Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games

Víctor Bucarey, Eugenio Della Vecchia, Alain Jean-Marie, Fernando Ordóñez

Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games

Víctor Bucarey, Eugenio Della Vecchia, Alain Jean-Marie, Fernando Ordóñez

ISSN 0249-6399 ISRN INRIA/RR--9271--FR+ENG

RESEARCH REPORT
N° 9271
May 2019
Project-Teams NEO and INOCS

Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games

Víctor Bucarey[*†], Eugenio Della Vecchia[‡], Alain Jean-Marie[§], Fernando Ordóñez[¶]

Project-Teams NEO and INOCS

Research Report n° 9271 — May 2019 — 62 pages

Abstract: In this work we study strong Stackelberg equilibria in stationary policies for discounted stochastic games, named (SSSE). We provide classes of games where the SSSE exists, and we prove via counterxamples that SSSE does not exist in the general case. We define suitable dynamic programming operators and we study their fixed points, named FPE. We show that the FPE and SSSE coincides for some games. In particular, we introduce the class of games with Myopic Follower Strategy, which have this property. We study the behaviour of Value Iteration, Policy Iteration and Mathematical programming formulations for this problem. Finally, we show an application in security in order to test the solution concepts and the efficiency of the algorithms studied in this article.

Key-words: Stochastic games, Stackelberg Equilibrium, Optimal control

[*] Département d'Informatique, Université Libre de Bruxelles, Brussels, Belgium.
[†] Inria Lille-Nord Europe, Villeneuve d'Ascq, France. vbucarey@ulb.ac.be
[‡] Departmento de Matemática, Universidad Nacional de Rosario, Argentina. eugenio@fceia.unr.edu.ar
[§] Inria, LIRMM, University of Montpellier, CNRS, Montpellier, France. alain.jean-marie@inria.fr
[¶] Departamento de Ingeniería Industrial, Universidad de Chile, Chile. fordon@dii.uchile.cl

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Équilibre de Stackelberg Stationnaire Fort dans les jeux stochastiques actualisés

Résumé : Dans cet article, nous étudions les Équilibres Forts de Stackelberg en politiques stationnaires (*Strong Stationary Stackelberg Equilibria*, SSSE) pour les jeux stochastiques actualisés. Nous exhibons des classes de jeux pour lesquels le SSSE existe et nous montrons par des contre-exemples qu'il n'en existe pas dans le cas général. Nous définissons des opérateurs de programmation dynamique appropriés pour ce concept, et étudions leurs points fixes, nommés FPE (*Fixed Point Equilibria*). Nous montrons que le FPE et le SSSE coïncident pour certaines classes de jeux stochastiques. En particulier, nous introduisons la classe des jeux avec stratégie du "follower" myope (*Myopic Follower Strategy*, MFS) pour laquelle cette propriété est vraie. Nous étudions le comportement des algorithmes Value Iteration, Policy Iteration la formulation par Programmation Mathématique de ce problème. Finalement, nous décrivons une application dans le domaine de la sécurité afin de tester le concept de solution et l'efficacité des algorithmes introduits.

Mots-clés : Jeux stochastiques, Équilibre de Stackelberg, Contrôle optimal

1 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 can be extended to a multistage setting where leader and followers repeatedly make strategic decisions. Such dynamic Stackelberg models have been considered in applications in economics [1], marketing and supply chain management [11], dynamic congestion pricing [18], and security [5].

For example, in a dynamic security application a defender could decide on a patrolling strategy over multiple periods and the attackers would take this defender patrol into consideration when deciding whether to attack in each period. The literature, in this security application setting, seeks to compute the Strong Stackelberg Equilibrium solution [8] and Strong Stackelberg Equilibrium in *stationary strategies* (SSSE) in the case of stochastic games [5]. Previous work features algorithms for computing such equilibria, using bi-level mathematical programs. Such programs are notoriously difficult to solve in general, which motivates research on alternative solution methods.

In this paper, we propose to use *iterative* algorithms, based on the operator formalism, to compute the *strong stationary Stackelberg equilibrium* (SSSE) solution of stochastic games. Our proposal is to use the well-known algorithms for solving Markov Decision Processes, Value Iteration and Policy Iteration, based on a suitably defined *dynamic programming operator*. Such algorithms are both conceptually simple and have a small computational burden per iteration. However, their use raises the question of convergence: do they converge, and if so, do they converge to some SSSE? Answering these questions led us to: a) conclude that SSSE do not necessarily exist, something not obvious from the current litterature; b) identify classes of games where iterative algorithms converge to a SSSE. We then exploited this property in the analysis of a model of dynamic planning of police patrols on a transportation graph.

In the remainder of this introduction we clarify the problem under consideration, present related literature, describe the contributions of our work and introduce the notation that is used in this paper.

1.1 Problem Statement

We present now the notion 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, 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 state s and commits to a, possibly mixed, strategy f depending 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 there is an immediate payoff 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 another state s'. This dynamics is illustrated in the following scheme:

$$s_0 \rightsquigarrow \begin{array}{c}\text{Player } A\\ \text{chooses } f_0\end{array} \rightsquigarrow \begin{array}{c}\text{Player } B\\ \text{observes } f_0\\ \text{and chooses } g_0\end{array} \underbrace{\rightsquigarrow}_{\substack{Q^{f_0 g_0}(s_1|s_0)\\ r_A^{f_0 g_0}(s_0),\, r_B^{f_0 g_0}(s_0)}} s_1 \rightsquigarrow \begin{array}{c}\text{Player } A\\ \text{chooses } f_1\end{array} \rightsquigarrow \begin{array}{c}\text{Player } B\\ \text{observes } f_1\\ \text{and chooses } g_1\end{array} \rightsquigarrow \cdots$$

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 her own payoff. This general "Stackelberg" approach to the solution of the game is complemented with the rule that when the follower is indifferent between several strategies, she chooses the one that benefits the leader: this refinement is the *strong* Stackelberg solution.

1.2 Related bibliography

The study of Strong Stackelberg Equilibria (SSE) has received much attention in the recent literature due to its relevance in security applications [8]. 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 Breton *et al* [2] as a Weak Stackelberg Equilibrium, together with the definition of the optimistic generalization, the Strong Stackelberg Equilibrium. This reference also points out the relationship between SSE and bi-level optimization.

Stackelberg equilibria in multi-stage and dynamic games have been studied by Simaan and Cruz in [13, 14]. In particular, these authors propose in [13] to focus on *feedback* strategies that can be obtained via dynamic programming. The idea is reused in [2] which introduces Strong *Sequential* Stackelberg Equilibria, in a setting very 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 [6] and developed in Whitt [17], is essential. Our analysis uses this formalism as well.

The stochastic game model we study in this paper is also the topic of Vorobeychik and Singh in [15]. In that work, the authors show that SSSE always exist in stochastic games for *team* games where both players have the same rewards. They also propose mathematical programming formulations to find the SSSE, extending the analysis for Markov Decision Processes (see [12, ch. 6.]) and Nash equilibrium in stochastic games [7], for this case. Similar mathematical programming formulations are established in [5] and [16] 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 provide a discussion of the mathematical programming formulations and what they compute when a SSSE does not exist.

The complexity of computing a SSE is studied in Letchford et al. [10]. That work shows, by reduction to 3SAT, that it is NP-hard to determine a SSE for a Stackelberg Stochastic Game with any discount factor $\beta > 0$ common for both players. The possibility that a Stackelberg Stochastic Game does not have a SSE is not mentioned.

As mentioned above, security applications are an important motivation for research on dynamic games. It is interesting to single out the stochastic game model described in [7, Chapter 6.3]. 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. An attacker-defender Stackelberg security game is also considered in [3] for a repeated stochastic Markov chain game. This problem is represented as a potential game in terms of a suitable Lyapunov function, which is used to prove convergence results to compute the strong Stackelberg equilibrium [4].

1.3 Contribution

While previous works have 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 or not 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. Furthermore, no work to date has advocated the use of the operator method, neither for the mathematical analysis of the problem, nor for its algorithmic solution.

We contribute to the issue in the following ways.

First, we give a formal definition of the Strong Stationary Stackelberg Equilibrium (SSSE) in stochastic games (Section 1.4). 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. This is the topic of Section 2.

Next, we focus on the general question of existence of SSSE and FPE, and how they are related. We show that games with MFS 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, Team Games and Acyclic Games. This analysis is developed in Section 3.

We then illustrate different situations with specific examples. In a first case, a FPE and a SSSE exist and coincide, 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 exist, or no FPE exist, or both a SSSE and a FPE exist but they do not coincide. In a third case, a FPE exists but Value Iteration does not necessarily converge to it. These examples are summarized in Section 4 and described in more detail in Appendix C.

Finally, we take advantage of the convergence properties we have shown, to propose a solution methodology in a dynamic game of security, representing the patrolling of security forces in a transportation network. This is reported in Section 5.

1.4 Notation and definitions

We introduce now formally the elements of the model and the notation. A synthesis of this notation is presented in Appendix A.

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 $s' \in \mathcal{S}$. 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

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

Strategies. The set of mixed strategies for players A and B available in state s are, respectively, represented by $\mathbb{P}(\mathcal{A}_s)$ and $\mathbb{P}(\mathcal{B}_s)$: the sets of distribution functions over $\mathcal{A}_s$ and $\mathcal{B}_s$. Define the sets of feedback stationary strategies:

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

For $f \in W_A$, $f(s)$ is a probability measure on $\mathcal{A}_s$. In order to simplify the notation, we will note, $f(s, a) = (f(s))(\{a\}) = f(a|s)$ 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 $g \in W_B$ represents 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 sets W_A and W_B are assumed to be equipped with total orders $\prec_A$ and $\prec_B$.

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

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

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:

$$V_i^{fg}(s) = \mathbb{E}_s^{fg}\left[\sum_{k=0}^{\infty} \beta_i^k r_i^{f(S_k), g(S_k)}(S_k)\right] \tag{2}$$

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

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 Markov strategy, the follower faces a finite-state, finite-action, discounted Markov Decision Process (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 [12].

Accordingly, let:

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

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

$$\gamma_B(f) := \max_{\prec_B} SR_B(f) \tag{5}$$

$$R_A(s) := \{f \in W_A \mid V_A^{f\gamma_B(f)}(s) \geq V_A^{h\gamma_B(h)}(s), \forall h \in W_A\} . \tag{6}$$

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.

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 $W_A \times W_B$. It corresponds to the definitions in [10, 15].

Definition 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 a 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 4.3 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 a SSSE. We address in Section 3 this issue in special cases.

2 Operators, Fixed Points and Algorithms

In this section, we develop the formalism of operators, as commonly found in texts on MDPs [12], and also for games in [2, 6, 17]. 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.

2.1 Definition of operators

We start with the definition of one-step (or "return function" [6]) 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

$$\left(T_i^{fg} v\right)(s) = \sum_{a \in \mathcal{A}_s} f(s,a) \sum_{b \in \mathcal{B}_s} g(s,b) \left[r_i^{ab}(s) + \beta_i \sum_{z \in \mathcal{S}} Q^{ab}(z|s) v(z) \right] . \tag{7}$$

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 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 that turns out to be $V^{fg} = (V_A^{fg}, V_B^{fg})$, these functions being defined in (2).

Extended reaction sets. We now extend the definitions of reaction sets to involve value functions. 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 2.1 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:

$$R_B(s, f, v_B) := \{g \in \mathcal{B}_s \mid (T_B^{fg} v_B)(s) \geq (T_B^{fh} v_B)(s), \forall h \in \mathcal{B}_s\} \tag{8}$$

$$SR_B(s, f, v) := \{g \in R_B(s, f, v_B) \mid (T_A^{fg} v_A)(s) \geq (T_A^{fh} v_A)(s), \forall h \in R_B(s, f, v_B)\} \tag{9}$$

$$\gamma_B(s, f, v) := \max_{\prec_B} SR_B(s, f, v) \tag{10}$$

$$R_A(s, v) := \{f(s) \in \mathbb{P}(\mathcal{A}_s) \mid (T_A^{f\gamma_B(s,f,v)} v_A)(s) \geq (T_A^{h\gamma_B(s,h,v)} v_A)(s), \forall h \in \mathbb{P}(\mathcal{A}_s)\} \ . \tag{11}$$

The definition of Player B's response in (10) 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.

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 2 (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:

$$(Tv)_i(s) = \left(T_i^{R_A(s,v),\gamma_B(s,R_A(s,v),v)} v_i\right)(s) \tag{12}$$

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.

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

Definition 3 (Fixed Point Equilibrium, FPE)**.** A strategy pair $(f, g) \in W_A \times W_B$ is a FPE if the function $v^* = V^{fg}$, 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 $s \in \mathcal{S}$.

One of the purposes of this paper is to propose results concerning FPEs and SSSE: discuss whether they respectively exist, and when they do, whether they coincide or not.

2.2 Properties of operators

The following property is well-known (e.g. [12]) for finite-state, finite-action discounted Markov Reward Processes:

Lemma 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

$$V_i^{fg} \;=\; (I - \beta_i Q^{fg})^{-1} r_i^{fg} \;, \tag{13}$$

where r_i^{fg} is defined in (1), where the probability transition matrix Q^{fg} is defined similarly in Section 1.4, 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 4 (Myopic Follower Strategy, MFS)**.** A stochastic game $\mathcal{G}$ is said to be with *Myopic Follower Strategy* if $R_B(s, f, v_B) = R_B(s, f)$, for all $s \in \mathcal{S}$, $f \in W_A$ and $v_B \in \mathcal{F}(\mathcal{S})$.

When a game is with MFS, the reaction of the follower depends only on the leader's value v_A:

$$\forall f \in W_A, \forall v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S}), \forall s \in \mathcal{S}, \qquad \gamma_B(s, f, v) = \overline{\gamma}_B(s, f, v_A). \tag{14}$$

Then the following Lemma is relevant.

Lemma 2. Assume (14) holds. Then there exists an operator $\overline{T}_A$ from $\mathcal{F}(\mathcal{S})$ to $\mathcal{F}(\mathcal{S})$ such that for all $v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S})$,

$$(Tv)_A = \overline{T}_A v_A. \tag{15}$$

Proof. According to definition (11) and because of (14), we have $R_A(s, v) = \overline{R}_A(s, v_A)$ for all $v \in \mathcal{F}(\mathcal{S})$ and $s \in \mathcal{S}$. Then, from (12),

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

□

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:

$$(\overline{T}_A^f v_A)(s) = (T_A^{f,\overline{\gamma}_B(s,f,v_A)} v_A)(s) \;. \tag{16}$$

Another consequence of MFS is this property which follows from the definition of $\gamma_B(\cdot)$ in (10) and $SR_B(\cdot)$ in (9):

$$(\overline{T}_A^f v_A)(s) = \max_{g \in R_B(s,f)} (T_A^{fg} v_A)(s) \;. \tag{17}$$

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 $s \in \mathcal{S}$:

$$(\overline{T}_A v_A)(s) = \max_{f \in W_A} (\overline{T}^f_A v_A)(s) \ . \tag{18}$$

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 $\overline{T}^f_A : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$, defined in (16) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$ of modulus β_A.

b) The operator $\overline{T}_A$ defined in (18) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$, of modulus β_A.

c) For any stationary strategy $f \in W_A$, operator $\overline{T}^f_A$ is monotone.

Proof. The central argument of the proof is the following fact. Let h_1 and h_2 be two real functions defined on some set B, where they attain their maximum. Then for all $b_1 \in \arg\max_B\{h_1(b)\}$ and all $b_2 \in \arg\max_B\{h_2(b)\}$,

$$h_1(b_2) - h_2(b_2) \ \leq \ \max_B\{h_1(b)\} - \max_B\{h_2(b)\} \ \leq \ h_1(b_1) - h_2(b_1) \ . \tag{19}$$

To show a), take $v_A, u_A \in \mathcal{F}(\mathcal{S})$, a stationary strategy $f \in W_A$, and $s \in \mathcal{S}$. Then, using (17),

$$\begin{aligned}(\overline{T}^f_A v_A)(s) - (\overline{T}^f_A u_A)(s) = & \max_{b \in R_B(s,f)} \sum_{a \in \mathcal{A}_s} f(s,a) \left[r^{ab}_A(s) + \beta_A \sum_{z \in \mathcal{S}} Q^{ab}(z|s) v_A(z) \right] \\ & - \max_{b \in R_B(s,f)} \sum_{a \in \mathcal{A}_s} f(s,a) \left[r^{ab}_A(s) + \beta_A \sum_{z \in \mathcal{S}} Q^{ab}(z|s) u_A(z) \right] \ . \end{aligned} \tag{20}$$

Then from (19), there exists $b \in R_B(s,f)$ such that:

$$\begin{aligned}(\overline{T}^f_A v_A)(s) - (\overline{T}^f_A u_A)(s) & \leq \sum_{a \in \mathcal{A}_s} f(s,a) \Big[r^{ab}_A(s) - r^{ab}_A(s) \\ & \qquad + \beta_A \sum_{z \in \mathcal{S}} \left(Q^{ab}(z|s) v_A(z) - Q^{ab}(z|s) u_A(z) \right) \Big] \\ & = \sum_{a \in \mathcal{A}_s} f(s,a) \beta_A \sum_{z \in \mathcal{S}} Q^{ab}(z|s)(v_A(z) - u_A(z)) \\ & \leq \beta_A ||v_A - u_A||_\infty \ . \end{aligned} \tag{21}$$

By reversing the roles of v_A and u_A, then taking the maximum over $s \in \mathcal{S}$ we have that:

$$||\overline{T}^f_A v_A - \overline{T}^f_A u_A||_\infty = \max_{s\in\mathcal{S}} |(\overline{T}^f_A v_A)(s) - (\overline{T}^f_A u_A)(s)| \leq \beta_A ||v_A - u_A||_\infty ,$$

concluding that $\overline{T}^f_A$ is a contracting map of modulus β_A.

In order to show b), take $v_A, u_A \in \mathcal{F}(\mathcal{S})$, a state $s \in \mathcal{S}$ and f^* any optimal policy realizing $\max_{f\in W_A}(\overline{T}^f_A v_A)(s)$. Then,

$$\begin{aligned}
(\overline{T}_A v_A)(s) - (\overline{T}_A u_A)(s) &= \max_{f\in\mathbb{P}(A_s)} (\overline{T}^f_A v_A)(s) - \max_{f\in\mathbb{P}(A_s)} (\overline{T}^f_A u_A)(s) \\
&= (\overline{T}^{f^*}_A v_A)(s) - \max_{f\in\mathbb{P}(A_s)} (\overline{T}^f_A u_A)(s) \\
&\leq (\overline{T}^{f^*}_A v_A)(s) - (\overline{T}^{f^*}_A u_A)(s) \\
&\leq \beta_A ||v_A - u_A||_\infty .
\end{aligned}$$

Then, by reversing the roles of v_A, u_A and taking the maximum the result follows.

Consider now statement c). If $v_A \geq u_A$, then from (20) with (19) and (21), there exists some $b \in R_B(s, f)$ such that:

$$\begin{aligned}
(\overline{T}^f_A v_A)(s) - (\overline{T}^f_A u_A)(s) &\geq \sum_{a\in\mathcal{A}_s} f(s,a)\beta_A \sum_{z\in\mathcal{S}} Q^{ab}(z|s)(v_A(z) - u_A(z)) \\
&\geq 0 .
\end{aligned}$$

Then c) also holds. □

2.3 Value Iteration Algorithms

Value Iteration generally consists in applying a dynamic programming operator to some initial value function, until convergence is thought to occur. Specifically, given some $\varepsilon > 0$, Value Iteration applies some operator repeatedly until the distance between two functions v^n_A and v^{n+1}_A is less than ε.

In view of the preceding discussion, two variants of the algorithm will be used: one for the general situation (Algorithm 1) and one for the specific situation where Property (14) holds (Algorithm 2).

Algorithm 1 Value function iteration for infinite horizon; general case

Require: $\varepsilon > 0$

Initialize with $n = 0$, $v_A^0(s) = v_B^0(s) = 0$ for every $s \in \mathcal{S}$

repeat

  $n := n + 1$

  Compute v^n as

$$v^n(s) := (Tv^{n-1})(s), \qquad \forall s \in \mathcal{S} \tag{22}$$

  with T according to Definition (12)

until $||v^n - v^{n-1}||_\infty \leq \varepsilon$

Pick (f^*, g^*) such that $v^n(s) = (T^{f^*g^*}v^{n-1})(s)$ for all $s \in \mathcal{S}$

return Approximate Stationary Strong Stackelberg policies (f^*, g^*)

Algorithm 2 Value function iteration for infinite horizon; simplified case

Require: $\varepsilon > 0$

Initialize with $n = 0$, $v_A^0(s) = 0$ for every $s \in \mathcal{S}$

repeat

  $n := n + 1$

  Compute v_A^n as

$$v_A^n(s) := (\overline{T}_A v_A^{n-1})(s), \qquad \forall s \in \mathcal{S} \tag{23}$$

  with $\overline{T}$ as in (18)

until $||v_A^n - v_A^{n-1}||_\infty \leq \varepsilon$

Pick f^* such that $v_A^n(s) = (\overline{T}_A^{f^*} v_A^{n-1})(s)$ and g^* such that $g^*(s) = \gamma_B(s, f^*, v_A^{n-1})$, for all $s \in \mathcal{S}$

return Approximate Stationary Strong Stackelberg policies (f^*, g^*)

There is no guarantee in general that Algorithm 1 will converge, and we present in Section 4.4 an example where it does not. However, thanks to Theorem 2, we can state that Algorithm 2 does converge.

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

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

Proof. Let the pair of policies (f^*, g^*) be the ones returned by Algorithm 2 and $V_A^{f^*g^*}$

be the fixed point of $\overline{T}_A^{f^*}$. By Theorem 1 b) and Banach's Theorem, we know that $\overline{T}_A$ has a unique fixed point, v_A^*. Then,

$$||V_A^{f^*g^*} - v_A^*||_\infty \le ||V_A^{f^*g^*} - v_A^n||_\infty + ||v_A^n - v_A^*||_\infty \ . \tag{24}$$

In (24), the first term on the right-hand side is bounded as follows:

$$\begin{aligned} ||V_A^{f^*g^*} - v_A^n||_\infty &= ||\overline{T}_A^{f^*} V_A^{f^*g^*} - v_A^n||_\infty \\ &\le ||\overline{T}_A^{f^*} V_A^{f^*g^*} - \overline{T}_A v_A^n||_\infty + ||\overline{T}_A v_A^n - v_A^n||_\infty \\ &= ||\overline{T}_A^{f^*} V_A^{f^*g^*} - \overline{T}_A^{f^*} v_A^n||_\infty + ||\overline{T}_A v_A^n - \overline{T}_A v_A^{n-1}||_\infty \\ &\le \beta_A ||V_A^{f^*g^*} - v_A^n||_\infty + \beta_A ||v_A^n - v_A^{n-1}||_\infty, \end{aligned}$$

where the first equality is by definition, the inequality right after is the triangular inequality. The third line is because of the definition of $\overline{T}_A^{f^*}$ and the inequality is because $\overline{T}_A^{f^*}$ and $\overline{T}_A$ are contracting maps of modulus β_A. This last inequality implies that

$$||V_A^{f^*g^*} - v_A^n||_\infty \le \frac{\beta_A}{1-\beta_A} ||v_A^n - v_A^{n-1}||_\infty.$$

For the second term on the right-hand side of (24), we have:

$$\begin{aligned} ||v_A^n - v_A^*||_\infty &= \lim_{t\to\infty} ||v_A^n - v_A^t||_\infty \\ &\le \lim_{t\to\infty} \sum_{k=0}^{t-n-1} ||v_A^{n+k} - v_A^{n+k+1}||_\infty \\ &\le \lim_{t\to\infty} \sum_{k=0}^{t-n-1} \beta_A^k ||v_A^{n-1} - v_A^n||_\infty \\ &= \frac{\beta_A}{1-\beta_A} ||v_A^{n-1} - v_A^n||_\infty. \end{aligned}$$

Then, we have that the policies returned by the algorithm satisfy:

$$||V_A^{f^*g^*} - v_A^*||_\infty \le 2\frac{\beta_A}{1-\beta_A} ||v_A^{n-1} - v_A^n||_\infty = \frac{2\beta_A}{1-\beta_A}\varepsilon \ .$$

Furthermore, given that

$$||v_A^{n-1} - v_A^n||_\infty \le \beta_A^{n-1} ||v_A^0 - v_A^1||_\infty = \beta^{n-1} ||v_A^1||_\infty \le \beta_A^{n-1} ||r_A||_\infty$$

the result follows. □

2.4 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 *Evaluation Phase* and *Improvement Phase*.

As in the previous section, two variants of the algorithm will be used: one for the general situation (Algorithm 3) and one for the specific situation where Property (14) holds (Algorithm 4).

Algorithm 3 Policy Iteration (PI); general case

Require $\varepsilon > 0$
Initialize with $n = 0$
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 $s \in \mathcal{S}$
Compute $u^0 = (u_A^0, u_B^0)$ fixed point of $T^{f_0 g_0}$
repeat
  $n := n + 1$
  Improvement Phase: Find a pair of strategies (f_n, g_n) such that $T^{f_n g_n} u^n = T u^n$ with $g_n(s) = \gamma_B(s, f_n, u^{n-1})$ for all $s \in \mathcal{S}$
  Evaluation Phase: Find $u^n = (u_A^n, u_B^n)$, fixed point of the operator $T^{f_n g_n}$
until $||u^n - u^{n-1}||_\infty \leq \varepsilon$
$f^* := f_n$; $g^*(s) := \gamma_B(s, f_n, u^n)$ for all $s \in \mathcal{S}$
return Approximate Stationary Strong Stackelberg policies (f^*, g^*)

Algorithm 4 Policy Iteration (PI); simplified case

Require $\varepsilon > 0$
Initialize with $n = 0$
Choose an arbitrary pair of strategies $(f_0, g_0) \in W_A \times W_B$ with $g_0(s) = \overline{\gamma}_B(s, f_0, \mathbf{0})$ for all $s \in \mathcal{S}$
Compute u_A^0 fixed point of $\overline{T}_A^{f_0}$
repeat
  $n := n + 1$
  Improvement Phase: Find a distribution f_n such that $\overline{T}_A^{f_n} u_A^{n-1} = \overline{T}_A u_A^{n-1}$
  Evaluation Phase: Find u_A^n fixed point of the operator $\overline{T}_A^{f_n}$
until $||u_A^n - u_A^{n-1}||_\infty \leq \varepsilon$
$f^* := f_n$; $g^*(s) := \overline{\gamma}_B(s, f_n, u_A^n)$ for all $s \in \mathcal{S}$
return Approximate Stationary Strong Stackelberg policies (f^*, g^*)

The Evaluation Phase in Algorithm 3 (respectively Algorithm 4) requires to solve two (resp. one) linear systems of size $|\mathcal{S}| \times |\mathcal{S}|$. On the other hand, the Improvement

Phase can be implemented by solving a static Strong Stackelberg equilibrium for each state $s \in \mathcal{S}$. Now we prove that Algorithm 4 converges to the SSSE. In other words, the PI algorithm converges to the SSSE for stochastic games with MFS.

Lemma 3. If a function $v_A \in \mathcal{F}(\mathcal{S})$ satisfies $v_A \leq \overline{T}^f_A v_A$, for some $f \in \mathbb{P}(\mathcal{A})$ then $v_A \leq v^f_A$, where v^f_A is the unique fixed point of $\overline{T}^f_A$ in $\mathcal{F}(\mathcal{S})$.

Proof. By hypothesis we have that

$$v_A \leq \overline{T}^f_A v_A \ ,$$

that implies by Theorem 1 c),

$$\overline{T}^f_A v_A \leq (\overline{T}^f_A)^2 v_A \ ,$$

and then

$$v_A \leq (\overline{T}^f_A)^2 v_A \ .$$

In the same way, for each n we have

$$v_A \leq (\overline{T}^f_A)^n v_A \ ,$$

and by Theorem 1 a), when $n \to \infty$,

$$(\overline{T}^f_A)^n v_A \longrightarrow v^f_A \ .$$

The result follows. □

Theorem 3. Suppose that Condition (14) holds. The sequence of functions u^n_A in Algorithm 3 verifies $u^n_A \uparrow v^*_A$. Further, if for any $n \in \mathbb{N}$, $u^n_A = u^{n+1}_A$, then it is true that $u^n_A = v^*_A$.

Proof. For each $s \in \mathcal{S}$, we have that

$$u^0_A(s) \ = \ \overline{T}^{f_0}_A(u^0_A)(s) \ \leq \ \overline{T}_A(u^0_A)(s) \ = \ \overline{T}^{f_1}_A(u^0_A)(s) \ .$$

Then the value function u^0_A satisfies

$$u^0_A \leq \overline{T}^{f_1}_A u^0_A \ ,$$

and by Lemma 3

$$u^0_A \leq v^{f_1}_A = u^1_A \ .$$

Iterating over n, we have that

$$u^n_A \leq \overline{T}_A(u^n_A) \leq u^{n+1}_A \ . \tag{25}$$

Now the sequence $\{u_A^n\}_{n\in\mathbb{N}}$ being non-decreasing and bounded by $||r_A||_\infty/(1-\beta_A)$, there exists a value function u_A such that for any $s \in \mathcal{S}$

$$u_A(s) = \lim_{n\to\infty} u_A^n(s) \ .$$

Taking $n \to \infty$ in (25), $u_A \leq T_A(u_A) \leq u_A$ and therefore $u_A = \overline{T}_A(u_A)$, and by uniqueness of the fixed point

$$u_A = v_A^* \ ,$$

and we have the first claim of the theorem: $u_A^n \uparrow v_A^*$. Also, if it is verified for some n that $u_A^n = u_A^{n+1}$, then, using (25),

$$u_A^{n+1} = u_A^n \leq \overline{T}_A u_A^n \leq u_A^{n+1} \ ,$$

which implies

$$u_A^n = \overline{T}_A u_A^n = v_A^* \ ,$$

where the second equality is again given by the uniqueness of the fixed point. The second claim follows. □

The results exposed in this section strongly rely on the fact that $\gamma_B(s, f, v)$ is independent on v_B. In the Section 3 we show that MFS is a sufficient condition for the existence of a FPE but all the results here may fail in the general case.

2.5 Mathematical Programming Formulations

In this section we develop the discussion of Mathematical Programming (MP) formulations, as the one proposed in [15]. 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 1.4, 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 induces 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:

$$0 \leq v_B(s) - (T_B^{fg} v_B)(s) \leq M_B(1 - g_{sb}) \qquad s \in \mathcal{S},\ b \in \mathcal{B}_s \qquad (26)$$

$$\sum_{b\in\mathcal{B}_s} g_{sb} = 1 \qquad s \in \mathcal{S} \qquad (27)$$

$$g_{sb} \in \{0, 1\} \qquad s \in \mathcal{S}, b \in \mathcal{B}_s. \qquad (28)$$

The variable of the program g_{sb} is meant to represent the probability $g(s, b)$.

For each $f \in W_A$, the deterministic best response set of the follower is determined by constraints (26)–(28). In (26) 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 [15] propose the following formulation:

$$
\begin{aligned}
\text{(MP)} \quad & \max \quad \sum_{s \in \mathcal{S}} \alpha_s v_A(s) && && (29)\\
\text{s.t.} \quad & \text{Constraints (26), (27), (28)} \\
& \sum_{a \in \mathcal{A}_s} f_{sa} = 1 && s \in \mathcal{S} && (30)\\
& v_A(s) - (T_A^{fg} v_A)(s) \le M_A(1 - g_{sb}) && s \in \mathcal{S},\, b \in \mathcal{B}_s && (31)\\
& f_{sa} \ge 0, && \forall s \in \mathcal{S},\, a \in \mathcal{A}_s. && (32)
\end{aligned}
$$

where $\alpha \in \mathbb{R}_+^{|\mathcal{S}|}$ is a non-negative vector of coefficients.

This Mathematical program is build using strongly the analogy with MDPs. In particular, it uses the reduction to a single objective using a vector of weights in (29). For MDPs, this choice is arbitrary. As it turns out in the experiments presented in Section 4, the result of (MP) sometimes depends on the vector α_s, and sometimes it is not a 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 seem to occur. We therefore conjecture that for the correctness of (MP), it is necessary to have monotonicity of the operator.

3 Existence Results for SSSE and FPE.

We gather in this section existence results about Stationary Strong Stackelberg Equilibria and Fixed-Points Equilibria of the dynamic programming operator. The general idea underlying these results is that, under certain assumptions, it can be proved that some operator, typically T or $\overline{T}_A$ defined in Section 2, is contractive. Using Banach's theorem, it then has a fixed point with which the solution is constructed. Then, still under some assumptions, this FPE solution is shown to be a SSSE.

3.1 Single-state results

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

$$V_i^{fg}(s_0) \;=\; \frac{r_i^{fg}(s_0)}{1-\beta_i}\;. \tag{33}$$

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 a 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.

We start with a general and useful result that applies to any game.

Lemma 4. 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 [14]. 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 the set $R_A(s)$ is nonempty. □

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

Proof. Let $\mathcal{S}$ have a single state: $\mathcal{S} = \{s_0\}$. The existence of SSSE is a particular case of Lemma 4. The existence of a FPE follows from the observation that the game is MFS: Theorem 1 applies to it. Being a contraction, the operator $\overline{T}_A$ has a unique fixed point from which a FPE is constructed. There remains to show that this FPE coincides with the SSSE.

To that end, we first show that $R_B(f) = R_B(s_0, f)$ for all $f \in W_A$ (since the game is with MFS, this latter set does not depend on $v_B \in \mathcal{F}(\mathcal{S})$). If $g \in R_B(s_0, f)$, 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 $g \in R_B(s_0, f)$. 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 (6) and (11) that $R_A(s_0) = R_A(s_0, v)$ for all v, which means that SSSE and FPE coincide. □

An alternative algorithmic proof of the existence of the SSSE in Theorem 4 is provided in Appendix B. This is based in the Multiple LPs algorithm which also give as a polynomial algorithm to solve a SSSE in the static case.

3.2 Myopic Follower Strategies

Theorem 5 (FPE for MFS)**.** If the game $\mathcal{G}$ is with MFS then it admits a 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 a 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 (15) and that of T in (12), combined with the MFS property, see the proof of Lemma 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 5, 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: $R_B(s, f, v_B) = R_B(s, f)$. It follows that the follower's best response has the form (14).

Leader-Controller Discounted Games: This case is a particular case of the Single-controller discounted game described in Filar and Vrieze [7], 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:

$$\begin{aligned} (T_B^{fg} v_B)(s) &= \sum_{a \in \mathcal{A}_s} f(s,a) \sum_{b \in \mathcal{B}_s} g(s,b) \left[r_B^{ab}(s) + \beta_B \sum_{z \in \mathcal{S}} Q^a(z|s) v_B(z) \right] \\ &= r_B^{fg}(s) + \beta_B \sum_{a \in \mathcal{A}_s} f(s,a) \sum_{z \in \mathcal{S}} Q^a(z|s) v_B(z) \ . \end{aligned}$$

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, (14) 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 5. Let $\mathcal{G}$ be a game with MFS. Then one 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:

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}_s, \forall b \in \mathcal{B}_s, \forall z \in \mathcal{S}, \quad \beta_B \left(Q^{ab}(z|s) - Q^{ab'}(z|s)\right) = 0 \ , \tag{34}$$

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 (34) 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 (34) 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 a 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 $R_B(f^*) = \prod_{s \in \mathcal{S}} R_B(s, f^*)$.

According to Lemma 5, 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 (13)):

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

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, $\prod_s R_B(s,f) \subset R_B(f)$. To see this, pick $g \in \prod_s R_B(s,f)$. 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, $R_B(f) \subset \prod_s R_B(s,f)$. To see this, pick $g \in R_B(f)$ (the set is not empty). If it is not in $\prod_s R_B(s,f)$, then there is some s and some $b \in R_B(s,f)$ 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, $\prod_s R_B(s,f) = R_B(f)$, 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 Markov Decision Problem. 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:

$$\begin{aligned} V_A^f(s) &= \max_{g \in R_B(f)} \{ r_A^{fg}(s) + \beta_A (Q^{fg} V_A^f)(s) \} \\ &= \max_{b \in R_B(s,f)} \{ r_A^{fb}(s) + \beta_A (Q^{fb} V_A^f)(s) \} \end{aligned} \tag{35}$$

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 (35) the operator $\overline{T}_A^f$ defined in (17). In other words, V_A^f is the fixed point of $\overline{T}_A^f$.

Let then 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 $s \in \mathcal{S}$,

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

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 Lemma 3, $\overline{T}_A U_A \geq U_A$ 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 a SSSE. □

3.3 Zero-Sum Games

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

Theorem 7. If the game $\mathcal{G}$ is a Zero-Sum Game, then it admits a FPE.

The existence of a FPE follows from the contractivity of the operator associated, in a similar way as in [6, 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 (14). It is then possible to define the operator $\overline{T}_A^f$ as in (16). 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 (8), 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 (17), 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. □

3.4 Team Games

A result of [15] (Proposition 1) is that Team Games have a SSSE. Team Games (also known as Identical-Goal Games in [14]) are such that both players seek to maximize the same metric. This is a property of reward functions only. We slightly generalize the definition of these games and state a similar result for FPE.

Definition 5 (Team Game)**.** The game is a Team Game if $\beta_A = \beta_B$ and there exists real constants μ and $\nu > 0$ such that: $r_B^{ab}(s) = \mu + \nu r_A^{ab}(s)$.

Theorem 8. If the game $\mathcal{G}$ is a Team Game, then it admits a FPE.

Proof. We adapt the proof of Theorem 7. Consider a function v in the set $\mathcal{W} = \{v \in \mathcal{F}(\mathcal{S}) \times \mathcal{F}(\mathcal{S}) | v_B = \frac{\mu}{1-\beta} + \nu v_A\}$. Since v_B can be expressed as a function of v_A, then

$SR_B(s, f, v)$ and $\gamma_B(s, f, v)$ can be made dependent only on v_A; in other words, it satisfies (14). It is then possible to define the operator $\overline{T}_A^f$ as in (16), mapping $\mathcal{F}(\mathcal{S})$ to $\mathcal{F}(\mathcal{S})$.

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 $g \in R_B(s, f, v_A)$. A straightforward calculation concludes that for every f, g, $T_B^{fg} v_B = \frac{\mu}{1-\beta} + \nu T_A^{fg} v_A$. The operator T^{fg} then maps $\mathcal{W}$ to $\mathcal{W}$ and so does $\overline{T}_A^f$. And by definition (8), 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 $(T_A^{fg} v_A)(s) \geq (T_A^{fh} v_A)(s)$ since $\nu > 0$. In other words,

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

so that

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

As it was the case in (17), the maximization set in the right-hand side does not depend on v_A. The proof concludes as in the proof of Theorem 7. □

3.5 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 6 (Acyclic Games)**.** The game is an Acyclic Game if the state space $\mathcal{S}$ admits the partition $\mathcal{S} = \mathcal{S}_\perp \cup \mathcal{S}_1$, 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 $(s, s') \in \mathcal{S}_1 \times \mathcal{S}_1$, 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 9. If the stochastic game $\mathcal{G}$ is an Acyclic Game, then it admits a 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 a FPE exist, 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 $(f, g) \in W_A \times W_B$,

$$V_i^{fg}(s_k) = r_i^{fg}(s_k) + \beta_i \sum_{\ell=0}^{k-1} \sum_{s' \in \mathcal{S}_\ell} Q^{fg}(s'|s) V_i^*(s') \ . \tag{36}$$

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: a FPE exist, resulting in a pair of strategies $(f^*_{s_k}, g^*_{s_k}) \in \mathbb{P}(\mathcal{A}_{s_k}) \times \mathbb{P}(\mathcal{B}_{s_k})$. When $k = K$, we have defined this way a strategy (f^*_s, g^*_s) for each $s \in S$.

We now prove that this strategy is a 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 $s_k \in \mathcal{S}_k$:

$$g^*(s_k) = \gamma_B(s_k, f^*, V^*)$$
$$f^*(s_k) = R_A(s_k, V^*)$$

where $V_i^* = V_i^{f^*g^*}$ is the unique fixed point of operator $T_i^{f^*g^*}$ for $i = A, B$. With Definition 3, the result will follow.

When $s_0 \in \mathcal{S}_0$, the local reaction set $LR_B(s_0, f, V^*)$ does not depend on V^* and $\{g^*_{s_0}\} \in LR_B(s_0, f^*, V^*)$, 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 (36). Then since $(f^*_{s_k}, g^*_{s_k})$ solves (locally) the SSE for the subgame defined by (36), 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 a FPE. □

In contrast with Theorem 4, the existence of SSSE is not guaranteed for acyclic games. In Section 4.3 we study a game without a SSSE. 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 MFS and it does have a SSSE according to Theorem 6.

4 Numerical Examples.

In this section, we present examples illustrating the different situations. We compare solutions returned by the different algorithms stated in Section 2.

In the example exposed in Section 4.2, VI converges to some FPE. Also we show that the FPE is a SSSE and the solution returned by (MP). The model involved does not satisfy the sufficient conditions for existence and convergence identified in Sections 2.3 and 3.

In Section 4.3, we describe an example where, depending on the discount factors β_A, β_B: either a SSSE exists and does not coincide with the FPE, or a SSSE does not exist, or a FPE does not exist.

Finally, in Section 4.4, we describe an example where a FPE is shown to exist, but VI does not necessarily converge to it.

4.1 Experimental setup

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. In order to solve (MP) we use the KNITRO solver provided by NEOS server[1].

4.2 Example 1

This first example is to illustrate the fast convergence of Value Iteration in a case where a FPE exists. Consider $\beta_A = \beta_B = 9/10$ and the data in Table 1.

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

State 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 1: Transition matrix and payoffs for each player.

In this example, Value Iteration, Policy Iteration, the SSSE and the optimal solution returned by (MP) coincide. The application of Value Iteration, starting with the null function, results in the evolution displayed in Figure 1. Given that Value Iteration converges, a FPE exists. The policies and values are given in Table 2. It is checked that these values also form a SSSE. Details are provided in Appendix C.1.

[1] https://neos-server.org/neos/.

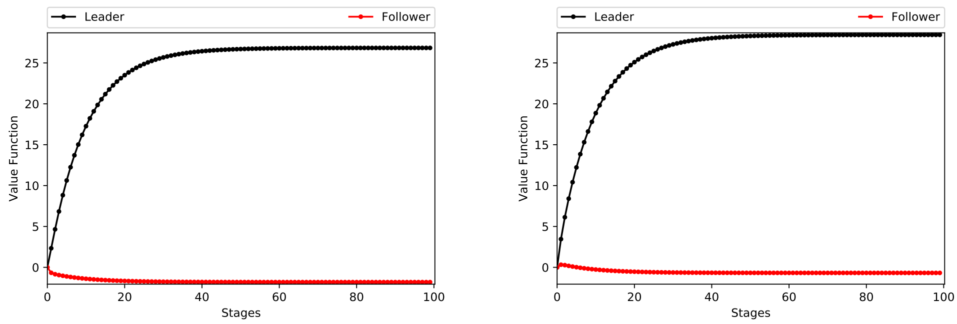


Figure 1: Value Iteration applied to Example 1

	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 2: Policies and values of the SSSE in Example 1

4.3 Example 2

In this section we will study the stochastic game given by the data in Table 3. This game has two states and two actions per state. Figure 2 shows a diagram of the possible transition between states. State s_1 is absorbing for any combination of actions.

State s_1

	b_1	b_2
a_1	(1, 0) / $(-1, -2)$	(1, 0) / $(-2, 1)$
a_2	(1, 0) / $(0, 0)$	(1, 0) / $(2, 0)$

State 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 3: Transition matrix and payoffs for each player in Example 2

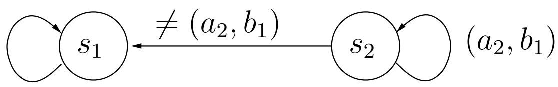


Figure 2: Transition structure of Example 2

Computation of values. For the sake of clarity, we will denote with f_i and g_i the probability of playing actions a_1 and b_1 in state s_i. Given that s_1 is absorbing, the value of both players in that state only depend on f_1 and g_1. For Player A this value is given by:

$$\begin{aligned} v_A(s_1) =& f_1 g_1[-1+\beta_A v_A(s_1)] + f_1(1-g_1)[-2+\beta_A v_A(s_1)] \\ &+ (1-f_1)g_1[0+\beta_A v_A(s_1)] + (1-f_1)(1-g_1)[2+\beta_A v_A(s_1)] \\ =& \frac{1}{1-\beta_A}(3f_1g_1 + 2 - 4f_1 - 2g_1) \end{aligned}$$

and analogously,

$$v_B(s_1) = \frac{f_1}{1-\beta_B}(1-3g_1). \tag{37}$$

In state s_2, the value functions depends on both the actions performed in s_1 and s_2. These values are given by

$$v_A(s_2) = \frac{1}{1-(1-f_2)g_2\beta_A}[2f_2g_2 - 3f_2 - g_2 + 1 + (1-(1-f_2)g_2)\beta_A v_A(s_1)] \tag{38}$$

$$v_B(s_2) = \frac{1}{1-(1-f_2)g_2\beta_B}[-3f_2 + 1 + (1-(1-f_2)g_2)\beta_B v_B(s_1)]. \tag{39}$$

In the following, we will discuss briefly that depending on the values of β_A and β_B this game does not have a FPE, a SSE, or having both, they do not coincide. We summarize these results in Table 4. The details of this analysis are provided in Appendix C.2.

β_A	SSSE	FPE
$[0, \frac{1}{5})$	No	Yes
$(\frac{1}{5}, \frac{1}{3})$	Yes	Yes
$(\frac{1}{3}, 1)$	Yes	No

Table 4: Existence of FPE and SSSE for different values of β_A and $\beta_B > 0$. Even when both exist they may not coincide.

No SSSE exists. The non-existence of SSSE for $\beta_A < \frac{1}{5}$ and $\beta_B > 0$ comes from the fact that $R_A(s_1) \cap R_A(s_2) = \emptyset$ in this case. The leader plays different strategies if she starts in state s_1 than if she starts in s_2. We explain briefly the case for $\beta_A = 0$.

Note that if the game starts in s_1, then the optimal policy for the leader, will be the same than in the static case, that is a game with only state s_1. The set of best strategies for the leader is to play $f_1 = 0$ (play a_2) in state s_1 and any arbitrary $f_2 \in [0, 1]$. The follower reacts with b_2 in state s_1 and an arbitrary response in s_2. In this case, the value functions are $v_A(s_1) = 2$ and $v_B(s_1) = 0$.

If the game starts in s_2, the value of Player A is $v_A(s_2) = 2f_2g_2 - 3f_2 - g_2 + 1$. It is maximized when $f_2 = g_2 = 0$, that is, when (a_2, b_2) is played. If Player A performs a_2, Player B has immediate rewards equal to 1 and she has to arbitrate between staying forever in state s_2, and receive an additional $\frac{\beta_B}{1-\beta_B}$, or jump to state s_1 and gain $\beta_B v_B(s_1)$. Player A prefers this second situation, and she has to incentivise Player B to play b_2.

From (37) and (39), and assuming $f_2 = 0$, Player B's value is:

$$v_B(s_2) = \frac{1 - f_1 + 3f_1g_1}{1 - g_2\beta_B} + f_1\frac{1 - 3g_1}{1 - \beta_B} \ .$$

Note that if $f_1 = 0$, this value becomes $\frac{1}{1-g_2\beta_B}$, and Player B maximises it when $g_2 = 1$, that is by playing b_1. The leader then gains 0. On the other hand, if $f_1 = 1$, this value becomes:

$$v_B(s_2) = \frac{1}{1 - \beta_B} + \beta_B\frac{3g_1(g_2 - 1)}{(1 - \beta_B)(1 - g_2\beta_B)} \ .$$

This value is maximised when $g_1 = 0$, $g_2 = 1$ or both. This tie is broken when $g_2 = 0$, giving the leader a value of 1.

In summary, when $\beta_A = 0$, $R_A(s_1) = \{a_1\} \times \{(f_2, 1 - f_2) : f_2 \in [0, 1]\}$ and $R_A(s_2) = \{a_2\} \times \{a_2\}$. Clearly, $R_A(s_1) \cap R_A(s_2) = \emptyset$, and therefore there is no SSSE. Note that the previous analysis does not hold anymore if $\beta_B = 0$. In this case, both players play, in each state, the SSSE of the static case and an arbitrary policy in the other state.

Observe here in passing that the "myopic" case does not always bring a simplification to the problem. Despite the fact that the myopic follower condition guarantees the existence of FPE and SSSE, the myopic leader condition does not guarantee the existence of SSSE.

No FPE exists. First we note that there is only one FPE for the operator associated to s_1. This fixed point is given by the SSE of the static game played in this state. These values are:

$$v_A(s_1) = \frac{2}{1 - \beta_A} \quad v_B(s_1) = 0. \tag{40}$$

Now we only focus in the sub-operator that is applied in state s_2. In order to clarify formulas, we use the symbols $w := v_A(s_2)$ and $z := v_B(s_2)$. The images $(Tv)_A(s_2)$ and $(Tv)_B(s_2)$ are denoted with w' and z' respectively. By using the values in (40), we arrive at a single-state game parametrised by the values of w and z:

	b_1	b_2
a_1	$\frac{3\beta_A-1}{1-\beta_A}$, -2	$\frac{4\beta_A-2}{1-\beta_A}$, -2
a_2	$\beta_A w$, $1 + \beta_B z$	$\frac{1+\beta_A}{1-\beta_A}$, 1

Then the sub-operator for this game is as follows:

$$(w',\, z') = \begin{cases} \left(\frac{1+\beta_A}{1-\beta_A},\, 1\right) & \text{if } z < 0 \\ \left(\frac{3\beta_A-1}{1-\beta_A}, -2\right) & \text{if } w < \frac{3\beta_A-1}{\beta_A(1-\beta_A)} \text{ and } z > 0 \\ T_{21}(w,z) = (\beta_A w, 1+\beta_B z) & \text{if } w > \frac{3\beta_A-1}{\beta_A(1-\beta_A)} \text{ and } z > 0 \\ (\beta_A w, 1-3f_2+(1-f_2)\beta_B z) & \text{if } w = \frac{3\beta_A-1}{\beta_A(1-\beta_A)} \text{ and } z > 0 \\ (\beta_A w, 1) & \text{if } w > \frac{1+\beta_A}{\beta_A(1-\beta_A)} \text{ and } z = 0 \\ \left(\frac{1+\beta_A}{1-\beta_A}, 1\right) & \text{if } w \leq \frac{1+\beta_A}{\beta_A(1-\beta_A)} \text{ and } z = 0. \end{cases} \tag{41}$$

We show that if $\beta_A > \frac{1}{3}$, then there is no FPE by explaining the dynamics of this operator.

- The iterations starting with $z < 0$ will be mapped in one iteration to the point $(\frac{1+\beta_A}{1-\beta_A}, 1)$. Then it will be mapped via the linear contractive operator T_{21}. This operator has a unique fixed point given by $(0, \frac{1}{1-\beta_B})$. Then, the sequence $T_{21}^k(\frac{1+\beta_A}{1-\beta_A}, 1)$ will enter to the zone where $w < \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$. In one more iteration, the operator will arrive at the point $(\frac{3\beta_A-1}{1-\beta_A}, -2)$, and there is a cycle.

- Iterations of T starting in $w < \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$, will map to $(\frac{3\beta_A-1}{1-\beta_A}, -2)$, and continue as above.

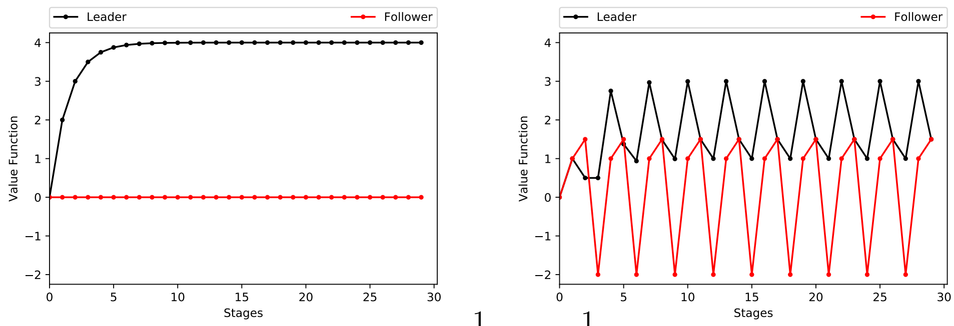
- Iterations starting with $w < \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$ will follow a sequence of points in the direction of the fixed point of the unique fixed point of T_{21} and then enter the zone where $w < \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$ and continue as above.

- If the iterations start with (w, z) such that $w = \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$, depending on the value of f_2, the mapping will lead to one of the two zones mentioned above. Then, the iterations will cycle.

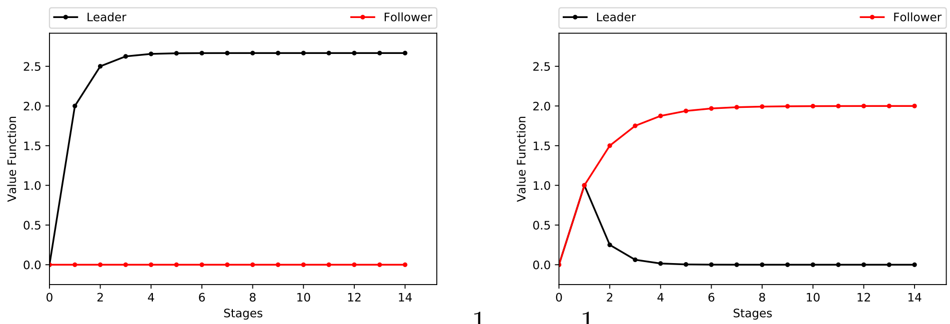
- In all cases, there is a limit cycle. The length of this cycle can range from 3 to ∞ depending on the proximity of β_A to $\frac{1}{3}$.

In the case where $\beta_A \leq \frac{1}{3}$, the fixed point of T_{21} is located in the zone where $w > \frac{3\beta_A-1}{\beta_A(1-\beta_A)}$ and $z > 0$, and then it is the only FPE of this game. The dynamics of Value Iteration for both cases are represented in Figure 3.

SSSE and FPE exist, but they do not coincide. For $\beta_A > \frac{1}{5}$, the values of the SSSE are given by $(v_A(s_1), v_A(s_2)) = (\frac{2}{1-\beta_A}, \frac{2\beta_A}{1-\beta_A})$ and $(v_B(s_1), v_B(s_2)) = (0, 0)$. On the other hand, for $\beta_A < \frac{1}{3}$ the FPE identified above is given by $(v_A(s_1), v_A(s_2)) = (\frac{2}{1-\beta_A}, 0)$ and $(v_B(s_1), v_B(s_2)) = (0, \frac{1}{1-\beta_B})$. Thus, when $\beta_A \in (\frac{1}{5}, \frac{1}{3})$, there exist both a SSSE and a FPE, and they do not coincide.



$$\beta_A = \frac{1}{2}, \beta_B = \frac{1}{2}$$



$$\beta_A = \frac{1}{4}, \beta_B = \frac{1}{2}$$

Figure 3: Value Iteration applied to Example 2: state s_1 (left) and s_2 (right)

Policy iteration and (MP). We test the Mathematical Programming (MP) 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 5 summarizes the results obtained.

		(MP)						(PI)	
β_A		$\alpha_{s_1} = 100$	$\alpha_{s_2} = 1$	$\alpha_{s_1} = 1$	$\alpha_{s_2} = 100$	$\alpha_{s_1} = 1$	$\alpha_{s_2} = 1$	s_2	s_1
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	2	2	~ 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 5: Results for (MP) and (PI) with different values of α and β_A with $\beta_B = 0.5$ fixed

Whenever the SSSE exists, (MP) computes it correctly. When no SSSE exist, (MP) returns the value function of the game that is started in the state s with greater weight α_s. Policy iteration finds a SSSE in none of the cases.

4.4 Example 3: A FPE exists but VI does not converge to it

We now develop an example where a FPE does exist, but Value Iteration does not necessarily converge to it. The data of this example is listed in Table 6.

	b_1	b_2
a_1	(1, 0) / (1, −1)	(0, 1) / (0, 1)
a_2	(0, 1) / (−1, 1)	(0, 1) / (−1, −1)

State s_1

	b_1	b_2
a_1	(0, 1) / (−1, 0)	(1, 0) / (0, 1)
a_2	(1, 0) / (0, 1)	(0, 1) / (1, −1)

State s_2

Table 6: Transition matrix and payoffs for each player in Example 3

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 7 constitute *both* a SSSE and a FPE. We provide in Appendix C.3 justifications for this claim.

	s_1	s_2
Play of A	$(1, 0)$	$(5-\sqrt{19}, -4+\sqrt{19})$
Play of B	b_2	b_2
v_A	$\frac{-3+\sqrt{19}}{5} \sim 0.27178$	$\frac{-6+2\sqrt{19}}{5} \sim 0.54356$
v_B	$\frac{16-2\sqrt{19}}{5} \sim 1.45644$	$\frac{22-4\sqrt{19}}{5} \sim 0.91288$

Table 7: Values and Policies forming a SSSE and FPE.

When applying Value Iteration with the null function as a starting point, we get however the evolution in Figure 4. Values obtained with Policy Iteration have a similar behavior. Finally, (MP) returned as the optimal solution the SSSE (and FPE).

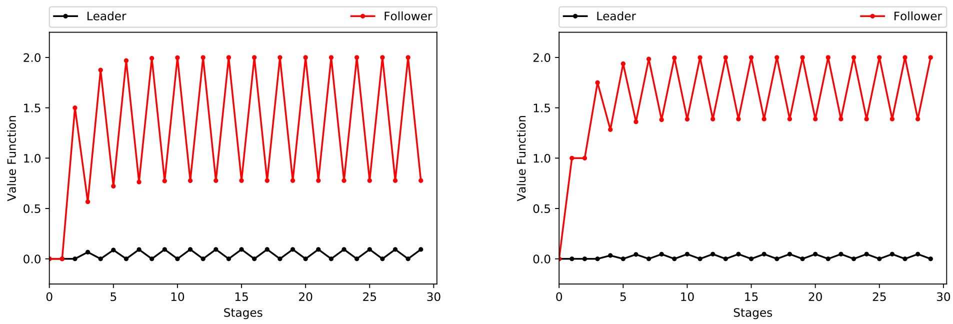


Figure 4: Value Iteration applied to Example 3: state s_1 (left) and s_2 (right)

5 Application: Surveillance in a graph.

In this section we present an example of a stochastic game modeling a security patrol. This example shows an application of the models and algorithms presented in this paper. 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 maximising their expected rewards.

The leader knowing the position of the attacker perform his policy. Then, the attacker observing the position of the defender, decides two things: where to move and if he performs an attack or not. Once the attack is performed the game ends in a set of terminal states. This game model situations where the defender has the information where the possible attacker it is located, but they cannot perform any action if an attack is not performed. In real situations, it can be the case of demonstrations or high-risk football matches.

The rewards mainly depend on the place where the attack is performed and whether the location is being covered or not. The effectiveness of the player's movements is influenced by random factors: when they decide to move to some location this move may fail due to external factors, in which case they remain in their current location.

5.1 Game description

We introduce now the elements of the model and the notation. A synthesis of this notation is presented in Appendix A. Formally, we consider a set of locations to patrol/targets $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_n\}$. There are some connections allowed between locations represented by edges (denoted by $\mathcal{E}$), so the board of the game is actually a graph $(\mathcal{L}, \mathcal{E})$. 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 ($\ell_A \in \mathcal{L}$) and the attacker's location ($\ell_B \in \mathcal{L}$). The binary parameter α takes the value 1 if the attacker is committing an attack or 0 if he is not, thereby being unnoticed in that period. There are also two special fictitious states $\perp_0$, $\perp_1$ representing the state of the game once the attack was performed. $\perp_1$ represents the case where the attack is successful and $\perp_0$ when the attacker is caught. These two are absorbing states in our game.

The action space $\mathcal{A}_s \subset \mathcal{L}$ for the leader represents all the possible location that he can achieve from its current position (given by the state s). For the follower, $\mathcal{B}_s \subset \mathcal{L} \times \{0, 1\}$ represents all the possible locations that the attacker can achieve in the state s with and the decision whether to attack or stay unnoticed. We use the notation $\ell \in \mathcal{L}$ to represent the action of "*move to ℓ*" and $\alpha \in \{0, 1\}$ to represent the action of *"attacking"* or *"stay unnoticed"* respectively. In states $s \in \{\perp_0, \perp_1\}$, actions are irrelevant: we can pick $\mathcal{A}_s = \mathcal{L}$ and $\mathcal{B}_s = \mathcal{L} \times \{0, 1\}$ for such states by convention.

In order to represent the transitions of the states, we define for all possible locations the function $q_i^{\ell'}(\ell''|\ell)$ as the probability of player player $i \in \{A, B\}$ reaching location ℓ'' from ℓ given that he decides to move to ℓ'. In particular, if there is no possible failure in the movements $q_i^{\ell'}(\ell''|\ell) = 1$ if $\ell'' = \ell'$, and 0 otherwise. This will be a data of the problem. We will assume independence of the transitions of both players. The transition probabilities $Q^{ab}(z|s)$ are defined by expressions (42)–(44) as follows:

$$Q^{ab}(z|\perp_i) = \begin{cases} 1 & z = \perp_i, \quad i \in \{0, 1\}, (a, b) \in \mathcal{A}_{\perp_i} \times \mathcal{B}_{\perp_i} \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

$$Q^{\ell'_A,(\ell'_B,\alpha)}(\ell''_A, (\ell''_B, \alpha)|\ell_A, (\ell_B, 0))) = \begin{matrix} q_A^{\ell'_A}(\ell''_A|\ell_A)\; q_B^{\ell'_B}(\ell''_B|\ell_B) & \alpha \in \{0, 1\}, \ell'_A \in \mathcal{A}_{\ell_A}, \\ & (\ell'_B, \alpha) \in \mathcal{B}_{\ell_B}, \end{matrix} \tag{43}$$

$$Q^{\ell'_A,(\ell'_B,\alpha)}(z|\ell_A, (\ell_B, 1))) = \begin{cases} 1 & z = \perp_0 \text{ and } \ell_A = \ell_B \\ 1 & z = \perp_1 \text{ and } \ell_A \neq \ell_B \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

Rewards result from the interaction, or lack thereof, between the defender and the attacker. Following the notation used in security games, we denote $U_A^u(\ell) < 0$ and $U_A^c(\ell) > 0$ the penalty and the benefit for the defender if an attack is performed in ℓ, which depends only if the target is uncovered (superscript u) or covered (superscript c).

Similarly, we define $U_B^u(\ell) > 0$ and $U_B^c(\ell) < 0$ the reward (and penalty respectively) of the attacker if the location attacked is uncovered or not. 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, which depends on the current state of the system s, the actions (a, b) performed by the players and the future state of the system z. This technique is fairly standard, as it is shown in [12, Ch. 2, pp.20]. The expressions for $R_i^{ab}(z|s)$, $i = A, B$ are listed in (45), in which s is any state and $z = (\ell_A, (\ell_B, \alpha))$.

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

(45)

The first two lines represent the payoffs when the attack is performed, $z = (\ell_A, (\ell_B, 1))$. In the third line, $P_B(\ell_B) < 0$ represents the opportunity cost and risk for the attacker of being in location ℓ_B and not perform an attack. In the two last lines P_A and P_B represents the residual value 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. From the definition of R_A and R_B we obtain instants rewards r_A and r_B as follows:

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

The dynamics of the game is summarized as follow: First, at the start of any epoch, the system is in a state formed by the location the both players and the behaviour of the attacker. Then, the defender knowing the state of the game chooses a strategy f (probably mixed) over the locations reachable from his current location. The attacker observes the strategy and chooses where to move and whether to attack or not. We denote this action as g. Note, that if the attacker decide to attack, the success or failure of his strategy 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.

In the next section we test the algorithms proposed in this work. First, we describe the instances we test. In order to test the efficiency of Value Iteration and Policy Iteration for the myopic case, that is $\beta_B = 0$, we will measure the running time to find the FPE - SSSE. We will compare the solutions obtained, with policies obtained heuristically. Finally, we evaluate the solution returned by the FPE in the case with non myopic followers, in the case the algorithm can detect them.

5.2 Computational study

In order to test the performance of the algorithms proposed in this work we generate instances of different structure and size. In particular we test the model on paths, cycles, T-shaped graphs and complete graphs. We limit the size of $\mathcal{A}_s$ and $\mathcal{B}_s$ by limiting

n	k	type	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{B}\rvert$
5	2	Cycle	52	4,846	9,653
		Line	52	3,692	7,346
		T	52	4,077	8,115
		Complete	52	4,846	9,654
	3	Cycle	52	4,846	9,654
		Line	52	4,462	8,885
		T	52	4,846	9,654
		Complete	52	4,846	9.654

n	k	type	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{B}\rvert$
10	2	Cycle	202	4,960	9,911
		Line	202	4,366	9,911
		T	202	4,564	8,723
		Complete	202	9,911	9,119
	3	Cycle	202	6,941	13,871
		Line	202	5,752	11,495
		T	202	6,347	12,683
		Complete	202	9,911	19,812

Table 9: Size of the instances by changing the graph structure, the amount of nodes n and the limit in the displacement

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 $q_A^{\ell'}(\ell''|\ell)$ in function of the nodes that are in the shortest path between ℓ and ℓ''. We denote this set of nodes as $SP(\ell, \ell'')$. Probabilities q_A are defined as follows:

$$q_A^{\ell'}(\ell''|\ell) = \begin{cases} 1-\epsilon & \ell' = \ell'' \\ \dfrac{\epsilon}{|SP(\ell,\ell')|-1} & \ell'' \in SP(\ell,\ell')\backslash\{\ell'\} \\ 0 & \text{otherwise.} \end{cases} \tag{46}$$

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 8. 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$deg(\ell)$
$U_B^u(\ell)$	$2^{deg(\ell)}$	$U_A^c(\ell)$	$-2^{deg(\ell)}$
$P_A(\ell_B)$	0	$P_A(\ell_B)$	$deg(\ell)$
$P_A(\bot_0)$	1	$P_A(\bot_1)$	-1
$P_B(\bot_0)$	-1	$P_B(\bot_1)$	1

Table 8: Payoff functions description

We test our models in instances with $n \in \{5, 10, 15\}$ and $k \in \{2, 3\}$ for each type of graph. In order to understand how the instances increases with n and k, in Table 9, the size of the state set and the average size of the set of actions for each type of instance. The precision parameter for the stopping criterion of iterative algorithm is set to $\varepsilon = 10^{-4}$ for both Value Iteration and Policy Iteration.

The experimental setup is the one described in Section 4.1. For the sizes of instances that we tested, the solver for (MP) runs out of memory, so we do not include this method in our analysis. Figure 5 shows the solution times in a performance profile in logarithmic scale comparing Value Iteration and Policy Iteration. Policy iteration has better performance for all the instances tested. On the other hand, we measure in a disaggregated fashion the computational times. Figure 6 we shows also the time spent in create the instances. This preprocessing time measures the time required tor building the stochastic game from a given graph. The results show that Value Iteration in general takes more time to be performed independently of the graph structure, while Policy Iteration performs faster in the complete graph. Also, that the preprocessing time is comparable to the time to perform Policy Iteration in some instances.

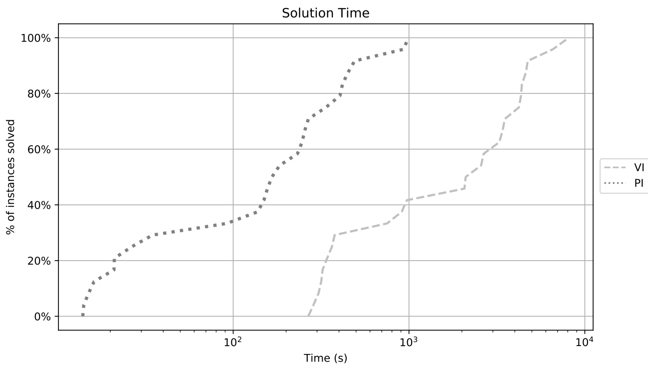


Figure 5: Solution times aggregated.

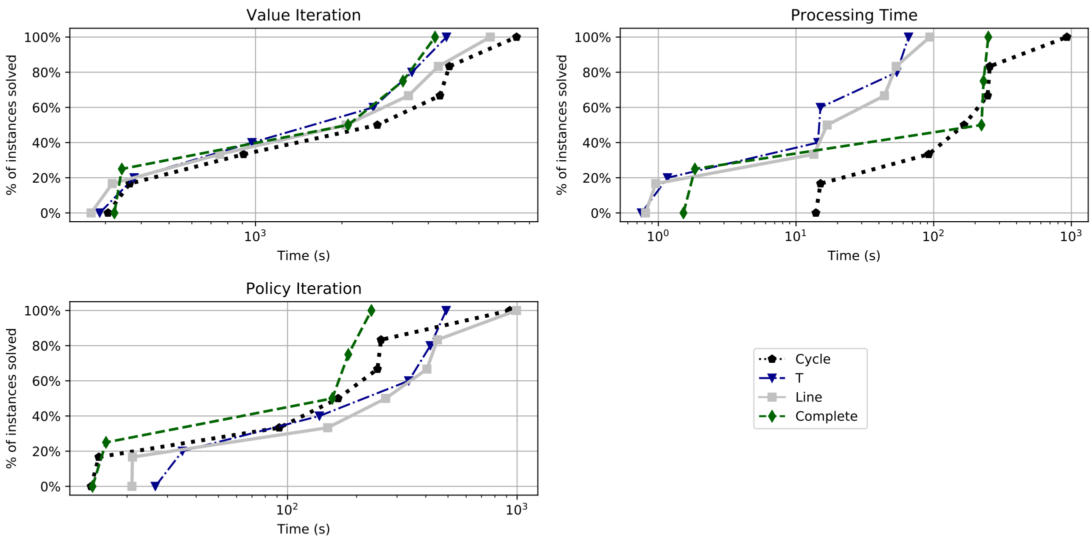


Figure 6: Solution times disaggregated.

In order to validate the model, we wish to compare the policies generated by this solution concept and the ones that are generated in a static fashion, that is, without considering the dynamics of the model. We will call this second heuristic type of policies Myopic policies.

However, we do not have a method for computing the SSSE of the game in general, and if we do, the computation time is prohibitive. So we use the FPE as a proxy that can be computed quickly. Accordingly, we compute the values of equilibrium $v^* = (v_A^*, v_B^*)$, with the respective equilibrium policies f^* and g^*.

As for myopic policies, 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 (13)) 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 10 shows the comparison of the values for the different types of graph structures mentioned before, and with the parameters $n = 10$, $k = 2$. We use Algorithm 2, Value Iteration, with $\beta_A = 0.9$ and $\beta_B = 0$. Remind that in this case, this algorithm converges to a 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 rest of the 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 10: Evaluation of the solution concept with $\beta_B = 0$.

We have repeated the same evaluation, but now with $\beta_B = 0.9$. Note that in this case there are no guarantees that Value Iteration will converge, neither that the FPE policy is a SSSE. Value Iteration found a FPE (i.e. converged) in all the instances. Table 11 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 11: 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 is because the change of the behaviour of the follower, but also because the FPE is not a SSSE anymore. The Myopic policy outperforms the FPE. Only in the case of the graph with the highest density, the FPE found is competitive with respect to the heuristic solution.

5.3 Conclusion

Computational results shows, when Policy Iteration converges to a FPE, this can be achieved by Policy Iteration in a more efficient way than Value Iteration. The Mathematical Programming formulation does not scale up to the size of instances considered in this experiment, therefore they were not included.

On the other hand, FPE performs very well when they coincide with the SSSE. In other cases, it can perform poorly. For these cases, heuristic myopic policies outperform in average value to the ones obtained by the FPE.

6 Conclusions and Further Work

In this paper, we have demonstrated 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 have first defined a suitable operator acting on the set of value functions for both players. We have introduced the concept of Fixed-Point Equilibrium, FPE, as the fixed points of this operator. We have then investigated the relationship between SSSE and FPE. We have shown 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, the class of games with Myopic Follower Strategies, MFS, is particularly promising for applications in the domain of security. We have shown that Value Iteration and Policy Iteration algorithms are able to approximate SSSE for such games, with a good algorithmic complexity. We have tested these algorithms on a model of mobile attackers and defenders on a transportation graph and shown that the SSSE, computed as a FPE in the case where the game is MFS, outperforms the naive myopic policy.

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. Finally, 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.

Acknowledgments.

Victor Bucarey has been partially supported by the Fonds de la Recherche Scientifique -FNRS under Grant(s) no PDR T0098.18.

A Notation Summary

We list in the following tables the principal notation used in the paper. Table 12 groups the notation related to the general model. Table 13 is relative to the application of Section 5.

Stochastic Games	
$\mathcal{G}$	Stochastic game.
$\mathcal{S}$	Set of states.
$\mathcal{A}_s$, $\mathcal{B}_s$	Set of actions for the Leader and the Follower respectively available in state s.
$r_i^{ab}(s)$	Immediate reward for Player i when actions a and b are performed in state s.
$Q^{ab}(z\|s)$	Probability of reaching state z from state s when actions a and b are performed.
β_i	Discount factor for Player i.
W_i	Set of feedback policies for Player i.
f, g	Policies for the Leader (Player A) and the Follower (Player B) respectively.
$V_i^{fg}(s)$	Expected discounted sum of all rewards for Player i, when policies f and g is applied and the starting state is s.
R_B, SR_B	Set of optimal responses (and strong respectively) for the Follower.
γ_B	Best response which is performed for the Follower.
$R_A(s)$	Best strategies for the Leader starting in state s.
v_i	Value function for Player i.
$\mathcal{F}(\mathcal{S})$	Set of value functions $\mathcal{S} \to \mathbb{R}$.
T_i^{fg}	One-step operator for any fixed pair of policies f and g.
T	Dynamic programming operator.
$\overline{\gamma}_B$	Best response of the Follower when it does not depend on v_B.
$\overline{T}_A^f$, $\overline{T}$	One-step operator and Dynamic programming operator when T does not depend on v_B.
M_i	Upper bound in value functions for Player i.

Table 12: Table of general notation

Application: Surveillance in a graph	
$\mathcal{L}$, $\mathcal{E}$	Set of locations and conecctions bewteen locations.
$\perp_0$, $\perp_1$	Absorbing states of the game.
α	Binary decision of attack or not.
$U_i^u(\ell), U_i^c(\ell)$	Reward or penalty for player i when an attack is performed in location ℓ and it is being protected (c) or unprotected (u).
P_i	Rewards and costs for player i when an attack is not performed.
$q_i^{\ell'}(\ell''\|\ell)$	Probability for player i of reaching location ℓ'' from ℓ given that he decides to move to ℓ'.
n	Number of locations in the graph.
k	Maximum geodesic distance which the defender can move in each step time.
$\overline{v}^*$	Average value function when a the policy obtained via Value Iteration is performed.
$\overline{v}^M$	Average value function when a myopic policy is performed.

Table 13: Table of notation for the surveillance application

B Algorithmic Proof for the Static Case.

We provide an algorithmic proof for the existence of SSSE in Theorem 4. This proof is also useful because it allows to compute a SSE in the static case, which is an important step in Algorithms 1, 2 and 4.

Lemma 6. If the game $\mathcal{G}$ is static (i.e. has only one state), then it has a SSSE.

Proof. Proof. Since the state space has only one state, we omit the reference to states in the notation. For each action $b \in \mathcal{B}$, define the following problem LP(b):

$$\text{LP}(b) \quad \max \sum_{a\in\mathcal{A}} r_A^{ab} f(a) \tag{47}$$

$$s.t. \sum_{a\in\mathcal{A}} r_B^{ab} f(a) \geq \sum_{a\in\mathcal{A}} r_B^{ab'} f(a) \qquad \forall b' \in \mathcal{B} \tag{48}$$

$$\sum_{a\in\mathcal{A}} f(a) = 1 \tag{49}$$

$$f(a) \geq 0 \qquad \forall a \in \mathcal{A}. \tag{50}$$

Consider Algorithm 5. Each of these LPs is bounded. Therefore, they are either unfeasible or they have a finite optimal solution. Furthermore, at least one of them is feasible. In that case, the optimal value represents the maximum value achieved by incentivising to the follower to play b. By picking the maximum of these values, the optimal f of that LP with the action b is an SSE for the game. □

Algorithm 5 Multiple LPs Algorithm

Initialize with $n = 0$, $v_A^0(s) = 0$ for every $s \in \mathcal{S}$
for $b \in \mathcal{B}$ **do**
    Solve LP(b)
    if LP(b) is bounded **then**
        Retrieve $f(b)$ and $v^*(b)$, respectively the leader's optimal policy given that the follower plays b, and its value
    end if
end for
Set the policy g^*, in which the follower plays $b^* = \arg\max v^*(b)$
return SSE $(f^*(b^*), g^*)$ for the static game

Given that Algorithm 5 solves $|\mathcal{B}|$ LPs with $|\mathcal{A}|$ variables and $|\mathcal{B}|$ constraints, it finds a SSSE in polynomial time.

C Analysis of the examples

In this section, we provide details and justifications for the examples reported in Section 4. In doing so, we introduce some elements useful in the analysis of the operators involved in Value Iteration algorithms.

C.1 Analysis of Example 1

In this section, we provide details on the analysis of Example 1 introduced in Section 4.2. We compute the Stationary Strong Stackelberg Equilibria (SSSE) and demonstrate their existence. We compute the one-step operator and show it has also a fixed point (FPE) which coincides with this SSSE.

C.1.1 Data

The data of this example is given in Table 14 (also Table 1). We will also specifically consider $\beta_A = \beta_B = 9/10$.

State s_1:

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

State 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 14: Transition matrix and payoffs for each player in Example 1

C.1.2 Computation of the SSSE

Values of stationary strategies. Given a stationary strategy (f,g), let us compute the value V_i^{fg} of this strategy for each player $i = A, B$. The notation is $f_i = f(s_i, a_1)$ and $g_i = g(s_i, b_1)$. We have first, observing that transition probabilities do not depend on the state,

$$Q^{fg} \;=\; \begin{pmatrix} \frac{5}{4}f_1g_1 - f_1 - \frac{3}{4}g_1 + 1 & -\frac{5}{4}f_1g_1 + f_1 + \frac{3}{4}g_1 \\ \frac{5}{4}f_2g_2 - f_2 - \frac{3}{4}g_2 + 1 & -\frac{5}{4}f_2g_2 + f_2 + \frac{3}{4}g_2 \end{pmatrix} .$$

Also,

$$r_{AB}^{fg} \;=\; \begin{bmatrix} (29\,g_1 - 11)\,f_1 - 14\,g_1 + 6 & (-24\,g_1 + 10)\,f_1 + 8\,g_1 - 4 \\ (13\,g_2 - 3)\,f_2 - 5\,g2 + 2 & (-31\,g_2 + 16)\,f_2 + 20\,g_2 - 10 \end{bmatrix} .$$

From these elements, one derives the values for Player B in each state, when she plays the four possible strategies: with $V_B^{f,xy}(s)$ the value at state s when x is played in state s_1 and y played in state s_2:

$$V_B^{f,b_1b_1}(s_1) = 10\,\frac{9\,f_1f_2 - 272\,f_1 - 369\,f_2 + 322}{-9\,f_1 + 9\,f_2 + 40} \qquad V_B^{f,b_1b_1}(s_2) = 10\,\frac{9\,f_1f_2 - 216\,f_1 - 429\,f_2 + 346}{-9\,f_1 + 9\,f_2 + 40}$$

$$V_B^{f,b_1b_2}(s_1) = 20\,\frac{180\,f_1f_2 - 235\,f_1 + 144\,f_2 - 55}{-9\,f_1 - 36\,f_2 + 67} \qquad V_B^{f,b_1b_2}(s_2) = 20\,\frac{180\,f_1f_2 - 207\,f_1 + 176\,f_2 - 83}{-9\,f_1 - 36\,f_2 + 67}$$

$$V_B^{f,b_2b_1}(s_1) = 20\,\frac{225\,f_1f_2 - 245\,f_1 + 18\,f_2 + 26}{-36\,f_1 - 9\,f_2 - 13} \qquad V_B^{f,b_2b_1}(s_2) = 20\,\frac{225\,f_1f_2 - 225\,f_1 + 48\,f_2 - 2}{-36\,f_1 - 9\,f_2 - 13}$$

$$V_B^{f,b_2b_2}(s_1) = 20\,\frac{27\,f_1f_2 + 5\,f_1 + 18\,f_2 - 20}{9\,f_1 - 9\,f_2 + 10} \qquad V_B^{f,b_2b_2}(s_2) = 20\,\frac{27\,f_1f_2 + 26\,f_2 - 23}{9\,f_1 - 9\,f_2 + 10}.$$

Analysis of values. The objective is to compute, for every state, the best strategy $f \in W_A$ for Player A: the set $R_A(s)$. A SSSE will exist if and only if the intersection of these sets is nonempty. First of all, we identify the sets $R_B(f)$, that is, the solutions to Player B's MDP problem. This is done by calculating the "zones" where some policy $g = (b_i, b_j) \in W_B$ is optimal, formally defined as $\mathcal{Z}_{ij} = \{f \in W_A, (b_i, b_j) \in R_B(f)\}$.

We successively compute the differences $V_B^{f,b_1b_1}(s) - V_B^{f,b_1b_2}(s)$, $V_B^{f,b_1b_2}(s) - V_B^{f,b_2b_2}(s)$, etc. and we identify four critical lines separating the four zones $\mathcal{Z}_{ij}$ $i,j = 1,2$, as represented in Figure 7.

C.1.3 Optimization for Player A

According to the preferences of Player B, we have identified in the previous section a covering of W_A in four sets $\mathcal{Z}_{ij}$, on which we proceed to find Player A's maximum (this is a covering and not a partition, since the sets $\mathcal{Z}_{ij}$ are not disjoint).

The plot of the functions $f \mapsto V_A^{f,\gamma_B(f)}(s)$ when $f \in W_A$, are displayed in Figure 8. We deduce from this that the global maximum of both $V_A^{f,\gamma_B(f)}(s_1)$ and $V_A^{f,\gamma_B(f)}(s_2)$ is

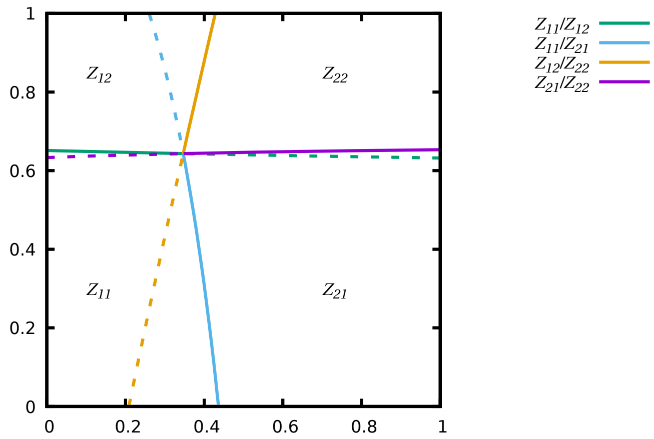


Figure 7: Zones $\mathcal{Z}_{ij}$ for the SSSE and their boundaries in Example 1

located at the point where all four zones meet. The zone that realizes the maximum is $\mathcal{Z}_{21}$. The coordinates of this point are obtained e.g. by solving for (f_1, f_2) in the equations $V_B^{f,b_1b_1}(s_1) = V_B^{f,b_1b_2}(s_1) = V_B^{f,b_2b_1}(s_1)$. The solution provides the value of f_2^* as the root of polynomial $p(f_2) = 3465\, f_2^3 - 22604\, f_2^2 + 26345\, f_2 - 8516$ that belongs to $[0,1]$, and f_1^* as a rational function of it. Finally, there exists a SSSE, given by the elements in Table 2.

C.1.4 One-step Value Iteration from the SSSE

Since there is a natural candidate for a FPE which is the SSSE computed in Table 2, we can check whether this strategy satisfies the conditions for FPE. Replacing $v_B(s_1)$ and $v_B(s_2)$ by the numerical values of Table 2, we obtain the functions (up to rounding of floating-point values):

$$g_B(s_1, f, b_1) = -14.25f + 3.13 \qquad g_B(s_1, f, b_2) = 11.01f - 5.62$$
$$g_B(s_2, f, b_1) = -15.25f + 9.13 \qquad g_B(s_2, f, b_2) = 17.01f - 11.62.$$

Likewise for Player A, with the scrap value function $v_A(s_1) = 26.841$, $v_A(s_2) = 28.437$ (see again Table 2), the gains are given by:

$$g_A(s_1, f, b_1) = 17.64f + 17.23 \qquad g_A(s_1, f, b_2) = -9.56f + 30.15$$
$$g_A(s_2, f, b_1) = 9.64f + 22.23 \qquad g_A(s_2, f, b_2) = -1.56f + 26.15.$$

Finally, the decision problem of Player A is represented in Figure 9: for each state s, we plot the function $g_A(s, f)$:

$$f \mapsto g_A(s, f, b_1)\ \mathbf{1}_{\{g_B(s,f,b_1) \geq g_B(s,f,b_2)\}} + g_A(s, f, b_2)\ \mathbf{1}_{\{g_B(s,f,b_1) < g_B(s,f,b_2)\}}.$$

It is checked that the maximum is attained at the values of f that are given in Table 2. This means that the SSSE is a FPE.

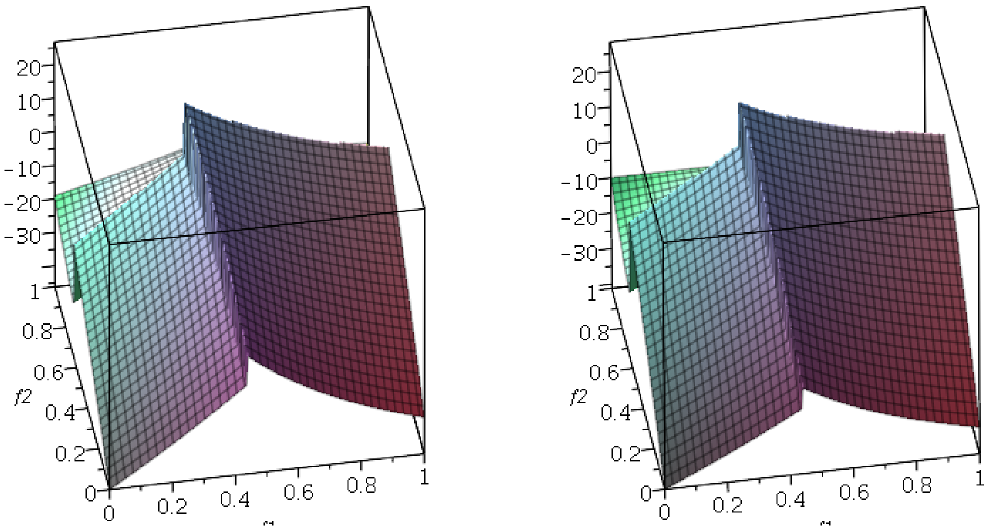


Figure 8: Value $V_A^{f,\gamma_B(f)}(s)$ in Example 1, for $s = s_1$ (left) and $s = s_2$ (right)

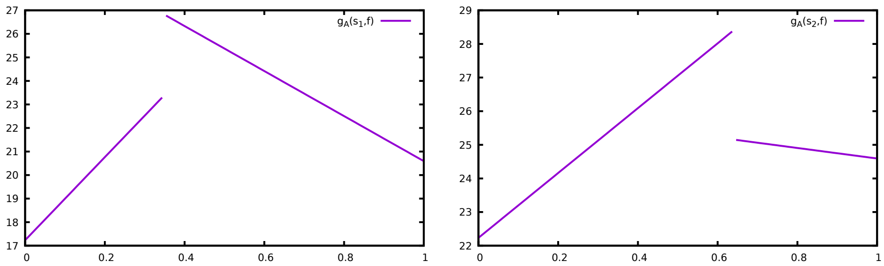


Figure 9: Value $g_A(s, f)$ in Example 1, for $s = s_1$ (left) and $s = s_2$ (right)

C.2 Analysis of Example 2

In this section, we provide details on the analysis of Example 2 introduced in Section 4.3. We compute the Stationary Strong Stackelberg Equilibria (SSSE) and discuss their existence. We compute the one-step operator and discuss its fixed points.

C.2.1 Data

The data is presented in Table 15 (also Table 3).

	b_1	b_2
a_1	(1, 0) / $(-1,-2)$	(1, 0) / $(-2,1)$
a_2	(1, 0) / $(0,0)$	(1, 0) / $(2,0)$

State s_1

	b_1	b_2
a_1	(1, 0) / $(-1,-2)$	(1, 0) / $(-2,-2)$
a_2	(0, 1) / $(0,1)$	(1, 0) / $(1,1)$

State s_2

Table 15: Transition matrix and payoffs for each player in Example 2

Observe that state s_1 is absorbing whatever the players play, and state s_2 is followed by state s_1, unless players play the combination (a_2, b_1). This is represented in Figure 2. Also, in state s_2, Player B's actions are indifferent to her.

C.2.2 Computation of the SSSE

Values of stationary strategies. Given a stationary strategy (f, g), let us compute the value V_i^{fg} of this strategy for each player i =A,B. The notation is $f_i = f(s_i, a_1)$ and $g_i = g(s_i, b_1)$. We have:

$$
\begin{aligned}
v_A(s_1) &= f_1 g_1[-1 + \beta_A v_A(s_1)] + f_1(1 - g_1)[-2 + \beta_A v_A(s_1)] \\
&\quad + (1 - f_1)g_1[0 + \beta_A v_A(s_1)] + (1 - f_1)(1 - g_1)[2 + \beta_A v_A(s_1)] \\
&= \frac{1}{1 - \beta_A}(3f_1 g_1 + 2 - 4f_1 - 2g_1) \qquad (51) \\
v_B(s_1) &= f_1 g_1[-2 + \beta_B v_B(s_1)] + f_1(1 - g_1)[1 + \beta_B v_B(s_1)] \\
&\quad + (1 - f_1)g_1[0 + \beta_B v_B(s_1)] + (1 - f_1)(1 - g_1)[0 + \beta_B v_B(s_1)] \\
&= \frac{f_1(1 - 3g_1)}{1 - \beta_B}\ . \qquad (52) \\
v_A(s_2) &= f_2 g_2[-1 + \beta_A v_A(s_1)] + f_2(1 - g_2)[-2 + \beta_A v_A(s_1)] \\
&\quad + (1 - f_2)g_2[0 + \beta_A v_A(s_2)] + (1 - f_2)(1 - g_2)[1 + \beta_A v_A(s_1)] \\
&= \frac{1}{1 - (1 - f_2)g_2\beta_A}(2f_2 g_2 - 3f_2 - g_2 + 1 + (1 - (1 - f_2)g_2)\beta_A v_A(s_1)) \qquad (53) \\
v_B(s_2) &= f_2 g_2[-2 + \beta_B v_B(s_1)] + f_2(1 - g_2)[-2 + \beta_B v_B(s_1)] \\
&\quad + (1 - f_2)g_2[1 + \beta_B v_B(s_2)] + (1 - f_2)(1 - g_2)[1 + \beta_B v_B(s_1)]
\end{aligned}
$$

$$= \frac{1}{1-(1-f_2)g_2\beta_B}\left(1-3f_2+(1-(1-f_2)g_2)\beta_B v_B(s_1)\right). \tag{54}$$

Analysis of values. The objective is to compute, for every state, the best strategy $f \in W_A$ for Player A: the set $R_A(s)$. A SSSE will exist if and only if the intersection of these sets is nonempty. We start with the absorbing state s_1 and proceed with s_2.

State s_1. We look at the maximum of $V_B^{fg}(s_1) = v_B(s_1)$ with respect to (g_1, g_2). Because state s_1 is absorbing, the expression of $v_i(s_1)$, $i =$ A,B, depends only on f_1 and g_1. The reaction of Player B will be of the form $R_B(f) = R_B(f, s_1) \times \{b_1, b_2\}$, where $R_B(f, s_1) \subset \{b_1, b_2\}$.

Clearly with (52), Player B prefers b_2 ($g_1 = 0$) when $f_1 \neq 0$. When $f_1 = 0$, she is indifferent between b_1 and b_2. She will break the tie in favor of Player A, whose gain is, from (51):

$$\frac{2-2g_1}{1-\beta_A}\,.$$

In that case also, B will play $g_1 = 0$.

So Player A gets to choose f_1 knowing that $g_1 = 0$ in (51). The optimum is reached at $f_1 = 0$. So the best for Player A is when she plays a_2 and B reacts with b_2. The values are then:

$$v_A^*(s_1) \;=\; \frac{2}{1-\beta_A} \qquad v_B^*(s_1) \;=\; 0. \tag{55}$$

State s_2. We look now at the maximum of $V_B^{fg}(s_2)$ with respect to (g_1, g_2), the values of (f_1, f_2) being fixed. As before, we define $\mathcal{Z}_{ij}$ as the subset of W_A where playing (b_i, b_j) is optimal for Player B. First we replace in (53) and (54) the terms $v_i(s_1)$ by their value from (51) and (52):

$$\begin{aligned} v_A(s_2) &= \frac{1}{1-(1-f_2)g_2\beta_A}\Big(2f_2g_2 - 3f_2 - g_2 + 1 \\ &\qquad\qquad + (1-(1-f_2)g_2)\frac{\beta_A}{1-\beta_A}(3f_1g_1 + 2 - 4f_1 - 2g_1)\Big) \end{aligned} \tag{56}$$

$$\begin{aligned} v_B(s_2) &= \frac{1}{1-(1-f_2)g_2\beta_B}\left(1-3f_2+(1-(1-f_2)g_2)\beta_B\frac{f_1(1-3g_1)}{1-\beta_B}\right) \\ &= \frac{1-3f_2-f_1+3f_1g_1}{1-(1-f_2)g_2\beta_B} + f_1\,\frac{1-3g_1}{1-\beta_B}\,. \end{aligned} \tag{57}$$

We deduce:

$$\frac{\partial}{\partial g_1}v_B(s_2) = -\,\frac{3\beta_B f_1(1-(1-f_2)g_2)}{(1-\beta_B)(1-(1-f_2)g_2\beta_B)}\,.$$

As a function of (g_1, g_2), $v_B(s_2)$ is strictly decreasing with respect to g_1, when $\beta_B \neq 0$, $f_1 \neq 0$ and $1-(1-f_2)g_2 \neq 0$. Its maximum is then attained at $g_1 = 0$ for all values of g_2.

We now assume $\beta_B \neq 0$. The case $\beta_B = 0$ will be handled separately.

If $f_1 = 0$, Player B is indifferent between playing $g_1 = 0$ or $g_1 = 1$ when starting in state s_2, just as she was when starting in state s_1. The tie is broken in favor of Player A, whose gain in this situation is, from (56):

$$\frac{1}{1-(1-f_2)g_2\beta_A}\left(2f_2g_2 - 3f_2 - g_2 + 1 + (1-(1-f_2)g_2)\frac{\beta_A}{1-\beta_A}(2-2g_1)\right) .$$

Again, Player B strictly prefers $g_1 = 0$, unless $1-(1-f_2)g_2 = 0$ or $\beta_A = 0$.

At this point, we conclude that g_1 must be 0 unless $1-(1-f_2)g_2 = 0$, which never happens when $f_2 > 0$, or $\beta_A = 0$. We focus first on the case where $f_2 > 0$, then come back to the case $f_2 = 0$. We also assume $\beta_A > 0$, the case $\beta_A = 0$ will be handled separately.

The case $f_2 > 0$. As we just concluded, $g_1 = 0$ in this case. Points (f_1, f_2) with $f_2 > 0$ belong to either $\mathcal{Z}_{21}$ or $\mathcal{Z}_{22}$, or both. Setting $g_1 = 0$ in (57), we have:

$$v_B(s_2) \;=\; \frac{1-3f_2-f_1}{1-(1-f_2)g_2\beta_B} + \frac{f_1}{1-\beta_B} .$$

The derivative of this function of g_2 is:

$$\frac{\partial}{\partial g_2}v_B(s_2) = \frac{(1-3f_2-f_1)\beta_B(1-f_2)}{(1-(1-f_2)g_2\beta_B)^2} \tag{58}$$

and therefore is either 0 if $f_2 = 1$, or else has the sign of $1-f_1-3f_2$. We study the case $f_2 = 0$ separately. If $f_1 < 1-3f_2$, Player B will prefer $g_2 = 1$ (play b_1): those points are in $\mathcal{Z}_{12}$. If $f_1 > 1-3f_2$, Player B will prefer $g_2 = 0$ (play b_0): those points are in $\mathcal{Z}_{22}$.

If equality holds, Player B is indifferent. She will break the tie in favor of Player A. The value of A on the line $f_1 = 1-3f_2$ is obtained from (56) as:

$$v_A(s_2) = \max\left\{\frac{f_2(\beta_A(12f_2-1)-1)}{(1-\beta_A)(1-\beta_A(1-f_2))},\; 3f_2\,\frac{5\beta_A-1}{1-\beta_A} + \frac{1-3\beta_A}{1-\beta_A}\right\} , \tag{59}$$

according to whether zone $\mathcal{Z}_{12}$ or zone $\mathcal{Z}_{22}$ is considered, respectively. It is possible to determine which term realizes the maximum for each value of f_2, but since the ultimate goal is to optimize with respect to (f_1, f_2), we skip this step here.

The case $f_2 = 1$. According to (58), Player B is indifferent between $g_2 = 0$ and $g_2 = 1$. She must break the tie in favor of Player A, whose value is, from (56):

$$g_2 - 2 + \frac{\beta_A}{1-\beta_A}(2-2f_1) .$$

Playing $g_2 = 1$ will maximize this value.

The case $f_2 = 0$. Setting f_2 in (57), we arrive at:

$$v_B(s_2) \;=\; \frac{1}{1-\beta_B} - \frac{\beta_B}{1-\beta_B}\,\frac{(1-g_2)(3f_1g_1 - f_1 + 1)}{1-\beta_B g_2}\;.$$

It is seen that the second term in the right-hand side is always positive, and is zero when either $g_2 = 1$, or $f_1 = 1$ and $g_1 = 0$. The maximum possible is therefore $v_B(s_2) = 1/(1-\beta_B)$, and it is attained when either $g_2 = 1$ (independently of f_1 and g_1: in that case, the point is in $\mathcal{Z}_{11} \cap \mathcal{Z}_{21}$), or when $f_1 = 1$ and $g_1 = 0$ (independently of g_2: in that case, the point is in $\mathcal{Z}_{22}$ as well).

Optimization for Player A. According to the preferences of Player B, we have identified a partition of W_A in five sets, on which we proceed to find Player A's maximum.

1. $f_2 = 0$. As we have just seen, the whole line is in $\mathcal{Z}_{11} \cap \mathcal{Z}_{21}$. In both cases, $g_2 = 1$ and Player A's value is 0 for all f_1. The point $(f_1, f_2) = (1, 0)$ is also in $\mathcal{Z}_{22}$: in that case Player A's value is $(1-3\beta_A)/(1-\beta_A)$, see the third case.

2. $f_2 = 1$. Here, $g_1 = 0$, $g_2 = 1$: the whole line is in $\mathcal{Z}_{21}$. Player A's value is:
$$\frac{\beta_A(3-4f_1)-1}{1-\beta_A}.$$
It is therefore maximized at $f_1 = 0$.

3. $(f_1 - 1)/3 < f_2 < 1$. Here, the zone is $\mathcal{Z}_{22}$ ($g_1 = 0$ and $g_2 = 0$). The value of A is
$$1 - 3f_2 + \frac{2\beta_A(1-2f_1)}{1-\beta_A}.$$
It is therefore maximized when f_1 is as small as possible, which is at $f_1 = 0$, when $f_2 > 1/3$, and $f_1 = 1 - 3f_2$ otherwise, see the next case.

4. $f_1 = 1 - 3f_2$. The value of A is given by (59). The second term in this expression, corresponding to zone $\mathcal{Z}_{22}$, is maximized when $f_2 = 0$ (and $f_1 = 1$) when $\beta_A < 1/5$, maximized when $f_2 = 1/3$ (and $f_1 = 0$) when $\beta_A > 1/5$, and constant when $\beta_A = 1/5$. The first term in (59) corresponds to zone $\mathcal{Z}_{12}$ and is always dominated by other values as we see in the next case.

5. $0 < f_2 < (f_1 - 1)/3$. This is zone $\mathcal{Z}_{21}$ ($g_1 = 0$ and $g_2 = 1$). The value of A is
$$\frac{f_2}{1-\beta_A(1-f_2)}\,\frac{\beta_A(3-4f_1)-1}{1-\beta_A}.$$
This is strictly decreasing with respect to f_1 and therefore maximal at $f_1 = 0$.

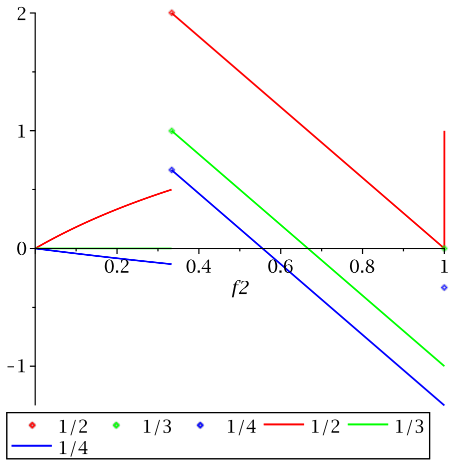


Figure 10: Gain of A as a function of f_2 for $\beta_A = 1/2, 1/3, 1/4$

The case $f_1 = 0$. From this list, we conclude that when $\beta_A \geq 1/5$, the maximum value for A is found when $f_1 = 0$. We focus on this case now. Letting $f_1 = 0$ in the value of B, we arrive at:

$$g_B = \frac{1 - 3f_2}{1 - (1 - f_2)g_2\beta_B} .$$

As a function of g_2, this is strictly increasing if $1-3f_2 > 0$, strictly decreasing if $1-3f_2 < 0$ and $f_2 \neq 1$, and constant if $1-3f_2 = 0$ or $f_2 = 1$. The optimal choice of g_2 is respectively 1 (B plays b_1), 0 (B plays b_2) and indifferent. In these two cases, the tie-breaking between b_1 and b_2 in favor of Player A amounts to comparing the two values, respectively:

$$\frac{f_2}{1 - \beta_A(1 - f_2)} \frac{3\beta_A - 1}{1 - \beta_A} \qquad \text{and} \qquad \frac{1 - 3f_2 + \beta_A(1 + 3f_2)}{1 - \beta_A} .$$

For these different cases, the play of B and the value of A are given in the table:

range	B's play	A's value
$0 \leq f_2 < 1/3$	b_1	$\frac{f_2}{1 - \beta_A(1 - f_2)} \frac{3\beta_A - 1}{1 - \beta_A}$
$1/3 \leq f_2 < 1$	b_2	$\frac{1 - 3f_2 + \beta_A(1 + 3f_2)}{1 - \beta_A}$
$f_2 = 1$	b_1	$\frac{3\beta_A - 1}{1 - \beta_A}$

The shape of this function of f_2 for different values of β_A is displayed in Figure 10.

Whatever the value of β_A, the maximum of this function is attained at $f_2 = 1/3$, with values

$$v^*_A(s_2) = \frac{2\beta_A}{1 - \beta_A} \qquad v^*_B(s_2) = 0. \tag{60}$$

The case $\beta_B = 0$. If $\beta_B = 0$, Player B's value in state s_2 does not depend on g_1 nor on g_2. She will therefore break this tie in favor of Player A. This is equivalent to have A choosing all four variables f_1, f_2, g_1, g_2 in (56). As in the general analysis, we have a dichotomy: either $\beta_A(1 - (1 - f_2)g_2) = 0$, either it is not 0.

In the second case, the maximization starts with maximizing with respect to (f_1, g_1), which yields $f_1 = g_1 = 0$ because it is the same as playing the static game in s_1. Then the analysis proceeds as in the case $f_1 = 0$ above and yields the value in (60).

In the first case, either $\beta_A = 0$, which we analyze below, or $f_2 = 0$ and $g_2 = 1$. In that situation, Player A's value is always 0 and in particular does not depend on (f_1, g_1). When $\beta_A > 0$, Player A will always prefer the other case.

The case $\beta_A = 0$. The necessity to consider this case appears when Player B has to break ties in favor of Player A. This is in particular the case when $f_1 = 0$, because B is indifferent between $g_1 = 0$ and $g_1 = 1$.

Player A's value is then: $v_A(s_2) = 2f_2g_2 - 3f_2 - g_2 + 1 = (2f_2 - 1)g_2 + 1 - 3f_2$ and is also indifferent to the choice of g_1.

When $f_1 > 0$, the analysis of Player B's response proceeds as in the general case and falls in the case $\beta_A < 1/5$.

Synthesis. Finally, we have four cases:

$\beta_A = 0$**:** The sets $R_A(s)$ are given by the following tables:

$R_A(s_1)$	s_1	s_2
Play of A	a_2	$\{a_1, a_2\}$
Play of B	b_2	$\{b_1, b_2\}$
$v_A(s_1)$	2	
$v_B(s_1)$	0	

$R_A(s_2)$	s_1	s_2
Play of A	a_1	a_2
Play of B	$\{b_1, b_2\}$	b_2
$v_A(s_2)$	$\dfrac{1 - 3\beta_A}{1 - \beta_A}$	
$v_B(s_2)$	$\dfrac{1}{1 - \beta_B}$	

$0 < \beta_A < 1/5$**:** The sets $R_A(s)$ are given by the following tables:

$R_A(s_1)$	s_1	s_2
Play of A	a_2	$\{a_1, a_2\}$
Play of B	b_2	$\{b_1, b_2\}$
$v_A(s_1)$	$\dfrac{2}{1 - \beta_A}$	
$v_B(s_1)$	0	

$R_A(s_2)$	s_1	s_2
Play of A	a_1	a_2
Play of B	b_2	b_2
$v_A(s_2)$	$\dfrac{1 - 3\beta_A}{1 - \beta_A}$	
$v_B(s_2)$	$\dfrac{1}{1 - \beta_B}$	

The intersection $R_A(s_1) \cap R_A(s_2)$ is empty: there exist no SSSE.

$\beta_A = 1/5$**:** The set $R_A(s_1)$ is as above. The set $R_A(s_2)$ is described in the following table:

$R_A(s_2)$	s_1	s_2
Play of A	$(1-3\phi, 3\phi)$	$(\phi, 1-\phi)$
Play of B	b_2	$\{b_1, b_2\}$
$v_A(s_2)$	$1/2$	
$v_B(s_2)$	$\dfrac{\phi}{1-\beta_B}$	

where $\phi \in [0, 1/3]$. When $\phi = 1/3$, the corresponding element of $R_A(s_2)$ is equal to the element of $R_A(s_1)$. There is therefore an intersection, which is a SSSE.

$\beta_A \geq 1/5$**:** The set $R_A(s_1)$ is as above and the set $R_A(s_2)$ coincides with it. There exists a unique SSSE with strategies and values as:

	s_1	s_2
Play of A	a_2	$(1/3, 2/3)$
Play of B	b_2	b_2
v_A	$\dfrac{2}{1-\beta_A}$	$\dfrac{2\beta_A}{1-\beta_A}$
v_B	0	0

C.2.3 The one-step Value Iteration operator

Computation of the operator. We compute here the operator T, which maps pairs of value functions to pairs of value functions. We restrict our attention to those pairs where the value at state s_1 is the fixed-point obtained in (55). There are therefore two remaining variables: $v_A(s_2)$ and $v_B(s_2)$. In order to clarify formulas a bit, we use the symbols $w := v_A(s_2)$ and $z := v_B(s_2)$. The images $(Tv)_A(s_2)$ and $(Tv)_B(s_2)$ will be denoted with w' and z' respectively.

Case $z < 0$**:** here, B always plays b_2 and A plays a_2. We have the fixed mapping:

$$(w', z') = \left(\frac{1+\beta_A}{1-\beta_A}, 1\right) =: P_{22} \ .$$

Case $z > 0$**:** here, B always plays b_1.

Case $w < (3\beta_A - 1)/\beta_A/(1-\beta_A)$**:** A plays a_1. The mapping is also fixed:

$$(w', z') = \left(\frac{3\beta_A - 1}{1-\beta_A}, -2\right) =: P_{11} \ .$$

Case $w > (3\beta_A - 1)/\beta_A/(1-\beta_A)$**:** A plays a_2. The mapping is linear:

$$(w', z') = (\beta_A w, 1 + \beta_B z) =: T_{21}(w, z) \ .$$

Case $w = (3\beta_A - 1)/\beta_A/(1-\beta_A)$: A plays anything she wants. Let ϕ be this strategy, the mapping is:

$$(w', z') = (\beta_A w, 1 - 3\phi + \beta_B(1-\phi)z) =: T_{\phi,0}(w, z) \ . \tag{61}$$

Case $z = 0$: B is indifferent with its own reward. Tie is broken in favor of A.

Case $w \leq (1+\beta_A)/\beta_A/(1-\beta_A)$: A prefers (a_2, b_2). The mapping is to P_{22}.

Case $w > (1+\beta_A)/\beta_A/(1-\beta_A)$: A prefers (a_2, b_1). The mapping is $T_{21}(w, z)$.

In summary, the plane (w, z) is partitioned in three zones. One maps to P_{22}, one maps to P_{11}. The third one has the map T_{21}. This mapping admits as fixed point:

$$P_{21} := \left(0, \frac{1}{1-\beta_B}\right) \ .$$

The frontier between $\mathcal{Z}_{21}$ and $\mathcal{Z}_{22}$, that is, the half-line $\{w = (3\beta_A - 1)/\beta_A/(1-\beta_A); z > 0\}$, has a special status: depending on the choice of ϕ in (61), an infinite number of mappings can be chosen.

Dynamics of the operator. Two main situations occur: either $\beta_A \geq 1/3$, either $\beta_A \leq 1/3$. Those are represented in Figure 11, on the left and the right, respectively. The three zones are represented: $\mathcal{Z}_{11}$ in green, $\mathcal{Z}_{22}$ in blue and $\mathcal{Z}_{21}$ in yellow. Point C is the triple point. Point S is a SSSE. The horizontal segment $(-\infty, L]$ belongs to zone $\mathcal{Z}_{22}$. The horizontal segment (L, ∞) belongs to zone $\mathcal{Z}_{21}$.

In all cases, P_{22} is in the zone $\mathcal{Z}_{21}$ and point P_{11} is in the zone $\mathcal{Z}_{22}$.

The solid red line is the place where the mapping $T_{\phi,0}$ of (61) applies. Its image is represented as the dashed red line.

Case $\beta_A > 1/3$. In this situation, the fixed-point of operator T_{21} is located in zone $\mathcal{Z}_{11}$. The iterations of T starting in $\mathcal{Z}_{22}$ will be mapped first to P_{22}, then to $T_{21}P_{22}$, then $T_{21}^2 P_{22}$ etc. Eventually, the sequence $T_{21}^k P_{22}$ will enter zone $\mathcal{Z}_{11}$. Then the value will be mapped to P_{11} and repeat the cycle.

Iterations of T starting in zone $\mathcal{Z}_{11}$ will map to P_{11} and continue as above.

Iterations of T starting in zone $\mathcal{Z}_{21}$ will follow a sequence of points in the direction of P_{21} and eventually enter $\mathcal{Z}_{11}$, then follow the cycle as above.

Depending on the choice of ϕ, the mapping from the frontier $\mathcal{Z}_{11} - \mathcal{Z}_{21}$ may lead to either $\mathcal{Z}_{11}$ or $\mathcal{Z}_{22}$. It then follows the patterns described.

In all cases, there is a limit cycle. The length of this cycle can range from 3 to ∞ depending on the proximity of β_A to 1/3.

Case $\beta_A \leq 1/3$. In this situation, the fixed-point P_{21} of operator T_{21} is located in zone $\mathcal{Z}_{21}$: it is a FPE. The iterations of T starting in $\mathcal{Z}_{21}$ will converge to it.

Iterations of T starting in $\mathcal{Z}_{22}$ will be mapped first to P_{22}, then converge to P_{21}. Likewise, iterations starting in $\mathcal{Z}_{11}$ will be mapped first to P_{11} in $\mathcal{Z}_{22}$, then converge to P_{21}. Iterations starting on the frontier $\mathcal{Z}_{11} - \mathcal{Z}_{21}$ either converge directly, or make one step in $\mathcal{Z}_{22}$.

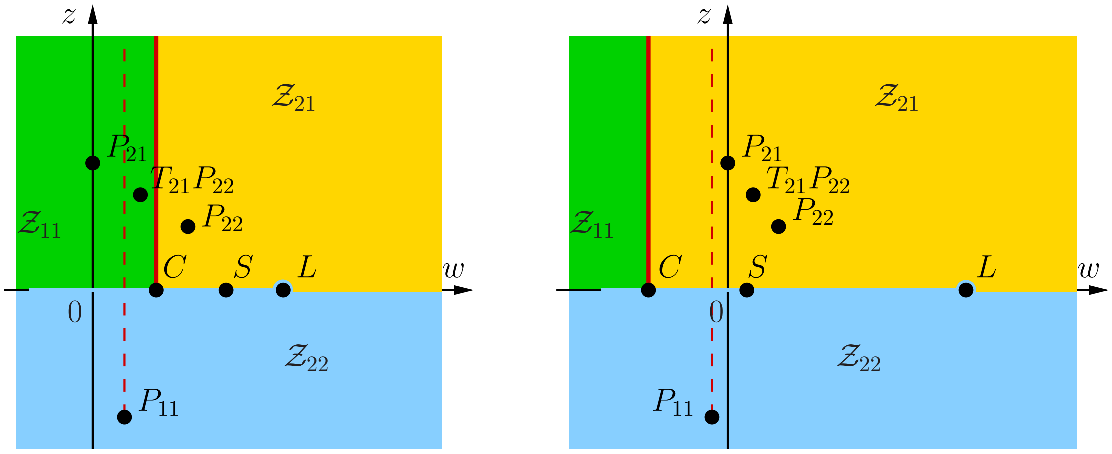


Figure 11: Zones of the mapping T: left, with $\beta_A > 1/3$; right, with $\beta_A < 1/3$

C.2.4 One-step VI from the SSSE

In this section, we compute the action of VI on the SSSE computed above, that is:

$$v_A^*(s_1) = \frac{2}{1-\beta_A} \qquad v_A^*(s_2) = \frac{2\beta_A}{1-\beta_A} \qquad v_B^*(s_1) = v_B^*(s_2) \ = \ 0. \tag{62}$$

Application of the operator formula. Using the operator defined in Section C.2.3, we find that $z = 0$. We must compare $w = (2\beta_A)/(1-\beta_A)$ to $L = (1+\beta_A)/\beta_A/(1-\beta_A)$ and find that $w \le L$. Therefore the FPE lies in the zone $\mathcal{Z}_{22}$ and it is mapped to P_{22} by T.

The value obtained is then:

$$(Tv^*)_A(s_2) = \frac{1+\beta_A}{1-\beta_A} \qquad (Tv^*)_B(s_2) = 1.$$

Direct computation. We provide a stand-alone direct derivation.

It is quickly checked that for state s_1, this function is a fixed point: the action of VI does not modify it.

For state s_2, we get (see the computation of (54)):

$$\begin{aligned} T^{fg}v_B^*(s_2) &= f_2g_2[-2] + f_2(1-g_2)[-2] + (1-f_2)g_2[1] + (1-f_2)(1-g_2)[1] \\ &= 1 - 3f_2 \ . \end{aligned} \tag{63}$$

Player B is indifferent. She will break this tie in favor of Player A, whose gain is (see (53)):

$$Tv_A^*(s_2) = f_2g_2[-1 + \beta_A\frac{2}{1-\beta_A}] + f_2(1-g_2)[-2 + \beta_A\frac{2}{1-\beta_A}]$$

$$+ (1 - f_2) g_2 [0 + \beta_A \frac{2\beta_A}{1 - \beta_A}] + (1 - f_2)(1 - g_2)[1 + \beta_A \frac{2}{1 - \beta_A}] \, . \qquad (64)$$

The choice between b_1 and b_2 amounts to comparing the respective gains:

$$g_A(b_1) = -f_2 + \beta_A \frac{2}{1 - \beta_A}(f_2 + \beta_A(1 - f_2))$$
$$g_A(b_2) = 1 - 3f_2 + \beta_A \frac{2}{1 - \beta_A} \, .$$

To avoid further calculations, note that both functions are decreasing wrt f_2. Therefore $\max\{g_A(b_1),\, g_A(b_2)\}$ is also decreasing, and its maximum obtains at $f_2 = 0$. The values of both players are then:

$$(Tv^*)_A(s_2) = \frac{1 + \beta_A}{1 - \beta_A} \qquad (Tv^*)_B(s_2) = 1.$$

Interpretation. Independently of β_A, the value of both players is improved by applying the operator once, starting from the SSSE.

It means that the policy in which Player A announces she:

- plays once a_2;
- plays the SSSE afterwards

gives a better reward to both players.

This policy actually *menaces* to play the SSSE, since at the first step players will play (a_2, b_2) and send the game to state s_1. The mixed strategy $(1/3, 2/3)$ of the SSSE will never be played.

C.3 Analysis of Example 3

We provide here the technical justifications for the claims of Section 4.4 on Example 3.

C.3.1 Data

The data of this example is presented in Table 16 (see also Table 6):

	b_1	b_2
a_1	(1, 0) / $(1, -1)$	(0, 1) / $(0, 1)$
a_2	(0, 1) / $(-1, 1)$	(0, 1) / $(-1, -1)$

State s_1

	b_1	b_2
a_1	(0, 1) / $(-1, 0)$	(1, 0) / $(0, 1)$
a_2	(1, 0) / $(0, 1)$	(0, 1) / $(1, -1)$

State s_2

Table 16: Transition matrix and payoffs for each player in Example 3

We claim that the following pair of strategies (f^*, g^*) and value functions (v_A^*, v_B^*) constitute *both* a SSSE and a FPE:

$$f^*(s_1, a_1) = 1 \quad f^*(s_2, a_1) = 5 - \sqrt{19} \quad v_A^*(s_1) = \frac{-3 + \sqrt{19}}{5} \quad v_A^*(s_2) = \frac{-6 + 2\sqrt{19}}{5} \tag{65}$$

$$g^*(s_1) = b_2 \quad g^*(s_2) = b_2 \quad v_B^*(s_1) = \frac{16 - 2\sqrt{19}}{5} \quad v_B^*(s_2) = \frac{22 - 4\sqrt{19}}{5}. \tag{66}$$

In order to support this claim, we need to construct the reaction functions $\gamma_B(f^*)$ (for the SSSE) and $\gamma_B(s, f^*, v^*)$ (for the FPE). As a common preliminary step, we first list the one-step rewards corresponding to a general strategy $f \in W_A$ and any possible action $g \in \mathcal{B}_s$, for all states s. The notation is $f_1 = f(s_1, a_2)$ and $f_2 = f(s_2, a_1)$.

$$\begin{aligned}
h_A(s_1, f, b_1, v_A) &= 2f_1 - 1 + \beta_A[f_1 v_A(s_1) + (1 - f_1)v_A(s_2)] \\
h_A(s_1, f, b_2, v_A) &= f_1 - 1 + \beta_A v_A(s_2) \\
h_A(s_2, f, b_1, v_A) &= -f_2 + \beta_A[f_2 v_A(s_2) + (1 - f_2)v_A(s_1)] \\
h_A(s_2, f, b_2, v_A) &= 1 - f_2 + \beta_A[f_2 v_A(s_1) + (1 - f_2)v_A(s_2)] \\
h_B(s_1, f, b_1, v_B) &= 1 - 2f_1 + \beta_B[f_1 v_B(s_1) + (1 - f_1)v_B(s_2)] \\
h_B(s_1, f, b_2, v_B) &= -1 + 2f_1 + \beta_B v_B(s_2) \\
h_B(s_2, f, b_1, v_B) &= 1 - f_2 + \beta_B[f_2 v_B(s_2) + (1 - f_2)v_B(s_1)] \\
h_B(s_2, f, b_2, v_B) &= 2f_2 - 1 + \beta_B[f_2 v_B(s_1) + (1 - f_2)v_B(s_2)].
\end{aligned}$$

C.3.2 Computation of the SSSE

Given a strategy $f \in W_A$, we know from MDP theory that the best response of Player B is found among the four pure strategies of $W_B \times W_B$. The value of each of these strategies for both players is obtained by solving the four equations of the list above, where the combination of s_i and b_j is the one sought, and where $h_i(s, f, b, v_i)$ is replaced with $v_i(s_i)$.

When setting $\beta_B = 1/2$, the four values of $V_B^{fg}(s_2)$ turn out to be:

$$V_B^{f,b_1 b_1}(s_2) = \frac{6(1 - f_2)(1 - f_1)}{3 - f_1 - f_2} \qquad V_B^{f,b_1 b_2}(s_2) = \frac{2(4f_1 f_2 - f_1 - 5f_2 + 2)}{f_1 - f_2 - 2}$$
$$V_B^{f,b_2 b_1}(s_2) = \frac{2(1 - f_2)(2f_1 + 1)}{3 - f_2} \qquad V_B^{f,b_2 b_2}(s_2) = \frac{2(2f_1 f_2 + 3f_2 - 2)}{f_2 + 2}\ .$$

The comparison of these four values determines four zones delimited by lines as in Figure 12.

When evaluating Player A's value on each of these zones, including the boundaries, it is found that A's optimum lies at the point specified in (65). This point lies at the boundary of B's best responses $g = (b_2, b_1)$ and (b_2, b_2) (line L6), where $f_1 = 1$. Player B breaks the tie in favor of Player A by playing b_2.

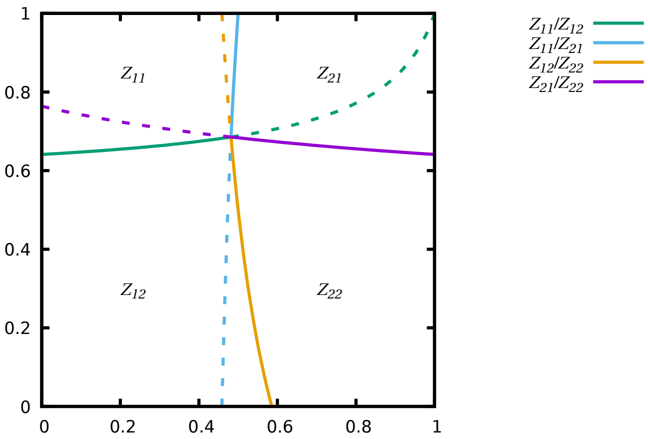


Figure 12: Best response zones for Example 3

C.3.3 Verification of the FPE

When replacing $v_B(s_1)$ and $v_B(s_2)$ by the values in (66), the functions $h_B(s, \cdot)$ of the variable $f(s, a_1)$ are as shown in Figure 13. The intersections occur for the values $f(s_1, a_1) = f_1^* := (23 + \sqrt{19})/51$ and $f(s_2, a_1) = f_2^* := f^*(s_2, a_1)$ as in (65). In addition, it is checked that for these values of $f(s, a_1)$,

$$h_A(s_1, f, b_1, v_A^*) > h(s_1, f, b_2, v_A^*) \qquad h_A(s_2, f, b_1, v_A^*) < h(s_2, f, b_2, v_A^*)$$

so that the "strong" response of Player B is:

$$\gamma_B(s_1, f, v) = \begin{cases} b_1 & \text{if } f_1 \leq f_1^* \\ b_2 & \text{if } f_1 > f_1^* \end{cases} \qquad \gamma_B(s_2, f, v) = \begin{cases} b_1 & \text{if } f_2 < f_2^* \\ b_2 & \text{if } f_2 \geq f_2^*. \end{cases}$$

In particular, this confirms that $\gamma_B(s, f^*, v) = \{b_2\}$ for $s = s_1, s_2$.

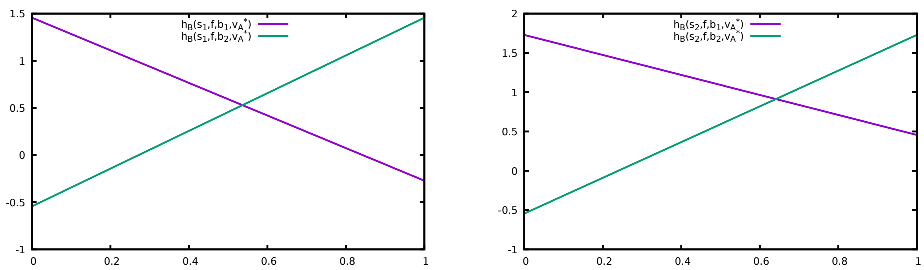


Figure 13: Player B's responses as a function of $f_1 = f(s, a_1)$ in Example 3; state s_1 (left) and state s_2 (right)

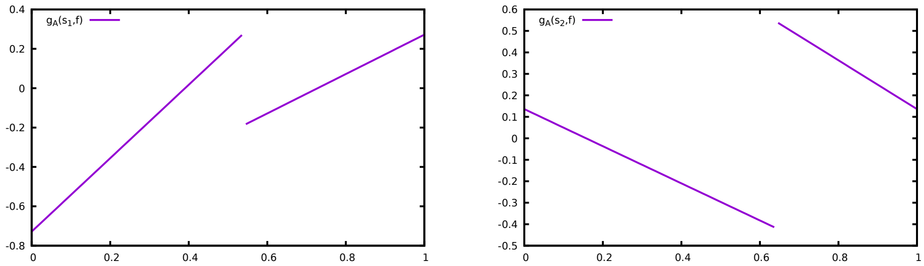


Figure 14: Value of Player A as a function of $f_1 = f(s, a_1)$ in Example 3; state s_1 (left) and state s_2 (right)

Then, the functions $g_A(s, f) := h_A(s, f, \gamma(s, f, v), v_A^*)$ for both states s are represented in Figure 14. It is concluded that

$$R_A(s_1, v) = \{(f_1^*, 1 - f_1^*), (1, 0)\} \qquad R_A(s_2, v) = \{(f_2^*, 1 - f_2^*)\}.$$

So indeed, the policy f^* of (65) is such that $f^*(s) \in R_A(s, v)$ for all s.

There remains to check that the values v_A^* and v_B^* provided are indeed the values of the policy (f^*, g^*). Indeed:

$$\begin{aligned}
h_A(s_1, f^*, b_2, v_A^*) &= \beta_A v_A^*(s_2) = \frac{-3+\sqrt{19}}{5} = v_A^*(s_1) \\
h_A(s_2, f^*, b_2, v_A^*) &= 1 - (5-\sqrt{19}) + \frac{1}{2}[(5-\sqrt{19})\frac{-3+\sqrt{19}}{5} + (1-5+\sqrt{19})\frac{-6+2\sqrt{19}}{5}] \\
&= \sqrt{19} - 4 + \frac{1}{2}\frac{-3+\sqrt{19}}{5}[-3+\sqrt{19}] = \sqrt{19} - 4 + \frac{1}{2}\frac{9+19-6\sqrt{19}}{5} \\
&= -\frac{6}{5} + \frac{2}{5}\sqrt{19} = v_A^*(s_2) \\
h_B(s_1, f, b_2, v_B^*) &= 1 + \beta_B v_B^*(s_2) = 1 + \frac{11-2\sqrt{19}}{5} = \frac{16-2\sqrt{19}}{5} = v_B^*(s_1) \\
h_B(s_2, f, b_2, v_B^*) &= 2(5-\sqrt{19}) - 1 + \frac{1}{2}[(5-\sqrt{19})\frac{16-2\sqrt{19}}{5} + (1-5+\sqrt{19})\frac{22-4\sqrt{19}}{5}] \\
&= 9 - 2\sqrt{19} + \frac{1}{5}[(5-\sqrt{19})(8-\sqrt{19}) - (4-\sqrt{19})(11-2\sqrt{19})] \\
&= 9 - 2\sqrt{19} + \frac{1}{5}[-4-19+6\sqrt{19}] = \frac{22}{5} - \frac{4}{5}\sqrt{19} = v_B^*(s_2).
\end{aligned}$$

We have therefore proved that $v^* = V^{f^* g^*}$, $g^*(s) = \gamma(s, f^*, v^*)$ and $f^* \in R_A(s, v^*)$ for all s. The solution proposed in then a FPE.

References

[1] Arunabha Bagchi. "Some Economic Applications of Dynamic Stackelberg Games." In: *Dynamic Games and Applications in Economics*. Ed. by Tamer Başar. Vol. 265. Lecture Notes in Economics and Mathematical Systems. Berlin, Heidelberg: Springer, 1986, pp. 88–102.

[2] Michèle Breton, Abderrahmane Alj, and Alain Haurie. "Sequential Stackelberg Equilibria in Two-Person Games." In: *J. Optim. Theory Appl.* 59.1 (1988), pp. 71–97.

[3] Julio B. Clempner and Alexander S. Poznyak. "Stackelberg security games: Computing the shortest-path equilibrium." In: *Expert Systems with Applications* 42.8 (2015), pp. 3967–3979.

[4] Julio Clempner and Alexander Poznyak. "Convergence method, properties and computational complexity for Lyapunov games." In: *International Journal of Applied Mathematics and Computer Science* 21.2 (2011), pp. 349–361.

[5] Francesco Maria Delle Fave, Albert Xin Jiang, Zhengyu Yin, Chao Zhang, Milind Tambe, Sarit Kraus, and John P Sullivan. "Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system." In: *Journal of Artificial Intelligence Research* 50 (2014), pp. 321–367.

[6] Eric W. Denardo. "Contraction Mappings in the Theory Underlying Dynamic Programming." In: *SIAM Review* 9.2 (1967), pp. 165–177.

[7] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[8] D. Kar, T.H. Nguyen, F. Fang, M. Brown, A. Sinha, M. Tambe, and A.X. Jiang. "Trends and Applications in Stackelberg Security Games." In: *Handbook of Dynamic Game Theory*. Ed. by T. Başar and G. Zaccour. Springer, 2018. Chap. 28, pp. 1223–1269.

[9] George Leitman. "On generalized Stackelberg strategies." In: *J. Optim. Theory Appl.* 26.4 (1978), pp. 637–643.

[10] Joshua Letchford, Liam MacDermed, Vincent Conitzer, Ronald Parr, and Charles L. Isbell. "Computing Optimal Strategies to Commit to in Stochastic Games." In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012.

[11] Tao Li and Suresh P. Sethi. "A Review of Dynamic Stackelberg Game Models." In: *Discrete and Continuous Dynamical Systems Series B* 22.1 (2017), pp. 125–159.

[12] Martin L. Puterman. *Markov decision processes*. Wiley-Interscience, 1994.

[13] Marwan Simaan and Jose Bejar Cruz Jr. "Additional Aspects on the Stackeberg Strategy in Nonzero-Sum Games." In: *J. Optim. Theory Appl.* 11.6 (1973), pp. 613–626.

[14] Marwan Simaan and Jose Bejar Cruz Jr. "On the Stackeberg Strategy in Nonzero-Sum Games." In: *J. Optim. Theory Appl.* 11.5 (1973), pp. 533–555.

[15] Yevgeniy Vorobeychik and Satinder Singh. "Computing Stackelberg Equilibria in Discounted Stochastic Games." In: *Twenty-Sicth AAAI Conference on Artificial Intelligence*. <https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4811/5686>. Corrected version retreived online on Oct. 19, 2018. 2012.

[16] Yevgeniy Vorobeychik, Bo An, Milind Tambe, and Satinder Singh. "Computing solutions in infinite-horizon discounted adversarial patrolling games." In: *Proc. 24th International Conference on Automated Planning and Scheduling (ICAPS 2014)(June 2014)*. 2014.

[17] Ward Whitt. "Representation and Approximation of Noncooperative Sequential Games." In: *SIAM J. Control and Optimization* 18.1 (1980), pp. 33–48.

[18] Byung-Wook Wie. "Dynamic Stackelberg equilibrium congestion pricing." In: *Transportation Research Part C: Emerging Technologies* 15.3 (2007), pp. 154–174.

Contents

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399