.

Lemma II.2. Assume (13) holds. Then there exists an operator $\overline{T}_A$ from $\mathcal{F}(\mathcal{S})$ to $\mathcal{F}(\mathcal{S})$ such that for all $v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$,

$$(Tv)_A = \overline{T}_A v_A. \quad (14)$$

Proof: According to Definition (10) and due to (13), we have $R_A(s, v) = \overline{R}_A(s, v_A)$ for all $v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$ and $s \in \mathcal{S}$. Then, from (11),

$$\begin{aligned}(Tv)_A(s) &= (T_A^{R_A(s,v), \gamma_B(s, R_A(s,v), v)} v_A)(s) \\ &= (T_A^{\overline{R}_A(s,v_A), \overline{\gamma}_B(s, \overline{R}_A(s,v_A), v_A)} v_A)(s) \\ &=: (\overline{T}_A v_A)(s).\end{aligned}$$

■

An alternate construction of operator $\overline{T}_A$ is as follows. It is possible to define, for each $f \in W_A$, the operator $\overline{T}_A^f$ from $\mathcal{F}(\mathcal{S})$ to $\mathcal{F}(\mathcal{S})$ as:

$$(\overline{T}_A^f v_A)(s) = (T_A^{f, \overline{\gamma}_B(s,f,v_A)} v_A)(s) \ . \quad (15)$$

Another consequence of MFS is this property which follows from the definition of $\gamma_B(\cdot)$ in (8) and $SR_B(\cdot)$ in (7):

$$(\overline{T}_A^f v_A)(s) = \max_{g \in \overline{R}_B(s,f)} (T_A^{fg} v_A)(s) \ . \quad (16)$$

In this equation, the maximization set on the right-hand side does not depend on v at all. Finally, define the operator $\overline{T}_A$ from $\mathcal{F}(\mathcal{S})$ to $\mathcal{F}(\mathcal{S})$ as, for all $s \in \mathcal{S}$:

$$(\overline{T}_A v_A)(s) = \max_{f \in W_A} (\overline{T}_A^f v_A)(s) \ . \quad (17)$$

Observe that the maximum is indeed attained, because the right-hand side is a linear combination of the finite set of values $f(s, a)$, $a \in \mathcal{A}_s$.

We can now state the principal tool of this paper for ascertaining the existence of FPE.

Theorem 1. Let $\mathcal{G}$ be a stochastic game with MFS, then it is true that:

a) For any stationary strategy $f \in W_A$, the operator $\overline{T}_A^f : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$, defined in (15) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$ of modulus β_A.
b) The operator $\overline{T}_A$ defined in (17) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$, of modulus β_A.
c) For any stationary strategy $f \in W_A$, operator $\overline{T}_A^f$ is monotone. Operator $\overline{T}_A$ is monotone as well.

Proof: See Theorem 1 in [21]. ■

We conclude with a result of general use in the remainder.

Proposition 1. If a function $v_A \in \mathcal{F}(\mathcal{S})$ satisfies $v_A \leq \overline{T}_A v_A$, then $v_A \leq v_A^*$, where v_A^* is the unique fixed point of $\overline{T}_A$ in $\mathcal{F}(\mathcal{S})$.

Proof: By hypothesis and Theorem 1 c) we have that

$$v_A \leq \overline{T}_A v_A \leq (\overline{T}_A)^2 v_A \ \leq \cdots \leq (\overline{T}_A)^n v_A.$$

Then by Theorem 1 b), $(\overline{T}_A)^n v_A \longrightarrow v_A^*$ when $n \to \infty$. The result follows. ■

C. Value iteration Algorithms

Value iteration (VI) generally consists in applying a dynamic programming operator to some initial value function, until some convergence criterion is met. Specifically, given some $\varepsilon > 0$, Value iteration applies some operator repeatedly until the distance between functions v^n and v^{n+1} is less than ε.

In view of the preceding discussion, two variants of the algorithm will be used: one for the general situation using operator T as in (11); and one for the specific situation of MFS, using $\overline{T}_A$ as in (14). Depending on the situation, (VI) works over $\mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$ or $\mathcal{F}(\mathcal{S})$ respectively. These two variants are summarized in Algorithm 1.

Algorithm 1 Value function Iteration (VI)

Require: $\varepsilon > 0$
1: Initialize with $n = 0$, $v^0(s) = 0$ for every $s \in \mathcal{S}$
2: **repeat**
3: $n := n + 1$
4: Compute v^n as $v^n(s) := (Tv^{n-1})(s)$, $s \in \mathcal{S}$ with T as in (11) or $\overline{T}_A$ as in (14) for games with MFS.
5: **until** $||v^n - v^{n-1}||_\infty \leq \varepsilon$
6: Pick (f^*, g^*) such that $v^n(s) = (T^{f^* g^*} v^{n-1})(s)$ for all $s \in \mathcal{S}$
7: **return** Approximate Stationary Strong Stackelberg policies (f^*, g^*)

Each iteration of Algorithm 1 requires to find a static strong Stackelberg equilibrium in a bimatrix game of size $|\mathcal{A}_s| \times |\mathcal{B}_s|$ for each $s \in \mathcal{S}$. This task can be performed in polynomial time by solving a linear optimization problem [22].

There is no guarantee in general that Algorithm 1 will converge, and we present in Section IV-C an example where it does not. However, thanks to Theorem 2, we can state that

this algorithm does converge for games with MFS by using operator $\overline{T}_A$.

Theorem 2. Let $\mathcal{G}$ be a stochastic game with MFS. Then the sequence of value functions v_A^n in Algorithm 1 converges to v_A^*, which is the fixed point of $\overline{T}_A$. Moreover the following bounds hold:

$$||v_A^* - v_A^n||_\infty \leq \frac{2\beta_A^n ||r_A||_\infty}{1-\beta_A} \qquad \text{for any } n \in \mathbb{N},$$
$$||v_A^* - V_A^{f^* g^*}||_\infty \leq \frac{2\beta_A \varepsilon}{1-\beta_A} \ .$$

Proof: See Theorem 2 in [21]. ■

D. Policy iteration

The policy iteration (PI) algorithm directly iterates in the policy space. This algorithm starts with an arbitrary policy f and then finds the optimal infinite discounted horizon values, taking into account the optimal response $g(f)$. These values are then used to compute new policies. These two steps of the algorithm can be defined as the *Evaluation Phase* and the *Improvement Phase*.

As in the previous section, two variants of the algorithm will be used: one for the general situation with operators T and function γ_B; and one for the specific situation of MFS, with operators $\overline{T}_A$ and function $\overline{\gamma}_B$. The algorithm is presented in Algorithm 2).

Algorithm 2 Policy Iteration (PI)

1: Require $\varepsilon > 0$
2: Initialize with $n = 0$
3: Choose an arbitrary pair of strategies $(f_0, g_0) \in W_A \times W_B$ with $g_0(s) = \gamma_B(s, f_0, \mathbf{0})$ for all $s \in \mathcal{S}$
4: Compute $u^0 = (u_A^0, u_B^0)$ fixed point of $T^{f_0 g_0}$
5: **repeat**
6: $n := n + 1$
7: **Improvement Phase**: Find a pair of strategies (f_n, g_n) such that $T^{f_n g_n} u^{n-1} = T u^{n-1}$ with $g_n(s) = \gamma_B(s, f_n, u^{n-1})$ for all $s \in \mathcal{S}$
8: **Evaluation Phase**: Find $u^n = (u_A^n, u_B^n)$, fixed point of the operator $T^{f_n g_n}$
9: **until** $||u^n - u^{n-1}||_\infty \leq \varepsilon$
10: $f^* := f_n$; $g^*(s) := \gamma_B(s, f_n, u^n)$ for all $s \in \mathcal{S}$
11: **return** Approximate Stationary Strong Stackelberg policies (f^*, g^*)

The Evaluation Phase in Algorithm 2 requires to solve two or one linear systems of size $|\mathcal{S}| \times |\mathcal{S}|$, depending whether the game has MFS or not. On the other hand, the Improvement Phase can be implemented by solving a static Strong Stackelberg equilibrium for each state $s \in \mathcal{S}$. As in Algorithm 1, this task can be performed in polynomial time by solving a linear optimization problem.

Now we prove that Algorithm 2 converges to the FPE when (13) holds. In other words, the (PI) algorithm converges to the FPE for stochastic games with MFS. Moreover, in the next section (Theorem 6) we prove that FPE and SSSE coincide in games with MFS. Furthermore, our computational results shows that (PI) is more efficient in computational times (VI) and mathematical programming formulations. So the following theorem give us a powerful algorithmic tool to compute the SSSE for games with MFS.

Theorem 3. Suppose that Condition (13) holds. The sequence of functions u_A^n in Algorithm 2 converges monotonically to v_A^* . Further, if for any $n \in \mathbb{N}$, $u_A^n = u_A^{n+1}$, then it is true that $u_A^n = v_A^*$.

Proof: See Theorem 3 in [21]. ■

The results exposed in this section strongly rely on the fact that $\gamma_B(s, f, v)$ is independent on v_B. In Section III we show that MFS is a sufficient condition for the existence of an FPE but all the results here may fail in the general case.

E. Mathematical Programming Formulations

In this section we develop the discussion of Mathematical Programming (MP) formulations, as the one proposed in [16]. To start the discussion we notice that for each $f \in W_A$ the follower solves an MDP with transition and rewards given by the expectation induced by f. Then, as argued in Section I-D, there exists (at least) one optimal policy in the set of deterministic stationary policies. This policy g can be retrieved by finding deterministic policies that induce a fixed point of the operator $T_B^{f\gamma(s,f,v)}$. This condition is modeled as the following set of non-linear constraints:

$$0 \leq v_B(s) - (T_B^{fg} v_B)(s) \leq M_B(1 - g_{sb}) \quad s \in \mathcal{S},\ b \in \mathcal{B}_s \tag{18}$$

$$\sum_{b \in \mathcal{B}_s} g_{sb} = 1 \qquad s \in \mathcal{S} \tag{19}$$

$$g_{sb} \in \{0, 1\} \qquad s \in \mathcal{S}, b \in \mathcal{B}_s \ , \tag{20}$$

where variable g_{sb} represents the deterministic policy $g(s, b)$.

For each $f \in W_A$, the deterministic best response set of the follower is determined by constraints (18)–(20). In (18) the constant M_B is chosen so that when $g_{sb} = 0$, the upper bound is not constraining. Since $||v_B||_\infty \leq ||r_B||_\infty/(1-\beta_B)$, the value $M_B = 2||r_B||_\infty/(1-\beta_B)$ is adequate. Now the leader's problem can be reduced to determine which f maximizes in each state its total expected reward. Vorobeychik and Singh in [16] propose the following formulation:

$$\text{(MP)} \quad \max \quad \sum_{s \in \mathcal{S}} \alpha_s v_A(s) \tag{21}$$

s.t. Constraints (18), (19), (20)

$$\sum_{a \in \mathcal{A}_s} f_{sa} = 1 \qquad s \in \mathcal{S}$$

$$v_A(s) - (T_A^{fg} v_A)(s) \leq M_A(1 - g_{sb}) \quad s \in \mathcal{S},\ b \in \mathcal{B}_s \tag{22}$$

$$f_{sa} \geq 0, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A}_s,$$

where $\alpha \in \mathbb{R}_+^{|\mathcal{S}|}$ is a non-negative vector of coefficients. Problem (MP) above is a non-linear optimization problem with integer variables. Such problems are challenging to solve in general. In particular, constraints (18) and (22) are non-convex quadratic constraints that involve integer variables.

This optimization problem is built based on an analogy to MDPs. In particular, it uses the reduction to a single objective using a vector of weights in (21). For MDPs, this choice is arbitrary. As it turns out in the experiments presented in Section IV, the result of (MP) sometimes depends on the vector α, and sometimes it is not an SSSE. We observe that when such anomalies occur, the operator T_A is not monotone. On the other hand, in cases where T_A is known to be monotone, no such problems have been observed. We therefore conjecture that for the correctness of (MP), it is necessary to have monotonicity of the operator.

III. Existence Results for SSSE and FPE

In this section we present existence results for Stationary Strong Stackelberg Equilibria and Fixed-Point Equilibria for different types of games. We present results for games with a single state, games that admit an MFS, team games, acyclic games and zero-sum games. The general approach is to show that, under assumptions, some operator is contractive, typically T or $\overline{T}_A$ defined in Section II. The fixed point of this operator, from Banach's theorem, is used to construct the FPE. Then, in certain situations, this FPE solution is shown to be an SSSE.

A. Single-state results

In the case where there is only one state, the game is equivalent to a static game ($\beta_A = \beta_B = 0$) or a repeated game ($\beta_A, \beta_B > 0$), so that SSE and SSSE coincide. Indeed, if $S = \{s_0\}$, it is clear that

$$V_i^{fg}(s_0) \;=\; \frac{r_i^{fg}(s_0)}{1-\beta_i}\;, \tag{23}$$

for all $(f,g) \in W_A \times W_B$, so that optimization of V_i^{fg} and of r_i^{fg} are equivalent.

The existence of an SSSE for single-state games is well accepted in the literature with however no clear reference. In this section, we state and prove this result and connect it to the FPE. By doing so, we also link our methodology with static and repeated games. We start with a general and useful result that applies to any game.

Lemma III.1. Let $\mathcal{G}$ be a stochastic game. For all $s \in \mathcal{S}$, the set $R_A(s)$ is nonempty.

Proof: We use the scheme of proof of Proposition 3.1 in [12]. Fix a state $s \in \mathcal{S}$. Define $D = \{(f,g) \in W_A \times W_B | g \in R_B(f)\}$. The value functions V_i^{fg} can be expressed as $V_i^{fg} = (I - \beta_i Q^{fg})^{-1} r_i^{fg}$. Due to the finiteness of $\mathcal{S}$, this is a rational function of f and g. It does not have singularities inside $W_A \times W_B$ and is therefore continuous. In particular, the mappings $(f,g) \mapsto V_i^{fg}(s)$ are continuous. Since W_A and W_B are compact, the maximum theorem applies: the maximum of this function over D exists. Therefore $R_A(s)$ is nonempty. ■

Theorem 4. If the game $\mathcal{G}$ has only one state, then it has an SSSE which is also an FPE.

Proof: Let $\mathcal{S}$ have a single state: $\mathcal{S} = \{s_0\}$. The existence of SSSE is a particular case of Lemma III.1. The existence of an FPE follows from the observation that the game is with MFS: Theorem 1 applies to it. Being a contraction, the operator $\overline{T}_A$ has a unique fixed point from which an FPE is constructed. There remains to show that this FPE coincides with the SSSE.

To that end, we first show that $R_B(f) = \overline{R}_B(s_0, f)$ for all $f \in W_A$ (since the game has MFS, this latter set does not depend on $v_B \in \mathcal{F}(\mathcal{S})$). If $g \in \overline{R}_B(s_0, f)$, then for all $h \in W_B$,

$$V_B^{fg}(s_0) \;=\; \frac{r_B^{fg}(s_0)}{1-\beta_B} \;\geq\; \frac{r_B^{fh}(s_0)}{1-\beta_B} \;=\; V_B^{fh}(s_0)$$

which means that $g \in R_B(f)$. By the same token, if $g \in R_B(f)$ then $g \in \overline{R}_B(s_0, f)$. So both reaction sets coincide. Since V_A^{fg} and r_A^{fg} are also proportional, breaking ties in favor of the leader is the same problem for SSSE and FPE: the sets $SR_B(f)$ and $SR_B(s_0, f, v_B)$ also coincide, and $\gamma_B(f) = \gamma_B(s_0, f, v)$ for all $f \in W_A$ and $v \in \mathcal{F}(\mathcal{S})$. It follows from (5) and (10) that $R_A(s_0) = R_A(s_0, v)$ for all v, which means that SSSE and FPE coincide. ■

B. Myopic Follower Strategies

Theorem 5 (FPE for MFS)**.** If the game $\mathcal{G}$ has MFS then it admits an FPE.

Proof: According to Theorem 1 b), the operator $\overline{T}_A$ is contractive. It therefore admits a fixed point v_A^*. Let $f^* \in W_A$ be defined by, for each $s \in \mathcal{S}$, $f^*(s) = \overline{R}_A(s, v_A^*)$. Let $g^* \in W_B$ be defined for each $s \in \mathcal{S}$, by $g^*(s) = \overline{\gamma}_B(s, f^*, v_A^*) = \gamma_B(s, f^*, v^*)$. We show that (f^*, g^*) is an FPE.

To avoid confusion in the notation, denote with $U = V^{f^* g^*}$, the unique fixed point of $T^{f^* g^*}$. We first check that $v_A^* = U_A$. We have successively: for every $s \in \mathcal{S}$,

$$\begin{aligned} v_A^*(s) &= (\overline{T}_A v_A^*)(s) \\ &= (T_A^{R_A(s,v_A^*), \overline{\gamma}_B(s, R_A(s,v_A), v_A^*)} v_A^*)(s) \\ &= (T_A^{f^*(s) g^*(s)} v_A^*)(s)\;. \end{aligned}$$

The first line is the definition of v_A^* as a fixed point. The second one is the definition of operator $\overline{T}_A$ in (14) and that of T in (11), combined with the MFS property, see the proof of Lemma II.2. The third one is by definition of f^* and g^*. This last line is equivalent to saying that v_A^* is the fixed-point of operator $T_A^{f^* g^*}$, hence $v_A^* = U_A$. As a consequence, $(TU)_A = U_A$.

There remains to be seen that $(TU)_B = U_B$. We have: $(TU)_B = T_B^{f^* g^*} U_B = U_B$ since by definition of U, U_B is the fixed point of $T^{f^* g^*}$. This completes the proof. ■

In the following Lemma III.2, we show that there are actually two main classes games which have MFS. We introduce now these classes of games with one important subclass.

Myopic follower: We define a game as a myopic follower game if $\beta_B = 0$. Note that in this case the follower at any step of the game does not take into account the future rewards, but only the instantaneous rewards.

In this case, the one-step operator of the follower is: $(T_B^{fg} v_B)(s) = r_B^{fg}(s)$ (see (1)) and it clearly does not depend on v_B. Therefore, the reaction set $R_B(s, f, v_B)$ defined

in (3) does not depend either on v_B: $R_B(s,f,v_B) = \overline{R}_B(s,f)$. It follows that the follower's best response has the form (13).

Leader-Controller Discounted Games: This case is a particular case of the Single-controller discounted game described in Filar and Vrieze [18], where the controller is the leader. In other words, the transition law has the form $Q^{ab}(z|s) = Q^a(z|s)$.

In that case, the one-step operator of the follower is:

$$(T_B^{fg} v_B)(s) = r_B^{fg}(s) + \beta_B \sum_{a \in \mathcal{A}_s} f(s,a) \sum_{z \in \mathcal{S}} Q^a(z|s) v_B(z) \ .$$

Then, for $g, h \in W_B$, we have:

$(T_B^{fg} v_B)(s) - (T_B^{fh} v_B)(s) = r_B^{fg}(s) - r_B^{fh}(s)$ and the difference does not depend on v_B. The reaction set $R_B(s,f,v_B)$ is defined as those g such that: $\forall h \in \mathcal{B}_s$, $(T_B^{fg} v_B)(s) - (T_B^{fh} v_B)(s) \geq 0$. It is therefore independent from v_B for any $s \in \mathcal{S}$, and as before, (13) holds.

Multi-stage games: in such games, the state evolves sequentially and deterministically through s_1, $s_2, \ldots, s_K$ and stops. This can be seen a particular case of Leader-Controlled Discounted Game, where the evolution is actually not controlled at all. An additional terminal state with trivial reward functions may be needed to model the end of a game with finitely many stages.

The reduction of MFS to these classes is the topic of the following lemma.

Lemma III.2. Let $\mathcal{G}$ be a game with MFS. Then one of the following statements is true:

i/ $\beta_B = 0$;
ii/ $Q^{ab}(z|s) = Q^a(z|s)$ for all $s, z \in \mathcal{S}$ and all $a \in \mathcal{A}_s$, $b \in \mathcal{B}_s$.

Proof: We prove by contradiction the following statement:

$$\forall s, z, a, b, b', \quad \beta_B \left(Q^{ab}(z|s) - Q^{ab'}(z|s)\right) = 0 \ , \qquad (24)$$

which itself is equivalent to the statement of the lemma.

For each $a \in \mathcal{A}_s$ and $s \in S$, consider the policy where the leader plays the pure strategy a, denoted by δ_a. Take $b^* \in R_B(s, \delta_a, v_B)$ for a given v_B (note that $R_B(s,f,v_B) \neq \emptyset$). Then it is true that for all $b \in \mathcal{B}_s$:

$$r_B^{ab^*}(s) - r_B^{ab}(s) + \sum_{z \in S} \beta_B \left(Q^{ab^*}(z|s) - Q^{ab}(z|s)\right) v_B(z) \geq 0.$$

Suppose by contradiction that (24) does not hold. Then there exists s, a, b and b' such that $\xi = \beta_B(Q^{ab^*}(z^*|s) - Q^{ab'}(z^*|s)) \neq 0$ for some z^*. Then by taking $v'_B(z^*)$ with the opposite sign of ξ big enough, and $v'_B(z) = 0$, for $z \neq z^*$, the inequality will turn negative. That would mean that b^* does not belong to $R_B(s,f,v'_B)$ with v'_B and then the game is not MFS. This is a contradiction, so such elements s, a, b, b', z do not exist, and (24) holds. ∎

We now state the principal results of this section: the MFS property implies the existence of both FPE and SSSE, and their coincidence.

Theorem 6. Let $\mathcal{G}$ be a stochastic game with MFS. Then $\mathcal{G}$ has an SSSE, which corresponds to its FPE.

Proof: Let (f^*, g^*) be the FPE of game $\mathcal{G}$, and $V^* = V^{f^* g^*}$. We know that the FPE exists by Theorem 5. From the proof of this result, we know that $V_A^* = \overline{T}_A V_A^* = \overline{T}_A^{f^*} V_A^*$.

We first prove that $R_B(f^*) = \prod_{s \in \mathcal{S}} \overline{R}_B(s, f^*)$. According to Lemma III.2, since the game has MFS, then either $\beta_B = 0$, or the game is Leader-Controlled Discounted. In both cases, the value of Player B has the form (see (12)):

$$V_B^{fg} \ = \ (I - \beta_B Q^f)^{-1} r_B^{fg} \ ,$$

where Q^f is the leader-controlled transition matrix, relevant only in case $\beta_B \neq 0$. We note that, given that the matrix $(I - \beta_i Q^f)^{-1}$ is positive, $r_B^{fg} \geq r_B^{fh}$, implies $V_B^{fg} \geq V_B^{fh}$. Additionally, if for some s, g, h, $r_B^{fg}(s) > r_B^{fh}(s)$, then $V_B^{fg}(s) > V_B^{fh}(s)$.

Let f be an arbitrary element of W_A. On the one hand, $\prod_s \overline{R}_B(s,f) \subset R_B(f)$. To see this, pick $g \in \prod_s \overline{R}_B(s,f)$. Then for all $h \in W_B$, $r_B^{fg} \geq r_B^{fh}$ and therefore $V_B^{fg} \geq V_B^{fh}$: this means $g \in R_B(f)$. The set $R_B(f)$ is therefore nonempty. On the other hand, $R_B(f) \subset \prod_s \overline{R}_B(s,f)$. To see this, pick $g \in R_B(f)$ (the set is not empty). If it is not in $\prod_s \overline{R}_B(s,f)$, then there is some s and some $b \in \overline{R}_B(s,f)$ such that $r_B^{fb}(s) > r_B^{fg}(s)$. Then the policy $h \in W_B$ which coincides with g except at state s where $h(s) = b$, is such that $V_B^{fh}(s) > V_B^{fg}(s)$, a contradiction. Therefore, $\prod_s \overline{R}_B(s,f) = R_B(f)$, for all $f \in W_A$.

At this point, we have shown that Player B reacts the same way to Player A's strategy f, in the SSSE problem or in the FPE problem with any scrap value function v. Nevertheless, we cannot conclude that the *strong* reaction is the same, since that of the FPE problem *does* depend on the scrap value v.

However, we know that Player B's tie-breaking problem in (4) is a MDP. This means that the value of Player A after Player B's strong reaction, say V_A^f, is given by a Bellman equation, namely:

$$\begin{aligned} V_A^f(s) &= \max_{g \in R_B(f)} \{r_A^{fg}(s) + \beta_A (Q^{fg} V_A^f)(s)\} \\ &= \max_{b \in \overline{R}_B(s,f)} \{r_A^{fb}(s) + \beta_A (Q^{fb} V_A^f)(s)\} \end{aligned} \qquad (25)$$

for all $s \in \mathcal{S}$. Here, we have used the fact that $R_B(f)$ is a cartesian product, and that MDPs can be solved state by state. We recognize in the right-hand side of (25) the operator $\overline{T}_A^f$ defined in (16). In other words, V_A^f is the fixed point of $\overline{T}_A^f$.

Lets define $U_A \in \mathcal{F}(\mathcal{S})$ as: $U_A(s) = \max_{f \in W_A} V_A^f(s) = V_A^{f_s}(s)$. Here, $f_s \in W_A$ realizes the maximum for state s. By construction, $U_A(s) \geq V_A^f(s)$ for any particular $f \in W_A$. We proceed to prove that $U_A = V_A^*$. First, consider the action of $\overline{T}_A$ on U_A: for $s \in \mathcal{S}$,

$$\begin{aligned} (\overline{T}_A U_A)(s) = \max_{f \in W_A} (\overline{T}_A^f U_A)(s) \ \geq \ (\overline{T}_A^{f_s} U_A)(s) \ \geq \\ (\overline{T}_A^{f_s} V_A^{f_s})(s) \ = \ V_A^{f_s}(s) \ = \ U_A(s) \ . \end{aligned}$$

The first equality is the definition of $\overline{T}_A$. The first inequality is clear. The second one results from the monotonicity of operator $\overline{T}_A^f$. The second equality is because V^{f_s} is the fixed point of $\overline{T}_A^{f_s}$. Then according to Proposition 1, $\overline{T}_A U_A \geq U_A$ implies $U_A \leq V_A^*$ since V_A^* is the fixed point of $\overline{T}_A$.

Now, since $V_A^* = V_A^{f^*}$, the fixed point of operator $\overline{T}_A^{f^*}$, then for all $s \in \mathcal{S}$, $U_A(s) = \max_f V_A^f(s) \geq V_A^{f^*}(s) = V_A^*(s)$. In other words, $U_A \geq V_A^*$. We conclude that indeed $U_A = V_A^*$.

As a consequence, we have shown that $f^* \in \cap_{s\in\mathcal{S}} R_A(s)$, that is, (f^*, g^*) is an SSSE. ∎

C. Team Games

Team Games (or Identical-Goal Games [12]) are such that both players seek to maximize the same metric. This is a property of reward functions only. Proposition 1 in [16] claims that Team Games have an SSSE. However, that result only shows the optimality of stationary policies, by reducing the game to an MDP, not the existence of an SSSE. Here we use this related MDP to show the existence of SSSE and FPE and their coincidence. We begin generalizing the definition of Team Games.

Definition III.1 (Team Game)**.** The game is a Team Game if $\beta_A = \beta_B$ and there exists real constants μ and $\nu > 0$ such that: $r_B^{ab}(s) = \mu + \nu r_A^{ab}(s)$.

We continue with the construction of a MDP from the team game. Its parameters are $(\tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \tilde{Q}, \tilde{r}, \beta)$. Here, $\tilde{\mathcal{S}} = \mathcal{S}$, $\tilde{\mathcal{A}} = \mathcal{A} \times \mathcal{B}$, $\beta = \beta_A = \beta_B$, and for $\tilde{a} = (a, b)$ define $\tilde{Q}^{\tilde{a}}(s'|s) = Q^{ab}(s'|s)$ and $\tilde{r}^{\tilde{a}}(s) = r_A^{ab}(s)$. Define a stationary policy for this MDP by $h(s, \tilde{a}) = h(s, (a, b))$ as the probability of selecting strategy $\tilde{a}$ in state s. Let $\widetilde{T}^h$ be the linear operator on $\mathcal{F}(\mathcal{S})$ for this MDP corresponding to a stationary policy h, and $\widetilde{V}^h$ be the corresponding value function (fixed point of $\widetilde{T}^h$). We also consider the dynamic programming operator $\widetilde{T} : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$ defined as follows:

$$\widetilde{T}v(s) = \max_{h\in\mathbb{P}(\tilde{A})} \sum_{\tilde{a}\in\tilde{A}_s} h(s,\tilde{a}) \left[\tilde{r}^{\tilde{a}}(s) + \beta \sum_{z\in S} Q^{\tilde{a}}(z|s)v(z)\right] .$$

From MDP theory we know that there exists a set of deterministic optimal stationary policies for the above MDP, which we denote $\mathcal{H}$. For $h \in \mathcal{H}$, $s \in \tilde{\mathcal{S}}$ and $\tilde{a} = (a, b) \in \tilde{\mathcal{A}}$ define $f^h \in W_A$ and $g^h \in W_B$ as:

$$f^h(s,a) = \begin{cases} 1 & \text{if } h(s,(a,b)) = 1 \text{ for some } b \\ 0 & \text{otherwise.} \end{cases}$$
$$g^h(s,b) = \sum_{a\in\mathcal{A}_s} f^h(s,a)h(s,(a,b)). \tag{26}$$

Given that $h \in \mathcal{H}$ is deterministic the pair (f^h, g^h) is well defined and $h(s,(a,b)) = f^h(s,a)g^h(s,b)$. From the optimality of $\mathcal{H}$, we have that $\widetilde{V}^h = \widetilde{V}^*$ for each $h \in \mathcal{H}$. Let $h^* \in \mathcal{H}$ denote the strategy with a maximal g^h:

$$h^* = \arg\max_{\prec_B} \{g^h : h \in \mathcal{H}\} . \tag{27}$$

We prove next that $(f^*, g^*) = (f^{h^*}, g^{h^*})$ is the FPE and SSSE, showing that if $\mathcal{G}$ is a Team Game, then it admits an FPE and an SSSE and they coincide.

Theorem 7. The pair (f^*, g^*) defined in (26) and (27) forms an SSSE and an FPE with value $v_A^* = \widetilde{V}^*$ for the leader and $v_B^* = \frac{\mu}{1-\beta} + \nu\widetilde{V}^*$ for the follower (unique fixed points of $T_i^{f^*g^*}$, $i = A, B$).

Proof: We start proving that (f^*, g^*) is an FPE. We show that $\widetilde{V}^*$ is the fixed point of $T_A^{f^*g^*}$ and that (f^*, g^*) satisfies conditions i/ and ii/ of Definition II.2. For any $s \in \mathcal{S}$,

$$\begin{aligned}
\widetilde{V}^*(s) &= \widetilde{T}\widetilde{V}^*(s) = \widetilde{T}^{h^*}\widetilde{V}^*(s) \\
&= \sum_{\tilde{a}\in\tilde{A}_s} h^*(s,\tilde{a})[\tilde{r}^{\tilde{a}}(s) + \beta \sum_{z\in S} \tilde{Q}^{\tilde{a}}(z|s)\widetilde{V}^*(z)] \\
&= \sum_{(a,b)\in\tilde{A}_s} f^*(s,a)g^*(s,b)[r_A^{ab}(s) + \beta \sum_{z\in S} Q^{ab}(z|s)\widetilde{V}^*(z)] \\
&= T_A^{f^*g^*}\widetilde{V}^*(s) .
\end{aligned}$$

Here we used that $\widetilde{V}^*$ is fixed point of $\widetilde{T}$, the definition of h^* and $T_A^{f^*g^*}$.

Consider now the operator T_B^{fg} applied to v_B^*, this function being defined in the statement of the theorem. From the definition of Team Games we have that, for all f, g,

$$\begin{aligned}
&(T_B^{fg}v_B^*)(s) \\
&= \sum_a \sum_b f(s,a)g(s,b) \left[\mu + \nu r_A^{ab}(s) + \beta \sum_z Q^{ab}(z|s)v_B^*\right] \\
&= \mu + \sum_a \sum_b fg \left[\nu r_A^{ab}(s) + \frac{\beta\mu}{1-\beta} + \beta\nu \sum_z Q^{ab}(z|s)v_A^*\right] \\
&= \mu + \frac{\beta\mu}{1-\beta} + \nu \left[r_A^{fg}(s) + \beta \sum_z Q^{fg}(z|s)v_A^*\right] \\
&= \frac{\mu}{1-\beta} + \nu(T_A^{fg}v_A^*)(s).
\end{aligned} \tag{28}$$

Then for each $g \in W_B$ and $s \in \mathcal{S}$:

$$\begin{aligned}
T_B^{f^*g^*}v_B^*(s) - T_B^{f^*g}v_B^*(s) =& \nu(T_A^{f^*g^*}v_A^*(s) - T_A^{f^*g}v_A^*(s)) \\
=& \nu(\widetilde{T}^{h^*}v_A^*(s) - \widetilde{T}^h v_A^*(s)) \geq 0,
\end{aligned}$$

where $h = f^*g$ is the policy induced by f^* and g. Therefore $g^* \in R_B(s, f, v_B^*)$. Since the strategy that maximizes the reward for Player B also optimizes the reward of Player A, we have $R_B(s, f^*, v_B^*) = SR_B(s, f^*, v^*)$. Finally, due to (27), $g^* = \gamma_B(s, f, v^*)$. With an analogous argument, we get that $f^* = R_A(s, v^*)$, proving that (f^*, g^*) is an FPE.

Now we show that (f^*, g^*) is an SSSE, that is (f^*, g^*) satisfies conditions i/ and ii/ of Definition I.1. For every $g \in W_B$ and $s \in \mathcal{S}$,

$$\begin{aligned}
\widetilde{V}^*(s) =& \widetilde{V}^{h^*}(s) = V_A^{f^*g^*}(s) = \frac{\mu}{1-\beta} + \nu V_B^{f^*g^*}(s) \\
\geq& \widetilde{V}^{f^*g}(s) = V_A^{f^*g}(s) = \frac{\mu}{1-\beta} + \nu V_B^{f^*g}(s) .
\end{aligned}$$

The inequality holds because policy f^*g may not belong to $\mathcal{H}$. Any best response to f^* must attain the value $\widetilde{V}^*(s)$. This shows that $g^* \in R_B(f^*) \subseteq \mathcal{H}$. The above derivation also implies

$$V_A^{f^*g^*}(s) - V_A^{f^*g}(s) = \nu(V_B^{f^*g^*}(s) - V_B^{f^*g}(s)).$$

So if g^* maximizes Player B's gain, it also maximizes Player A's gain, implying that $R_B(f^*) = SR_B(f^*)$. Finally, by (27)

$$g^* = \max_{\prec_B} \{g^h : h \in \mathcal{H}\} = \max_{\prec_B} SR_B(f^*) = \gamma_B(f^*) ,$$

proving condition i/. To prove condition ii/, assume by contradiction that $f^* \notin R_A(s')$ for some s'. This means there exists some $f' \in W_A$ such that $V_A^{f'\gamma_B(f')}(s') > V_A^{f^*\gamma_B(f^*)}(s') = V_A^{f^*g^*}(s') = \widetilde{V}^*(s')$. We have constructed a strategy h' with larger value at state s', contradicting the fact that h^* optimizes the MDP. This proves that (f^*, g^*) is an SSSE. ∎

We conclude this discussion of team games by highlighting the intuition behind the main result. Since both players have basically the same goal, they act as a single player and it is expected that games with identical goals would behave as MDPs, inheriting all the good properties: the existence of optimal stationary policies and the convergence of value iteration and policy iteration algorithms.

D. Acyclic Games

Acyclic games are such that state-to-state transitions do not lead back to a visited state, except for absorbing states. Acyclicity is a property of transition operators only.

We say that state s' is reachable from state s if there exist $k \in \mathbb{N}$, a sequence of states $s = s_0, s_1, \ldots, s_k = s'$ and actions $a_0, \ldots, a_{k-1}, b_0, \ldots, b_{k-1}$ with $Q^{a_0 b_0}(s_1|s_0) \times Q^{a_1 b_1}(s_2|s_1) \times \ldots \times Q^{a_{k-1} b_{k-1}}(s_k|s_{k-1}) > 0$.

Definition III.2 (Acyclic Games)**.** The game is an Acyclic Game if the state space $\mathcal{S}$ admits the partition $\mathcal{S} = \mathcal{S}_\perp \cup \mathcal{S}_\oplus$, with:

- for all $s \in \mathcal{S}_\perp$, $a \in \mathcal{A}_s$, $b \in \mathcal{B}_s$, $Q^{ab}(s|s) = 1$;
- for every pair $(s, s') \in \mathcal{S}_\oplus \times \mathcal{S}_\oplus$, if s' is reachable from s, then s is not reachable from s'.

The following theorem is based on Theorem 4 and generalizes it for the FPE part.

Theorem 8. If the stochastic game $\mathcal{G}$ is an Acyclic Game, then it admits an FPE.

Proof: The proof will proceed by successive reductions to static (or single-state) games.

The game being acyclic, it is possible to perform a topological sort of the state space. There exists a partition $\mathcal{S} = \cup_{k=0}^K \mathcal{S}_k$ with $\mathcal{S}_0 = \mathcal{S}_\perp$ and for every $s \in \mathcal{S}_k$, $k > 0$, if s' is reachable from s then $s' \in \mathcal{S}_{k'}$ with $k' < k$. In a first step, we construct a candidate strategy (f^*, g^*). Then we prove that this strategy solves the FPE problem.

For each $s_0 \in \mathcal{S}_0 = \mathcal{S}_\perp$, consider G_0, the single-state game with $S = \{s_0\}$ and same strategies, rewards and discount factors. Theorem 4 applies to this game. It states that an FPE exists, resulting in a pair of strategies $(f^*_{s_0}, g^*_{s_0}) \in \mathbb{P}(\mathcal{A}_{s_0}) \times \mathbb{P}(\mathcal{B}_{s_0})$ and a value $V^*(s_0)$.

We now construct the strategies $(f^*_{s_k}, g^*_{s_k})$ for $s_k \in \mathcal{S}_k$ with a recurrence on k. Assume this has been done up to $k-1$. Pick $s_k \in \mathcal{S}_k$. Because the game is acyclic, we have, for any $(f, g) \in W_A \times W_B$,

$$\begin{aligned} V_i^{fg}(s_k) &= r_i^{fg}(s_k) + \beta_i \sum_{s' \in \mathcal{S}} Q^{fg}(s'|s_k) V_i^*(s') \\ &= r_i^{fg}(s_k) + \beta_i \sum_{\ell=0}^{K} \sum_{s' \in \mathcal{S}_\ell} Q^{fg}(s'|s_k) V_i^*(s') \\ &= r_i^{fg}(s_k) + \beta_i \sum_{\ell=0}^{k-1} \sum_{s' \in \mathcal{S}_\ell} Q^{fg}(s'|s_k) V_i^*(s') \end{aligned} \quad (29)$$

since, by construction of the topological sort, $Q^{fg}(s'|s_k) = 0$ for all $s' \in \mathcal{S}_\ell$ when $\ell \geq k$, and for any (f, g).

Consider the static game (i.e. one-state game with null discount factors) with $S = \{s_k\}$ and rewards defined with this formula. Again, Theorem 4 applies to this game: an FPE exist, resulting in a pair of strategies $(f^*_{s_k}, g^*_{s_k}) \in \mathbb{P}(\mathcal{A}_{s_k}) \times \mathbb{P}(\mathcal{B}_{s_k})$ and values $V_i^*(s_k)$. When $k = K$, we have defined this way a strategy (f^*_s, g^*_s) for each $s \in S$.

We now claim that this strategy is an FPE. We prove this with a recurrence. More precisely, we prove that for all k, property P_k holds, which says that for and all $s_k \in \mathcal{S}_k$:

$$\begin{aligned} g^*(s_k) &= \gamma_B(s_k, f^*, V^*) \\ f^*(s_k) &= R_A(s_k, V^*) \end{aligned}$$

where $V_i^* = V_i^{f^*g^*}$ is the unique fixed point of operator $T_i^{f^*g^*}$ for $i = A, B$. With Definition II.2, the result will follow.

When $s_0 \in \mathcal{S}_0$, the local reaction set $R_B(s_0, f, V^*)$ does not depend on V^* and $\{g^*_{s_0}\} \in R_B(s_0, f^*, V^*)$, as in the proof of Theorem 4. In particular, it follows that $g^*(s_0) = \gamma_B(s_0, f^*, V^*)$ and $f^* = R_A(s_0, V^*)$. So P_0 holds.

Assume now that property P_ℓ holds for all $\ell < k$. Let $s_k \in \mathcal{S}_k$. Then

$$(T_B^{f^*g} V_B^*)(s_k) = r_i^{f^*g}(s_k) + \beta_i \sum_{\ell=0}^{k-1} \sum_{s' \in \mathcal{S}_\ell} Q^{f^*g}(s'|s) V_B^*(s'),$$

to be compared with (29). Then since $(f^*_{s_k}, g^*_{s_k})$ solves (locally) the SSE for the subgame defined by (29), then $g^*_{s_k} = \gamma_B(s_k, f^*, V^*)$ and $f^*_{s_k} \in R_A(s_k, V^*)$ by construction. So property P_k holds. By recurrence, P_K holds and (f^*, g^*) is an FPE. ∎

In contrast with Theorem 4, the existence of SSSE is not guaranteed for acyclic games. In Section IV-B we study a game without an SSSE. As we explain at the end of this section, this game is not acyclic, but it is possible to "approximate" it with an acyclic game which will have the same qualitative properties. On the other hand, if the transitions of a game are *deterministic*, in other words if the game is a multi-stage game, then it is with MFS and it does have an SSSE according to Theorem 6.

E. Zero-Sum Games

In zero-sum games, $\beta_A = \beta_B$ and $r_B = -r_A$.

Theorem 9. If the game $\mathcal{G}$ is a Zero-Sum Game, then it admits an FPE.

The existence of an FPE follows from the contractivity of the operator associated, in a similar way as in [14, Section 8] for Nash Equilibria in Stochastic Games. We include here an argument in the line of the proof of Theorem 1.

Proof: Consider a function v in the set $\mathcal{W} = \{v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S}) | v_B = -v_A\}$. Since v_B can be substituted with $-v_A$, it turns out that $SR_B(s, f, v) = R_B(s, f, v_B) = R_B(s, f, -v_A)$ and $\gamma_B(s, f, v)$ can be made dependent only on v_A; in other

words, it satisfies (13). It is then possible to define the operator $\overline{T}_A^f$ as in (15). This operator maps $\mathcal{W}$ to $\mathcal{W}$.

On the other hand, $(\overline{T}_A^f v_A)(s) = (T_A^{f,\gamma(s,f,v_A)} v_A)(s) = (T_A^{f,g} v_A)(s)$ for all $g \in R_B(s,f,v_A)$. But for every f, g, $T_A^{fg} v_A = -T_B^{fg} v_B$. And by definition (6), for all $g \in R_B(s,f,v_A)$, $h \in W_B$, $(T_B^{fg} v_B)(s) \geq (T_B^{fh} v_B)(s)$, which is equivalent to $(T_A^{fg} v_A)(s) \leq (T_A^{fh} v_A)(s)$. In other words,

$$R_B(s,f,v_A) = \arg\min_{g \in W_B} (T_A^{fg} v_A)(s)$$

so that

$$(\overline{T}_A^f v_A)(s) = \min_{g \in W_B} (T_A^{fg} v_A)(s) \ .$$

As it was the case in (16), the minimization set in the right-hand side does not depend on v_A. The proof of Theorem 1 then applies mutatis mutandis, to conclude that operator $\overline{T}_A$ is contractive on $\mathcal{W}$. It then admits a fixed point v_A^* in that set. Then the argument in the proof of Theorem 5 applies, and the FPE of the game is constructed from this fixed point. ■

In one-shot games, the Stackelberg equilibrium and the Nash equilibrium for zero-sum games coincide [12]. In addition, for stochastic zero-sum games, the Nash equilibrium is the fixed point of the dynamic programming operator. We have tested computationally the relationship between the output of (MP), value iteration and policy iteration in zero-sum games. In all instances tested, the output of MP and the FPE coincide. We conjecture that for zero-sum games, FPE and SSSE coincide. The formal proof of this relationship is left for future study.

IV. Numerical Examples

In this section, we present examples illustrating different situations that can occur by comparing the solutions returned by the different algorithms presented in Section II.

The example of Section IV-A does not satisfy the sufficient conditions for existence and convergence identified in Sections II-C and III, however (VI) converges to some FPE which is an SSSE and the solution returned by (MP). In Section IV-B, we describe an example where, depending on the discount factors β_A, β_B: either an SSSE exists and does not coincide with the FPE, or an SSSE does not exist, or an FPE does not exist. Finally, in Section IV-C, we describe an example where an FPE is shown to exist, but (VI) does not necessarily converge to it.

The numerical experiments involving value iteration and policy iteration were performed using Python 3.6, on a machine running under MacOS, a processor of 2,6 GHz Intel Core i5, and memory of 8 GB 1600 MHz DDR3. Operator T is implemented with Cplex 12.8. To solve the linear systems involved in T^{fg} we use the Python library Numpy. In order to solve (MP) we use the KNITRO 12.0 solver combined with AMPL.

The data of the exemples will be displayed according to the following figure that shows the relevant parameters when the system is in state s, Player A performs action a and Player B performs action b.

s	b
a	$Q^{ab}(s_1\|s)$, $Q^{ab}(s_2\|s)$ $r_A^{ab}(s), r_B^{ab}(s)$

A. Example 1: FPE and SSSE coincide and (VI) converges

This first example shows the convergence of value iteration in a case where an FPE exists. Consider $\beta_A = \beta_B = \frac{9}{10}$ and the data in Table I.

s_1	b_1	b_2
a_1	$\frac{1}{2}, \frac{1}{2}$ $10, -10$	$0, 1$ $-5, 6$
a_2	$\frac{1}{4}, \frac{3}{4}$ $-8, 4$	$1, 0$ $6, -4$

s_2	b_1	b_2
a_1	$\frac{1}{2}, \frac{1}{2}$ $7, -5$	$0, 1$ $-1, 6$
a_2	$\frac{1}{4}, \frac{3}{4}$ $-3, 10$	$1, 0$ $2, -10$

TABLE I
TRANSITION MATRIX AND PAYOFFS FOR EACH PLAYER.

This example satisfies none of the sufficient conditions listed in Sections II and III. However, both value iteration and policy iteration converge to the FPE, and furthermore the FPE, the SSSE and the optimal solution returned by (MP) coincide. The policies and values are given in Table II. Detailed algebraic manipulations of the value functions given the data in this example allows us to show that the SSSE satisfies the values in Table II. Then, we show that this solution is a fixed point for operator T. Details are provided in Appendix C.1 in [21].

	s_1	s_2
Play of A	(0.3467, 0.6533)	(0.6434, 0.3566)
Play of B	b_1	b_2
v_A	26.841	28.437
v_B	-1.807	-0.679

TABLE II
POLICIES AND VALUES OF THE SSSE IN EXAMPLE 1

B. Example 2: FPE and SSSE are different

In this section we study the stochastic game given by the data in Table III. This game has two states and two actions per state. Figure 1 shows a diagram of the possible transition between states. State s_1 is absorbing for any combination of actions.

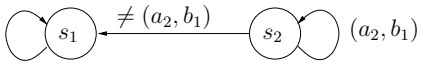

Fig. 1. Transition structure of Example 2

Whenever $\beta_B > 0$, depending on the values of β_A the existence and non-existence for both equilibrium concepts, FPE and SSSE, changes. For $\beta_A < \frac{1}{5}$ the existence of an FPE is guaranteed, but no SSSE exist. This comes from the fact that for these values of β_A, $R_A(s_1) = \{a_1\} \times \{(f_2, 1 - f_2) : f_2 \in [0,1]\}$ and $R_A(s_2) = \{a_2\} \times \{a_2\}$. Clearly,

s_1	b_1	b_2
a_1	$1,0$ / $-1,-2$	$1,0$ / $-2,1$
a_2	$1,0$ / $0,0$	$1,0$ / $2,0$

s_2	b_1	b_2
a_1	$1,0$ / $-1,-2$	$1,0$ / $-2,-2$
a_2	$0,1$ / $0,1$	$1,0$ / $1,1$

TABLE III
TRANSITION MATRIX AND PAYOFFS FOR EACH PLAYER IN EXAMPLE 2

$R_A(s_1) \cap R_A(s_2) = \emptyset$ and therefore there is no SSSE. When the game starts in the absorbing state s_1, the optimal policy for the leader is to play and announce the (static) SSE in s_1 and an arbitrary strategy in s_2. When the game starts in s_2, the leader has the incentive to announce a sub-optimal strategy in s_1 in order to remain in state s_2 and increase her expected reward. This could be done only for low values of β_A.

On the other hand, whenever $\beta_A > \frac{1}{3}$, there is no FPE. In the interval $\beta_A \in [\frac{1}{5}, \frac{1}{3}]$, both SSSE and FPE exists, but the strategies and values are different. This information is summarized in Table IV. The details of this analysis are provided in Appendix C.2 in [21].

Finally, we test the Mathematical Programming (MP) formulation and policy iteration (PI) algorithm, for the different values of the parameter α in (MP) and for values of β_A in $\{0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$. Table V summarizes the results obtained.

In this experiment whenever the SSSE exists ($\beta_A \in \{\frac{1}{4}, \frac{1}{2}\}$), (MP) computes it correctly and policy iteration converges to the FPE. When no SSSE exists ($\beta_A \in \{0, \frac{1}{8}\}$), (MP) returns a value that is influenced by the vector of weights α_s. In these cases policy iteration converges to an FPE that is different from the solution found by (MP).

The example we have presented is not acyclic in the sense of Definition III.2. However, it is possible to construct from it acyclic games that reproduce the absence of SSSE. It suffices to duplicate state s_2 in a sequence $s_2^{(0)}$, $s_2^{(1)}$, ..., such that transitions $s_2 \to s_2$ in the original game become transitions $s_2^{(i)} \to s_2^{(i+1)}$ while other transitions lead to absorbing state s_1. This transition structure is acyclic. Because of the discount factor β_A, the value of state s_2 for Player A in the original game can be approximated at any precision in the new game, with a suitable number of replications. Her behavior will then be the same: commit to a suboptimal strategy in s_1 in order to induce Player B to have the game stay in the sequence of "s_2" states, thereby leading to a lack of SSSE.

We conclude this section commenting about the Example 1 in [16]. Authors claim that this counterexample is an example of non-existence of stationary equilibrium, but it only shows the sub-optimality of stationary policies. Example 1 in [16] is actually a multi-stage game and it is easy to see that any stationary strategy for the leader and the trivial strategy for the follower is in fact an SSSE.

C. Example 3: FPE exists but (VI) does not converge to it

We now develop an example where an FPE does exist, but value iteration does not necessarily converge to it. The data of this example is listed in Table VI.

Consider $\beta_A = \beta_B = \frac{1}{2}$. We claim that the pair of strategies (f^*, g^*) and value functions (v_A^*, v_B^*) in Table VII constitute *both* an SSSE and an FPE. Justifications for this claim are provided in Appendix C.3 in [21].

When applying value iteration with the null function as a starting point, we get however the evolution in Figure 2. Values obtained with policy iteration have a similar behavior. Finally, (MP) returns as the optimal solution the SSSE (and FPE).

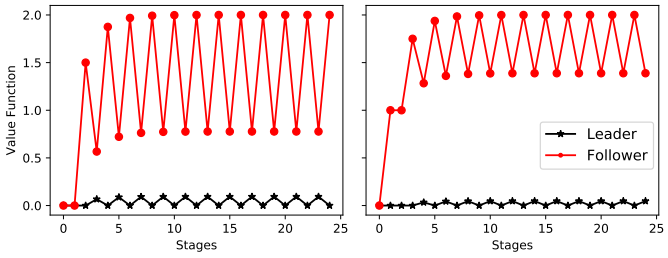

Fig. 2. Value iteration applied to Example 3: state s_1 (left) and s_2 (right)

V. APPLICATION: SURVEILLANCE IN A GRAPH

In this section we present an example of a stochastic game that models the interaction between a security patrol and an attacker. In this game a defender has to patrol (or "cover") a set of locations and an attacker wants to perform an attack in one of these locations, both maximizing their expected rewards.

The states of this game are defined by the locations of the defender and attacker and whether an attack occurred or not. At every iteration the leader decides where to patrol, according to a mixed strategy which is known to the attacker, who then decides two things: where to move and whether to attack or not. Once the attack is performed the game ends in one of two terminal states.

Both player's actions are affected by randomness, players may fail to reach the location they decided to move due to external factors, in which case they remain in their current location.

A. Game description

We introduce now the elements of the model and notation. We consider a graph $(\mathcal{L}, \mathcal{E})$, where the set of nodes represent n locations to patrol/targets $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_n\}$ that may be connected by edges in $\mathcal{E}$. Player A is the defender, Player B is the attacker. The state space is $\mathcal{S} = \mathcal{L} \times \mathcal{L} \times \{0,1\} \cup \{\perp_0, \perp_1\}$. A typical state $s = (\ell_A, (\ell_B, \alpha)) \in \mathcal{S}$ represents the defender's location ($\ell_A \in \mathcal{L}$), the attacker's location ($\ell_B \in \mathcal{L}$), and whether Player B is committing an attack $\alpha = 1$ or not $\alpha = 0$. There are also two special absorbing states $\perp_0$, $\perp_1$ representing the outcome of the attack. State $\perp_1$ represents the case where the attack is successful and $\perp_0$ when the attacker is caught.

	SSSE				FPE			
β_A	$v_A(s_1)$	$v_A(s_2)$	$v_B(s_1)$	$v_B(s_2)$	$v_A(s_1)$	$v_A(s_2)$	$v_B(s_1)$	$v_B(s_2)$
$[0, \frac{1}{5})$	No	No	No	No	$\frac{2}{1-\beta_A}$	0	0	$\frac{1}{1-\beta_B}$
$[\frac{1}{5}, \frac{1}{3}]$	$\frac{2}{1-\beta_A}$	$\frac{2\beta_A}{1-\beta_A}$	0	0	$\frac{2}{1-\beta_A}$	0	0	$\frac{1}{1-\beta_B}$
$(\frac{1}{3}, 1)$	$\frac{2}{1-\beta_A}$	$\frac{2\beta_A}{1-\beta_A}$	0	0	No	No	No	No

TABLE IV
EXISTENCE OF FPE AND SSSE FOR DIFFERENT VALUES OF β_A AND $\beta_B > 0$. EVEN WHEN BOTH EXIST THEY MAY NOT COINCIDE.

		(MP)						(PI)	
		α_{s_1}	α_{s_2}	α_{s_1}	α_{s_2}	α_{s_1}	α_{s_2}		
β_A		100	1	1	100	1	1	s_1	s_2
0	v_A	2	~ 0	-2	1	2	~ 0	2	0
	v_B	~ 0	~ 0	2	2	~ 0	~ 0	0	2
	f	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(1,0)	(0,1)	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	(0,1)
	g	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(1,0)
$\frac{1}{8}$	v_A	16/7	2/7	$-16/7$	5/7	16/7	2/7	16/7	0
	v_B	~ 0	~ 0	2	2	~ 0	~ 0	0	2
	f	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(1,0)	(0,1)	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	(0,1)
	g	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(1,0)
$\frac{1}{4}$	v_A	8/3	2/3	8/3	2/3	8/3	2/3	8/3	0
	v_B	~ 0	~ 0	~ 0	~ 0	~ 0	~ 0	0	2
	f	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	(0, 1)
	g	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(1,0)
$\frac{1}{2}$	v_A	4	2	4	2	4	2	-	-
	v_B	~ 0	~ 0	~ 0	~ 0	~ 0	~ 0	-	-
	f	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	(0,1)	$(\frac{1}{3}, \frac{2}{3})$	-	-
	g	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	-	-

TABLE V
RESULTS FOR (MP) AND (PI) WITH DIFFERENT VALUES OF α AND β_A WITH $\beta_B = 0.5$ FIXED

s_1	b_1	b_2
a_1	1, 0 / $1, -1$	0, 1 / 0, 1
a_2	0, 1 / $-1, 1$	0, 1 / $-1, -1$

s_2	b_1	b_2
a_1	0, 1 / $-1, 0$	1, 0 / 0, 1
a_2	1, 0 / 0, 1	0, 1 / $1, -1$

TABLE VI
TRANSITION MATRIX AND PAYOFFS FOR EACH PLAYER IN EXAMPLE 3

	s_1	s_2
Play of A	$(1, 0)$	$(5 - \sqrt{19}, -4 + \sqrt{19})$
Play of B	b_2	b_2
v_A	$\frac{1}{5}(-3 + \sqrt{19})$	$\frac{1}{5}(-6 + 2\sqrt{19})$
v_B	$\frac{1}{5}(16 - 2\sqrt{19})$	$\frac{1}{5}(22 - 4\sqrt{19})$

TABLE VII
VALUES AND POLICIES FORMING AN SSSE AND FPE.

The action space $\mathcal{A}_s \subset \mathcal{L}$ for the leader represents all possible locations that can be reached from its current position (given by state s). For the follower, $\mathcal{B}_s \subset \mathcal{L} \times \{0, 1\}$ represents all possible locations that can be reached by the attacker from state s and the decision whether to attack or not. Actions are irrelevant for states $s \in \{\perp_0, \perp_1\}$, we denote these action states by $\mathcal{A}_{\perp_i}$ for $i = \{0, 1\}$. We denote the action of "*move to* ℓ*"* by $\ell \in \mathcal{L}$ and the action of "*attack/not attack"* by $\alpha \in \{0, 1\}$.

The probability of transitions between states is constructed using the function $q_i^{\ell'}(\ell''|\ell)$ which denotes the probability that player $i \in \{A, B\}$ reaches location ℓ'' from ℓ having decided to move to ℓ'. We assume that these probabilities are independent between players. In particular, if there is no failure in player's i movements, then $q_i^{\ell'}(\ell''|\ell) = 1$ if $\ell'' = \ell'$, and 0 otherwise. The transition probabilities $Q^{ab}(z|s)$ are then defined as follows:

$$Q^{ab}(z|\perp_i) = \begin{cases} 1 & z = \perp_i, \quad i \in \{0,1\}, (a,b) \in \mathcal{A}_{\perp_i} \times \mathcal{B}_{\perp_i} \\ 0 & \text{otherwise} , \end{cases}$$

$$Q^{\ell'_A,(\ell'_B,\alpha)}(\ell''_A, (\ell''_B, \alpha)|\ell_A, (\ell_B, 0)) = q_A^{\ell'_A}(\ell''_A|\ell_A) q_B^{\ell'_B}(\ell''_B|\ell_B) ,$$

for $\alpha \in \{0, 1\}$, $\ell'_A \in \mathcal{A}_{\ell_A}$ and $(\ell'_B, \alpha) \in \mathcal{B}_{\ell_B}$, and

$$Q^{\ell'_A,(\ell'_B,\alpha)}(z|\ell_A, (\ell_B, 1))) = \begin{cases} 1 & z = \perp_0 \text{ and } \ell_A = \ell_B \\ 1 & z = \perp_1 \text{ and } \ell_A \neq \ell_B \\ 0 & \text{otherwise} . \end{cases}$$

Rewards result from the interaction, or lack thereof, between the defender and the attacker. If an attack is performed in ℓ, we denote by $U_A^u(\ell) < 0$ and $U_A^c(\ell) > 0$ the penalty and benefit for the defender when ℓ is uncovered (superscript u) or covered (superscript c), respectively. Similarly for Player B the values $U_B^u(\ell) > 0$ and $U_B^c(\ell) < 0$ are the benefit and penalty for the attacker when ℓ is uncovered or covered. Instant rewards r_A and r_B are defined as the expected values of the rewards $R_i = R_i^{ab}(z|s)$, $i \in \{A, B\}$, of the dynamics between

players, see [17, Ch. 2, p. 20]. These rewards depend on the current state of the system s, the actions (a, b) performed by the players and the future state of the system z, and for this game are given in (30).

$R_A^{ab}(z\|s)$	$R_B^{ab}(z\|s)$	$z = (\ell_A, (\ell_B, \alpha))$
$U_A^u(\ell_B)$	$U_B^u(\ell_B)$	$\ell_A \neq \ell_B$ and $\alpha = 1$
$U_A^c(\ell_B)$	$U_B^c(\ell_B)$	$\ell_A = \ell_B$ and $\alpha = 1$
$P_A(\ell_A)$	$P_B(\ell_B)$	$\alpha = 0$
$P_A(\perp_0)$	$P_B(\perp_0)$	$z = \perp_0$
$P_A(\perp_1)$	$P_B(\perp_1)$	$z = \perp_1$

(30)

The value $P_B(\ell_B) < 0$ represents the opportunity cost and risk for the attacker of being in location ℓ_B and not perform an attack. Values $P_A(\perp_i)$ and $P_B(\perp_i)$ represent the residual values of being in an absorbing state. We assume $P_A(\perp_0) > 0$ and $P_A(\perp_1) < 0$ and the opposite for the attacker. These definitions give the instants rewards r_A and r_B as follows:

$$r_i^{ab}(s) = \sum_{z \in \mathcal{S}} Q^{ab}(z|s) R_i^{ab}(z|s) \qquad i \in \{A, B\} \ .$$

Before reaching an absorbing state, the dynamics of the game are as follows: the system begins in a state defined by the location of both players and the decision to attack. Then, the defender chooses a strategy f (probably mixed) over the locations reachable from the current location. The attacker observes this strategy and chooses where to move and whether to attack or not. We denote this action as g. Note, that if the attacker decides to attack, the success or failure of this action will be revealed in the next state. The system evolves to the following state influenced by f, g, and Q. Both players receive their payoffs.

B. Computational study

Here we evaluate the solution algorithms presented in terms of solution time and quality of the solution obtained. We begin by describing the instances of the graph surveillance problem constructed for this computational study. We compare the solution times of value iteration and policy iteration on every instance considered. Solving the (MP) formulation was not a competitive solution approach for these problems, as it took a state-of-the-art non-linear optimization solver more than 5 hours of computational time for the smallest instance considered. We therefore do not present computational results for (MP). The experimental setup used is the one described in Section IV. To evaluate the quality of the FPE solutions obtained by these algorithms we compare them to heuristic policies, both in the myopic and non-myopic follower case.

We generate instances of different structure and size by considering different graphs to patrol on: paths, cycles, T-shaped graphs and complete graphs. We limit the size of $\mathcal{A}_s$ and $\mathcal{B}_s$ by limiting the distance that each player can travel from one time step to the next. To do so, we introduce the parameter k as the maximum geodesic distance that each player can travel through one time step.

Functions $q_A^{\ell'}(\ell''|\ell)$ are a function of the nodes that are in the shortest path between ℓ and ℓ'. We denote this set of nodes as $SP(\ell, \ell') = \{\ell, \ell_{i_2}, \ldots, \ell_{i_{k-1}}, \ell'\}$. Probabilities q_A are defined as follows: if $|SP(\ell, \ell')| = 1$ and $SP(\ell, \ell) = \{\ell\}$, then $q_A^{\ell}(\ell|\ell) = 1$, if $|SP(\ell, \ell')| \geq 2$ then

$$q_A^{\ell'}(\ell''|\ell) = \begin{cases} 1-\epsilon & \ell' = \ell'' \\ \dfrac{\epsilon}{|SP(\ell,\ell')|-1} & \ell'' \in SP(\ell,\ell')\backslash\{\ell'\} \\ 0 & \text{otherwise.} \end{cases}$$

In our experiments we set the probability of failing to $\epsilon = 0.25$. We assume $q_B^{\ell'}(\ell''|\ell) = \mathbf{1}_{\ell'=\ell''}$ are deterministic: the attacker always succeeds with its intended move.

The payoff functions are defined in Table VIII. The values of each parameter depend on the degree of the node, representing the fact that nodes with greater degree are more important in order to keep the connectivity of the graph.

Parameter	Value	Parameter	Value
$U_A^u(\ell)$	-10$deg(\ell)$	$U_A^c(\ell)$	10$deg(\ell)$
$U_B^u(\ell)$	$2^{deg(\ell)}$	$U_B^c(\ell)$	$-2^{deg(\ell)}$
$P_A(\ell_A)$	0	$P_B(\ell_B)$	$-deg(\ell_B)$
$P_A(\perp_0)$	1	$P_A(\perp_1)$	-1
$P_B(\perp_0)$	-1	$P_B(\perp_1)$	1

TABLE VIII
PAYOFF FUNCTIONS DESCRIPTION

We test our models in instances with $n \in \{5, 10\}$, $\beta_B = \{0, 0.9\}$ and $k \in \{2, 3\}$ for each type of graph. The instances considered range from 52 to 202 states. The actions spaces vary from an average of 3.7 to 9.9 actions per state for player A and from 7.3 to 19.8 for player B. The stopping criterion parameter is set to $\varepsilon = 10^{-3}$ for both value iteration and policy iteration.

Figure 3 shows the solution times in a performance profile in logarithmic scale comparing value iteration and policy iteration over all the instances considered. Policy iteration has faster solution times over the instances tested.

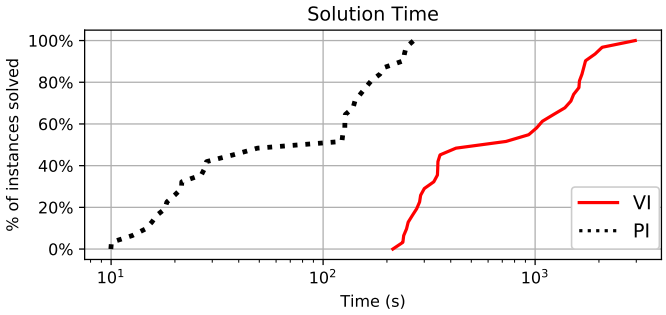

Fig. 3. Solution time performance profile.

We now aim to evaluate the quality of the Stackelberg equilibrium solution for this discounted game. However, since we do not have a method to compute the SSSE in general, we use the FPE that can be computed using value iteration as a proxy. Accordingly, we compute the values of equilibrium $v^* = (v_A^*, v_B^*)$, with the respective equilibrium policies f^* and g^* using the value iteration algorithm.

We compare the FPE solution to static policies obtained by ignoring the dynamic nature of the game. We refer to these heuristic policies as Myopic policies. To determine the Myopic

policy, for each state we compute the Strong Stackelberg policies, f^M, g^M of the static game, that is with $\beta_A = \beta_B = 0$. Finally, we evaluate this policy in the dynamic setting: we obtain the value $V^{f^M g^M} = v^M = (v_A^M, v_B^M)$ as the fixed point of operator $T^{f^M g^M}$ (see (12)) with real values of β_A and β_B.

Finally, in order to compare the policies obtained for both methods we compare the average value for applying each policy, denoted respectively as $\bar{v}_A^*, \bar{v}_B^*, \bar{v}_A^M$ and $\bar{v}_B^M$, where $\bar{v} = |\mathcal{S}|^{-1} \sum_{s \in \mathcal{S}} v(s)$.

Table IX shows the comparison of the values for the different types of graph structures mentioned before, and with the parameters $n = 10$, $k = 2$, $\beta_A = 0.9$ and $\beta_B = 0$. Recall that in this case, value iteration converges to an FPE (and SSSE) because the game is MFS. For the complete graph, the myopic strategy generates in average the same reward as the equilibrium strategy. In the other cases, the SSSE strategy outperforms the myopic-heuristic policy.

type	$\bar{v}_A^*$	$\bar{v}_A^M$	$\bar{v}_B^*$	$\bar{v}_B^M$
Cycle	9.957	8.376	1.485	2.079
Path	9.070	6.686	1.109	1.703
T	10.623	8.129	0.703	2.218
Complete	89.595	89.595	129774.653	129774.653

TABLE IX
EVALUATION OF THE SOLUTION CONCEPT WITH $\beta_B = 0$.

We repeat the same evaluation, but now with $\beta_B = 0.9$. Note that in this case there are no guarantees that value iteration will converge, or that the FPE policy is an SSSE. Value iteration found an FPE (i.e. converged) in all the instances. Table X shows the average values obtained applying the policies provided by the FPE and the myopic heuristic.

type	$\bar{v}_A^*$	$\bar{v}_A^M$	$\bar{v}_B^*$	$\bar{v}_B^M$
Cycle	-17.667	6.767	6.261	2.171
Path	-14.102	4.938	6.617	1.506
T	-12.955	6.143	6.739	2.249
Complete	89.595	89.604	129773.771	129773.762

TABLE X
EVALUATION OF THE SOLUTION CONCEPT WITH $\beta_B = 0.9$.

Note that in relative terms, both the FPE policy and the Myopic policy return a lower average value for Player A. This can be attributed to the follower's change of behavior. More significantly, the FPE solution performs worse than the Myopic strategy for all graphs.

VI. Conclusions and Further Work

This work demonstrates the relevance of the concept of Strong Stationary Stackelberg Equilibria, SSSE, and the related operator-based algorithms, for the computation of policies in the context of two-player discounted stochastic games.

For this, we first define a suitable operator acting on the set of value functions for both players. We introduce the concept of Fixed-Point Equilibrium, FPE, as the fixed points of this operator. We then investigate the relationship between SSSE and FPE. We show that neither need to exist in general, and that when they do, they do not necessarily coincide. We also show that the solution based on mathematical programming suggested in the literature, does not necessarily compute a correct answer. We have nevertheless identified several classes of games where SSSE and FPE do exist and do coincide. Among these is the class of games with Myopic Follower Strategies, MFS, which include games with myopic followers and games with leader controlled transitions.

We consider an application in a security domain, in which a moving defender protects locations on a graph that can be attacked by a moving attacker. We give a formulation of this problem as a Stackelberg equilibrium in a discounted stochastic game. The value iteration and policy iteration algorithms are able to compute efficiently the FPE for the instances considered of this security application. The instances considered are too difficult to solve if using the mathematical programming formulation of the problem. In the case of myopic follower, in which the FPE corresponds to the SSSE, we observe that the solution obtained is efficient and outperforms heuristic policies. However, in examples without MFS, we see that the FPE solution is worse than heuristic policies.

An important contribution of our work is the extension of convergence proofs for VI and PI that exist for MDPs to discounted Stackelberg stochastic games. This result is challenging since we had to identify appropriate operators and necessary hypotheses that guarantee that these operators are contractive. Proving the existence of the SSSE, done by directly checking that the definition holds, is also challenging because this process involves finding optimal solutions of non-convex problems. Another theoretical challenge in this work is to identify conditions that can guarantee the equivalence between the different notions of equilibrium (SSSE and FPE), question that we have partially addressed in this manuscript.

Future research will aim at identifying more general sufficient conditions for the two concepts, SSSE and FPE, to coincide. The problem of finding general methodologies to detect the existence of SSSE is still open. It is also important to determine algorithms to find the equilibrium in games which possess SSSEs but do not satisfy the MFS condition. It will also be interesting to determine whether the use of value iteration or policy iteration, when they do not converge, can nevertheless produce nonstationary strategies with a good performance. Given the difficulty of finding an SSSE, an interesting research direction would be to relax the notion of equilibrium to an $\epsilon-$SSSE, that is, solutions that increases the size of the sets of best response. The latter framework will imply new definitions of equilibria and new algorithmic approaches. Another relevant future research will focus on expand this methodology to the multiple attacker case. Finally, we should consider other algorithmic frameworks, such as the use of Q-learning in policy iteration. For instance, [23] uses an asynchronous implementation of it for MDPs, proving its correctness without the need for monotonicity, and showing good computational results.

ACKNOWLEDGMENT

Bucarey was funded by the CONICYT PFCHA/DOCTORADO NACIONAL/2013 - 21130556 and the Fonds de la Recherche Scientifique -FNRS under Grant(s) no PDR T0098.18. Ordóñez was supported for this research by the Complex Engineering Systems Institute through grant CONICYT-PIA-FB0816 and CONICYT through grant FONDECYT 1201844. .

REFERENCES

[1] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Society for Industrial and Applied Mathematics, 1998.

[2] A. Bagchi, "Some economic applications of dynamic Stackelberg games," in *Dynamic Games and Applications in Economics*, T. Başar, Ed. Berlin, Heidelberg: Springer, 1986, vol. 265, pp. 88–102, lecture Notes in Economics and Mathematical Systems.

[3] T. Li and S. P. Sethi, "A review of dynamic Stackelberg game models," *Discrete Cont. Dyn.-B*, vol. 22, no. 1, pp. 125–159, 2017.

[4] B.-W. Wie, "Dynamic Stackelberg equilibrium congestion pricing," *Transport. Res. C-Emer.*, vol. 15, no. 3, pp. 154–174, 2007.

[5] F. M. Delle Fave, A. X. Jiang, Z. Yin, C. Zhang, M. Tambe, S. Kraus, and J. P. Sullivan, "Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system," *J. Artif. Intell. Res.*, vol. 50, pp. 321–367, 2014.

[6] T. Bewley and E. Kohlberg, "On stochastic games with stationary optimal strategies," *Mathematics of Operations Research*, vol. 3, no. 2, pp. 104–125, 1978.

[7] D. Kar, T. Nguyen, F. Fang, M. Brown, A. Sinha, M. Tambe, and A. Jiang, "Trends and applications in Stackelberg security games," in *Handbook of Dynamic Game Theory*, T. Başar and G. Zaccour, Eds. Springer, 2018, ch. 28, pp. 1223–1269.

[8] J. B. Clempner and A. S. Poznyak, "Stackelberg security games: Computing the shortest-path equilibrium," *Expert Syst. Appl.*, vol. 42, no. 8, pp. 3967–3979, 2015.

[9] G. Leitman, "On generalized Stackelberg strategies," *J. Optim. Theory Appl.*, vol. 26, no. 4, pp. 637–643, 1978.

[10] M. Breton, A. Alj, and A. Haurie, "Sequential Stackelberg equilibria in two-person games," *J. Optim. Theory Appl.*, vol. 59, no. 1, pp. 71–97, 1988.

[11] S. Dempe, *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

[12] M. Simaan and J. B. Cruz Jr., "On the Stackeberg strategy in nonzero-sum games," *J. Optim. Theory Appl.*, vol. 11, no. 5, pp. 533–555, 1973.

[13] ——, "Additional aspects on the Stackeberg strategy in nonzero-sum games," *J. Optim. Theory Appl.*, vol. 11, no. 6, pp. 613–626, 1973.

[14] E. W. Denardo, "Contraction mappings in the theory underlying dynamic programming," *SIAM Review*, vol. 9, no. 2, pp. 165–177, 1967.

[15] W. Whitt, "Representation and approximation of noncooperative sequential games," *SIAM J. Control Optim.*, vol. 18, no. 1, pp. 33–48, 1980.

[16] Y. Vorobeychik and S. Singh, "Computing Stackelberg equilibria in discounted stochastic games," in *AAAI Conference*, 2012, https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4811/5686. Corrected version retreived Oct. 19, 2018.

[17] M. L. Puterman, *Markov decision processes*. Wiley-Interscience, 1994.

[18] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[19] Y. Vorobeychik, B. An, M. Tambe, and S. Singh, "Computing solutions in infinite-horizon discounted adversarial patrolling games," in *Proc. ICAPS*, 2014.

[20] J. Letchford, L. MacDermed, V. Conitzer, R. Parr, and C. L. Isbell, "Computing optimal strategies to commit to in stochastic games," in *AAAI Conference*, 2012.

[21] V. Bucarey, E. Della Vecchia, A. Jean-Marie, and F. Ordóñez, "Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games," INRIA, Research Report RR-9271 v3, Mar. 2021. [Online]. Available: https://hal.inria.fr/hal-02144095

[22] C. Casorrán, B. Fortz, M. Labbé, and F. Ordóñez, "A study of general and security Stackelberg game formulations," *European Journal of Operational Research*, vol. 278, no. 3, pp. 855–868, 2019.

[23] D. P. Bertsekas and H. Yu, "Q-learning and enhanced policy iteration in discounted dynamic programming," *Mathematics of Operations Research*, vol. 37, no. 1, pp. 66–94, 2012.

Victor Bucarey is a professor and researcher in the Institute of Engineering Sciences at Universidad de O'Higgins, Chile. He received the Ph.D. degree in Engineering Systems from Universidad de Chile in 2017.

Eugenio Della Vecchia is a professor and researcher in the Department of Mathematics at Universidad Nacional de Rosario (UNR), Argentine. He received his PhD in from UNR, in 2013.

Alain Jean-Marie is senior researcher at Inria, the French National Institute for Research in Computer Science and Automation. He received his PhD degree from the University of Paris XI at Orsay in 1987.

Fernando Ordoñez is a professor in the Department of Industrial Engineering at Universidad de Chile. He obtained his Ph.D. in Operations Research from MIT in 2002.