Solving strong Stackelberg equilibrium in stochastic games.

Víctor Bucarey
and Fernando Ordoñez
Departamento de Ingeniería Industrial
Universidad de Chile
Beauchef 851, Santiago, Chile
Email:vbucarey@ing.uchile.cl

Eugenio Della Vecchia
FCEIA
U. Nacional de Rosario
Pellegrini 250, Rosario, Argentina
Email:eugenio@fceia.unr.edu.ar

Alain Jean Marie
INRIA
Sophia Antipolis Méditarrenée
Email:alain.jean-marie@inria.fr

Abstract—**In this work we face the problem of finding strong Stackelberg equilibrium in stochastic games. We study a familiy of stochastic games where equilibrium in stationary policies exist and prove the convergence of value iteration and policy iteration procedures. Preliminary computational results evaluate the performance of these algorithms for stochastic games in the form of security games. Finally, we show that is not always possible to achieve strong Stackelberg equilibrium via dynamic programming.**

I. Introduction

In this work we face the problem of computing a strong Stackelberg equilibrium (SSE) in a stochastic game (SG). Given a set of states we model a two player perfect information dynamic where one of them, called *Leader* or player A, observes the current state and decides, possible up to probability distribution f, between a set of available actions. Then other player, called *Follower* or player B, observes the strategy of player A and plays his best response noted by g. We represent a two-person stochastic discrete game $\mathcal{G}$ by

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

where $\mathcal{S}$ represents the states of the games, $\mathcal{A}$ and $\mathcal{B}$ represents space of actions of both players. We denote $\mathcal{A}_s$ and $\mathcal{B}_s$ the available set of actions of both players in state s. $Q = Q^{ab}(z|s)$ represents the transition probability of going from a state s to a state z given the actions a and b were performed by players A and B respectively. $r_A^{ab}(s)$, $r_B^{ab}(s)$ represents the one-step rewards functions depending on the actual state, and the actions performed by the players. β_A and β_B are the discount factors for both players. τ represents the number of periods in the dynamic.

Stochastic games were first introduced by Shapley [1] who also gives the first algorithm to find Nash Equilibrium in zero-sum SG based on a dynamic programming algorithm. SG have been used to model interaction in economics [2], computer networks [3], and security [4], among others applications. Feedback policies are policies depending on the actual state s and time epoch t. Stationary policies can be defined as feedback policies that do not depend on the time step. In our setting, players aim to maximize their expected discounted sum of payoffs from 0 until τ considering feedback and stationary polcies. To the best of our knowledge there is no prior work on efficient algorithms to find stationary strategies in Stackelberg games for $\tau \to +\infty$ when they exist.

In the security applications that motivate this work, such as patrolling the streets to prevent crime, the decision of where to patrol next should not depend on the history of previous patrolling actions and a time independent policy is easy to communicate to real world security agents. Therefore, the focus of this work is computing Stackelberg equilibria in feedback or Markovian policies.

A complete review about Nash equilibria computation and Learning in SG can be found in [5]. In [6] authors study the relationship between Stackelberg strategies and correlated equilibrium in SG. They also shows that it is NP-hard to find Stackelberg equilibrium in SG. In contrast to MDP settings, Stackelberg equilibrium in stationary policies can be arbitrary suboptimal as is showed in [7] providing a Mixed integer non linear program to compute a SSE in general SG when players are restricted to stationary policies. They extend this formulation in [8] to policies that depends on history of bounded length.

Some applications of Stackelberg equilibrium in SG are the following:

- The problem of coordinate a group of robots for planet exploration is presented in [9]. They model it as a multi-objective SG and the solution concept used is the Stackelberg equilibrium.
- Adversarial patrolling games [10] and robotic patrolling games where a robot has to detect an intruder [11].
- Optimal policies to detect fare evasion under execution uncertainty is presented in [4] as solutions of Bayesian Stackelberg SG.

II. Algorithms and main results

A. Myopic follower strategies case ((MFS))

In this section we discuss algorithms for the case where the value function of the follower v_B do not affect in its best response. Let define the functional g of best response as follows:

$$g(f, v_B) = \arg\max_{b \in \mathcal{B}} \sum_{a \in \mathcal{A}_s} f(a) \left[r_B^{a,b}(s) + \beta_B \sum_{z \in \mathcal{S}} Q^{a,b}(z|s) v_B(z) \right] . \tag{1}$$

Here we follow the convention that the argmax is unique because in case of indifference between options, the follower select the one that favors the leader. In case the leader is also indifferent, then, in order for $g(f, v_B)$ to be well defined, the follower selects the action with the lowest index. Note that since $g(f, v_B)$ optimizes (1), $g(f, v_B)$ is at least as good as any mixed strategy.

We say that a stochastic game $\mathcal{G}$ has Myopic follower strategies (MFS) if at every step of the game the functional g is independent of v_B, that is $g(f, v_B) = g(f)$. In particular, we distinguish two important cases with MFS:

- Myopic follower: We define a game as having a myopic follower 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.
- Leader-Controller Discounted Games: This case is a particular case of the Single-controller discounted game where the controller is the leader. In other words, the transition law is in the form $Q^{ab}(z|s) = Q^a(z|s)$.

B. Stackelberg operator and Value function iteration

Let $\mathcal{F}(\mathcal{S})$ be the set of all bounded functions of the space state $\mathcal{S}$ into $\mathbb{R}$. Given a stochastic game $\mathcal{G}$ and a strategy f, we define the operator $T_A^f : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$ by the following expression:

$$T_A^f(v_A)(s) = \sum_{a \in \mathcal{A}_s} f(a) \left[r_A^{ag(f)}(s) + \beta_A \sum_{z \in S} Q^{ag(f)}(z|s) v_A(z) \right] . \tag{2}$$

Now we define the Stackelberg operator, $T_A^f : \mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$, for (MFS) case as follows:

$$T_A(v_A)(s) \;=\; \max_{f \in \mathbb{P}(\mathcal{A}_s)} \; T_A^f(v_A)(s). \tag{3}$$

This operator computes in every state the strong Stackelberg equilibrium value of being in state s for each $s \in \mathcal{S}$ and the future expected rewards are given by function v_A. Given the value function v_A and a fixed state $s \in \mathcal{S}$ the operator T_A can be computed with a mixed integer formulation or with a multiple LPs algorithm (see [12]).

In [13, Theorem 7.4] it is shown that when the algorithm for finite time horizon ends, it returns both the value for our game and a pair of Stackelberg feedback policies (π^*, γ^*) for the τ-finite horizon game. We propose Algorithm 1 in order to compute SSE in stationary policies, showing its convergence by proving Theorems II.1 and II.2.

Algorithm 1 Value function iteration: Infinite horizon

```
Require: $\varepsilon > 0$
Initialize with $n = 1$, $v_A^0(s) = 0$ for every $s \in \mathcal{S}$ and
  $v_A^1 = T_A(v_A^0)$
while $||v_A^n - v_A^{n-1}||_\infty > \varepsilon$ do
    Compute $v_A^{n+1}$ by
        $v_A^{n+1}(s) = T_A(v_A^n)(s)$ .
    Finding $f^*$ and $g^*(f)$ SSE strategies at stage $n$.
    n:=n+1
end while
return  Stationary Stackelberg policies $\pi^* = \{f^*, \ldots\}$
  and $\gamma^* = \{g^*, \ldots\}$
```

Theorem II.1. *Let $\mathcal{G}$ be a SG with MFS, then*

a) *For any stationary strategy f, the operator T_A^f : $\mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$, defined in (2) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$ of modulus β_A.*

b) *The operator T_A defined in (3) is a contraction on $(\mathcal{F}(\mathcal{S}), ||\cdot||_\infty)$, of modulus β_A.*

Theorem II.2. *Let $\mathcal{G}$ be a SG with MSF. Then the sequence of value functions v_A^n converges to v_A^*. Furthermore, v_A^* is the fixed point of T_A, and therefore, for any $n \in \mathbb{N}$,*

$$||v_A^* - v_A^n|| \leq \frac{||r_A||_\infty \, \beta_A^n}{1 - \beta_A}.$$

Given that the best response of the follower in this games are independent of the future expected value v_B ignore its behavior. The stationary pair of policies $(f^*, g(f^*))$ is guaranteed to exist and they are enough to compute the value function for SG with MFS as the fixed point:

$$v_B^* = \sum_{a \in A} f_a^* r_B^{a,g(f^*)} + \beta_B \sum_{z \in \mathcal{S}} Q^{ag(f^*)}(z|s) v_B^*.$$

C. Policy Iteration

The Policy Iteration (PI) algorithm directly iterates in the policy space. This algorithm starts with an arbitrary policy f^0 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 *Computation Phase*. This algorithm is described in Algorithm 2. We show convergence of PI by proving Lemma II.3 and Theorem II.4 for SG with MFS.

Algorithm 2 Policy Iteration

```
Choose a stationary Stackelberg pair $(f_0, g(f_0))$.
while $||u_{A,n} - u_{A,n+1}|| > \varepsilon$ do
    Evaluation phase: Find $u_{A,n}$ fixed point of the operator
      $T_A^{f_n}$.
    Improvement phase: Find a strategy $f_{n+1}$ such that
        $T_A^{f_{n+1}}(u_{A,n}) = T_A(u_{A,n})$ .
    n:= n+1
end while
return  Stationary Stackelberg policies $\pi^* = \{f^*, \ldots\}$
  and $\gamma^* = \{g(f^*), \ldots\}$
```

Lemma II.3. *If a value function v_A satisfies $v_A \leq T_A^f(v_A)$, then $v_A \leq v_A^f$, where v_A^f is the unique fixed point of $T_A^f(v_A)$.*

Theorem II.4. *The sequence of functions $u_{A,n}$ verifies $u_{A,n} \uparrow v_A^*$. Even more, if for any $n \in \mathbb{N}$, $u_{A,n} = u_{A,n+1}$, then the following it is true that $u_{A,n} = v_A^*$.*

The results exposed in this section strongly rely on the fact that $g(f, v_B)$ is independent on v_B. All the results exposed in this section may fail in the general case.

D. General Case

In general instances the main results for the (MFS) case do not hold. Operator T_A is not sufficient to describe the whole behavior of the values of both players, in particular, v_B has influence in best response in the response of leader and follower. Given a pair of stationary policies f, g, we define the following operators:

$$T_A^{f,g}(v_A)(s) = \sum_{\substack{a\in\mathcal{A}_s \\ b\in\mathcal{B}_s}} f(a)g(b) \sum_{z\in\mathcal{S}} Q^{a,b}(z|s) \left[r_A^{a,b}(s) + \beta_A v_A(z)\right], \tag{4}$$

$$T_B^{f,g}(v_B)(s) = \sum_{\substack{a\in\mathcal{A}_s \\ b\in\mathcal{B}_s}} f(a)g(b) \sum_{z\in\mathcal{S}} Q^{a,b}(z|s) \left[r_B^{a,b}(s) + \beta_B v_B(z)\right]. \tag{5}$$

Using that we can define the Stackelberg operator for the general case as:

$$(T(v_A, v_B))(s) = \Bigg(\max_{f\in\mathbb{P}(\mathcal{A}_s),} T_A^{f,g(f,v_B)}(v_A)(s), \\ T_B^{f^*,g(f^*,v_B)}(v_B)(s) \Bigg). \tag{6}$$

We show via a counterexample that this operator is not contractive in general. Anyway we computationally tested that for a special type of stochastic games, called security games, the VI algorithm converges to the stationary equilibrium policies and this operator is contractive. We adapt the VI procedure to detect if the algorithm will not converge.

III. Computational Results

Our computational tests give us for SG with MFS that PI and VI outperform any mathematical programming formulation in literature. Further, PI scale-up better than VI as the instance grows (see Figure 1).

IV. Conclusions and Future work

In this work we adapt dynamic programming based algorithms to find stationary policies forming SSE. We first show a family of SG where this type of equilibrium exists and is achievable via dynamic programming. In this case, we show that value function iteration and policy iteration converges. For the general case it may not possible. Our computational test show that PI outperforms VI is faster in MFS instances. Our experiments, also shows a special family of instances called Security Games that VI always converge. We do not provide a formal proof to the conjecture of T associated to a Security Game is contractive.

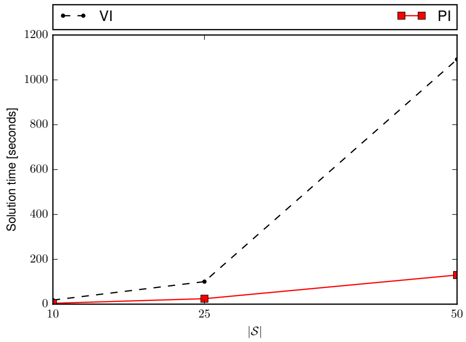

Fig. 1. Performance of PI against VI.

In future work we aim to extend the analysis to other families of stochastic games where the operator T is contractive. A second research line is to analyze the impact of approximate dynamic programming to calculate this type of equilibrium. The third line is to formalize the existence of cyclic policies that forms SSE.

Acknowledgment

This research has been partially supported by the project DyGaMe STIC AmSud.

References

[1] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[2] R. Amir, "Stochastic games in economics and related fields: an overview," in *Stochastic Games and Applications*. Springer, 2003, pp. 455–470.

[3] E. Altman, K. Avratchenkov, N. Bonneau, M. Debbah, R. El-Azouzi, and D. S. Menasché, "Constrained stochastic games in wireless networks," in *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*. IEEE, 2007, pp. 315–320.

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

[5] O. Sigaud and O. Buffet, *Markov decision processes in artificial intelligence*. John Wiley & Sons, 2013.

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

[7] Y. Vorobeychik and S. Singh, "Computing stackelberg equilibria in discounted stochastic games (corrected version)," 2012.

[8] Y. Vorobeychik, B. An, M. Tambe, and S. 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.

[9] A. Canu and A.-I. Mouaddib, "Collective decision-theoretic planning for planet exploration," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*. IEEE, 2011, pp. 289–296.

[10] Y. Vorobeychik, B. An, and M. Tambe, "Adversarial patrolling games," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 1307–1308.

[11] N. Basilico, N. Gatti, and F. Amigoni, "Leader-follower strategies for robotic patrolling in environments with arbitrary topologies," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 57–64.

[12] V. Conitzer and T. Sandholm, “Computing the optimal strategy to commit to,” in *Proceedings of the 7th ACM conference on Electronic commerce*. ACM, 2006, pp. 82–90.

[13] T. Basar, G. J. Olsder, and G. Clsder, *Dynamic noncooperative game theory*. SIAM, 1995, vol. 200.