2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Learning in Stochastic Stackelberg Games

Pranoy Das, Benita Nortmann, Lillian J Ratliff, Vijay Gupta and Thulasi Mylvaganam

Abstract— **We present a learning algorithm for players to converge to their stationary policies in a general sum stochastic sequential Stackelberg game. The algorithm is a two time scale implicit policy gradient algorithm that provably converges to stationary points of the optimization problems of the two players. Our analysis allows us to move beyond the assumptions of zero-sum or static Stackelberg games made in the existing literature for learning algorithms to converge.**

I. Introduction

Various notions of equilibrium solutions exist in game theory. Both due to the fact that such equilibrium strategies are usually difficult to compute, and as a justification for agents to arrive at strategies that constitute a given equilibrium solution, learning algorithms by which the agents can update their strategies based on past experience and (hopefully) converge to such equilibria have a long history in game theory [7], [15], [25]. In this paper, we are interested in general-sum stochastic games that have a sequential structure. In a one-shot version, these games are studied under the rubric of static Stackelberg games, considering the so-called Stackelberg equilibrium solution concept [14].

While such games arise naturally in many settings, learning algorithms that converge to Stackelberg equilibrium even in (repeated) one-shot games are less studied than the more common Nash equilbrium in (repeated) simultaneous move games. In the context of Generative Adversarial Networks (GANs) [18], where the optimization problem can be written as a zero sum game between the generator and the discriminator, [23] presented a simultaneous gradient based algorithm. In this class of algorithms, the leader and follower learn at the same rate which has been known to lead to cyclic behaviour in general [22], [26]. In the context of actor-critic algorithms in reinforcement learning, [27], [33] presented a Stackelberg game formulation and presented a policy gradient algorithm that provably locally converges to an equilibrium. Another relevant recent work is [13], which presented a two-timescale stochastic gradient algorithm and proved local convergence to differential Stackelberg equilibria in repeated static games.

As opposed to these works, we are interested in *stochastic* games [28]. Roughly speaking, stochastic games have a notion of state that evolves over time as a result of the actions by the agents. We consider two agents with the interaction between them occurring in a leader-follower manner. The leader agent commits to a policy. The policy becomes known to the follower who, in turn, plays the best response to the leader policy. The state evolves stochastically given the current state and the joint actions of the agents. The cycle repeats at the next time step. A more precise formulation is given later.

An example where stochastic Stackelberg games would be natural to consider is in security games. Security games are games between a defender who wishes to protect some targets through deployment of limited resources and attackers who wish to strategically attack the targets to benefit themselves. The hierarchical order of play arises naturally since the defender typically acts first and deploys a strategy. Attackers observe the strategy of the defender before attacking. Security games are widely modelled as stochastic games [19], [29]. While some formulations have considered zero-sum security games, particular classes of security games such as adversarial patrolling games have also been modeled as general-sum stochastic games [3]–[6]. Another example where the framework of stochastic Stackelberg games fits naturally is in testing during epidemics, where the leader (the government) sets testing policies and the follower (the citizens) decide at every time step whether to get tested. The government wishes to minimize the number of infected people in the population while the follower wishes to minimize the cost of getting sick and testing. Once again this setting leads to a general-sum game.

Learning algorithms for stochastic Stackelberg games have been considered only for some special cases in the literature. For instance, [17] provides a value iteration-based algorithm to converge to the Stackelberg equilibrium in zero-sum stochastic games. Similarly, [30] presents a mixed integer program for computing the Stackelberg equilibrium strategy when the follower plays their deterministic best response stationary policy. In [31], policy gradient algorithms for stochastic Stackelberg games are introduced and studied numerically, albeit without a convergence analysis.

In this paper, we consider the Stackelberg policy gradient algorithms introduced in [31] and analyse conditions guaranteeing convergence to a Stackelberg equilibrium. In stochastic games, it is common to consider Markov stationary policies for both the players [11], [20], [30] and we make this assumption as well. While this restriction may be limiting (e.g., [30] showed that the Stackelberg equilibrium strategy for the leader need not be a Markov stationary policy even if the follower policy is a Markov stationary policy), this

P. Das and V. Gupta are with the Elmore Family School of Electrical and Computer Engineering at Purdue University. {das211,gupta869}@purdue.edu. B. Nortmann and T. Mylvagnam are with the Department of Aeronautics at Imperial College. {benita.nortmann15,t.mylvagnam}@imperial.ac.uk. L. J. Ratliff is with the Department of Electrical and Computer Engineering at University of Washington. {ratliffl}@uw.edu. The work was partially supported by ARO through grants W911NF2310111 and W911NF2310266, AFOSR through grant F.10052139.02.005, ONR through grant 13001274, and NSF through grants 2300355 and 2222097.

assumption considerably simplifies our convergence analysis. Our convergence analysis provides theoretical guarantees on the convergence of the proposed two time-scale Stackelberg policy gradient algorithms, and thus, verifies the numerical demonstration in [31]. Specifically, our analysis builds on formulating a smooth version of general-sum stochastic games where the value function of the follower is entropy regularized. We show that if we keep the leader policy fixed, the follower iterates obtained through gradient ascent converge to the unique optimal policy. For the leader, we prove that its policy iterates given that the follower is playing the best response converges to a stationary point. We then use the above two results to prove the convergence of the Stackelberg policy gradient algorithm when both the leader and the follower update their policies simultaneously. Since the Stackelberg policy gradient is a first order method, convergence to only a stationary point of the optimization problem for the leader is the best we can hope for without further assumptions on the cost function.

The paper is organized as follows. Section II formulates the game and the equilibrium notion considered. We present the proposed learning algorithm in Section III. Section IV presents the properties of the best response and value function that aid in the convergence analysis. Finally, the convergence result for the overall algorithm is presented in Section V. The appendix provides some technical details on the computation of the gradients in the algorithm.

II. Problem Formulation

a) Game definition: In this paper, we are interested in a two-player stochastic game with a hierarchical order of play. In these games, called stochastic Stackelberg games, two players—a leader and a follower—share a Markov decision process (MDP). The leader announces (i.e., commits to) their policy before the game begins. At every time step, both the players play actions corresponding to the state of the MDP in order to maximize their own infinite horizon discounted reward. However, there is a hierarchy between the two players in the sense that the follower *responds* to the policy that the leader has committed to.

Formally, a stochastic Stackelberg game is defined as a tuple $\mathbb{G} = \langle S, N, (A_1, A_2), \mathbb{P}, (\mathbb{R}_1, \mathbb{R}_2), \gamma, \rho \rangle$, where

- S is a finite set of states given by $S = \{s_1, s_2, ...s_k\}$ of the underlying MDP;
- $N = \{1,2\}$ is the set of players. Without loss of generality, we will assume that player 1 is the leader and player 2 is the follower;
- A_1 and A_2 are finite sets of actions of the leader and follower respectively;
- $\mathbb{P}$ is the probability transition function of the underlying MDP with $\mathbb{P}(s'|s, a_1, a_2)$ specifying the probability of the state at the next time step being s' given that the current state and joint actions are given by (s, a_1, a_2);
- $\mathbb{R}_i : S \times A_1 \times A_2 \to [0,1],\ i \in \{1,2\}$ are the reward functions of the leader and the follower;
- γ is a discount factor;
- $\rho \in \Delta(S)$ is the distribution of the state at time $t = 0$, also termed as the initial state s^0, with the probability of state $s \in S$ given by $\rho(s)$.

At each stage (or time) τ and corresponding state s^τ, the leader and the follower take joint actions (a_1^τ, a_2^τ). They then receive the rewards $\mathbb{R}_i(s^\tau, a_1^\tau, a_2^\tau)$ $(i = 1,2)$. The state transitions to $s^{\tau+1} \sim \mathbb{P}(s'|s^\tau, a_1^\tau, a_2^\tau)$. The stage $\tau + 1$ then begins.

b) Policies and Value Functions: Given a stochastic Stackelberg game, the set of strategies for the leader and the follower that we concentrate on in this paper are Markov stationary policies. Specifically, we define the policies for the two players as functions $\pi_i : S \times A_i \to \Delta_{A_i}, i \in \{1,2\}$ where Δ_{A_i} is a probability simplex in $|A_i|$ dimensions. We also denote the actions a_i^τ selected according to policy π_i by $a_i^\tau \sim \pi_i$.

The value functions $V_i(\pi_1, \pi_2)(s)$ for the two players can now be defined. For the leader, the value function for a given initial state s^0 at time $\tau = 0$ and given that the players play policies π_1, π_2 is given by

$$\bar{V}_1(\pi_1, \pi_2)(s) := \mathbb{E}_{\mathscr{T}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{R}_1(s^\tau, a_1^\tau, a_2^\tau) | s^0 = s\right] \quad (1)$$

where $\mathbb{E}_{\mathscr{T}}[.]$ denotes that the expectation is over all the trajectories $\mathscr{T} = \{s^\tau, a_1^\tau, a_2^\tau\}_{\tau \geq 0}$ with the actions $a_1^\tau \sim \pi_1, a_2^\tau \sim \pi_2$. Similarly, the value function for the follower can be defined as the expected discounted reward for player 2 as given by

$$\bar{V}_2(\pi_1, \pi_2)(s) := \mathbb{E}_{\mathscr{T}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{R}_2(s^\tau, a_1^\tau, a_2^\tau) | s^0 = s\right]. \quad (2)$$

As explained further in Section IV, for technical reasons, we consider an entropy regularized version of this value function as given by

$$\bar{V}_2^\lambda(\pi_1, \pi_2)(s) := \bar{V}_2(\pi_1, \pi_2)(s) + \lambda H_{\pi_2}(\pi_1, s) \quad (3)$$

with $H_{\pi_2}(\pi_1, s)$ defined as the discounted entropy of the follower policy π_2 when the leader policy is fixed to π_1, as given by

$$H_{\pi_2}(\pi_1, s) := \mathbb{E}_{\mathscr{T}}\left[\sum_{\tau=0}^{\infty} -\gamma^\tau \log \pi_2(a_2^\tau | s_\tau) | s^0 = s, \pi_1\right]. \quad (4)$$

We will also use the value functions with respect to an initial state distribution ρ rather than a given initial state. Thus, we define

$$V_1(\pi_1, \pi_2)(\rho) := \sum_{s \in S} \rho(s) \bar{V}_1(\pi_1, \pi_2)(s) = \mathbb{E}_{s_0 \sim \rho}[\bar{V}_1(\pi_1, \pi_2)(s)] \quad (5)$$

$$\begin{aligned} V_2^\lambda(\pi_1, \pi_2)(\rho) &:= \sum_{s \in S} \rho(s) \bar{V}_2^\lambda(\pi_1, \pi_2)(s) \\ &= \mathbb{E}_{s \sim \rho}[\bar{V}_2^\lambda(\pi_1, \pi_2)(s)]. \end{aligned} \quad (6)$$

c) Stackelberg Equilibrium: To define the Stackelberg equilibrium, we first need to define the best response or the rational reaction set of the follower. Suppose the leader commits to a policy π_1. Then, the best response of the

follower is the set of policies given by

$$B(\pi_1) := \underset{\pi_2}{\operatorname{argmax}} V_2(\pi_1, \pi_2)(\rho).$$

The problem of computing a Stackelberg equilibrium is a bi-level optimization problem [2], [12]. Specifically, the policy for the leader can be obtained through the upper level problem given by

$$\pi_1^\star \in \underset{\pi_1}{\operatorname{argmax}} V_1(\pi_1, B(\pi_1))(\rho),$$

while the follower's policy can be obtained through the lower level optimization problem given by

$$\pi_2^\star \in \arg\max_{\pi_2} V_2(\pi_1, \pi_2)(\rho).$$

The problem we consider in this paper is to provide a learning algorithm for the leader and the follower to update their policies and analyze conditions under which such an algorithm converges to the Stackelberg equilibrium.

III. Proposed Algorithm

In this paper, we consider a two time-scale gradient ascent algorithm for the leader and the follower to converge to the stationary policies that achieve the Stackelberg equilibrium. We begin by presenting the parametrization of the policies that we consider.

a) Parameterization of the policies: We will utilize the player policies given in terms of the commonly used soft max policy parameterization [1], [8], [21]. Specifically, the policy of each player i is determined using a parameter $\theta_i \in \mathbb{R}^{|S||A_i|}$. Denoting the element of θ_i corresponding to the state $s \in S$ and j^{th} action of player i $a_{i,j} \in A_i$ be $\theta_i(s, a_{i,j})$, the policy is given by specifying the probabilities of taking the action $a_{i,j}$ given the state s as

$$\pi_i(a_{i,j}|s) = \frac{e^{\theta_i(s,a_{i,j})}}{\sum_{a_{i,j} \in A_i} e^{\theta_i(s,a_{i,j})}}. \tag{7}$$

Since the parameters θ_i will evolve with time in the learning algorithm, we denote the value of the parameter at time t by θ_i^t.

As is well known [1], any stochastic policy can be represented using the soft max parameterization. To avoid notational clutter, we use θ_i and π_{θ_i} interchangeably to denote the policies of player i. Similarly, we use $V_1(\theta_1, \theta_2)$ (resp. $V_2^\lambda(\theta_1, \theta_2)$) and $V_1(\pi_{\theta_1}, \pi_{\theta_2})$ (resp. $V_2^\lambda(\pi_{\theta_1}, \pi_{\theta_2})$) interchangeably to represent the value functions as a function of θ_1, θ_2.

b) The gradient ascent algorithm: We consider the Stackelberg policy gradient algorithm from [31], which is a gradient ascent algorithm for both the leader and the follower. The algorithm is given through the following iterations of the policy parameters:

$$\theta_1^{t+1} = \theta_1^t + \alpha_t \nabla V_1(\theta_1^t, \theta_2^t)(\mu) \tag{8}$$

$$\theta_2^{t+1} = \theta_2^t + \beta_t \nabla_{\theta_2} V_2^\lambda(\theta_1^t, \theta_2^t)(\mu), \tag{9}$$

where α_t and β_t are the step sizes and $\nabla_{\theta_i} V_i^\lambda(.,.)(\mu)$ represents the gradient of the value function of player i with respect to its policy, for a given policy of the other player w.r.t to some initial distribution μ such that $\mu(s) > 0, \forall s \in S$. $\nabla V_1(\theta_1^t, \theta_2^t)(\mu)$ is defined in Proposition 3. We will make the following assumption on the stepsizes:

Assumption 1: The stepsizes satisfy

$$\sum_{t=0}^{\infty} \alpha_t = \sum_{t=0}^{\infty} \beta_t = \infty, \tag{10}$$

$$\sum_{t=0}^{\infty} \alpha_t^2 = \sum_{t=0}^{\infty} \beta_t^2 < \infty. \tag{11}$$

$$\lim_{t \to \infty} \frac{\alpha_t}{\beta_t} = 0. \tag{12}$$

This assumption implies that there is a time-scale separation between the learning dynamics of the leader and the follower captured by (12). Specifically, the condition (12) implies that the follower is learning at a faster rate compared to the leader and is a commonly made assumption in learning algorithms for Stackelberg games even in a static setting [13], [14]. In the rest of the paper, we show that this algorithm converges to a stationary point of the objective function of the leader. The time step or stages in the game during an iteration of the learning algorithm is given by τ and the iteration of the learning algorithm is given by t or n.

IV. Properties of Best Response and Value Function

We now proceed to analyze the best response of the follower given a specific policy followed by the leader and the value function of the leader when the follower plays the best response to the policy of the leader.

Recall that the policy at the follower evolves according to (9). With a given leader policy parameter θ_1 and hence the leader policy π_{θ_1}, we can convert the two player stochastic game into a MDP from the point of the view of the follower. To this end, define the averaged rewards and probability transition functions as

$$\mathbb{R}_2^{\pi_{\theta_1}}(s, a_2) := \sum_{a_1 \in A_1} \pi_{\theta_1}(a_1|s) \mathbb{R}_2(s, a_1, a_2) \tag{13}$$

$$\mathbb{P}^{\pi_{\theta_1}}(s'|s, a_2) := \sum_{a_1 \in A_1} \pi_{\theta_1}(a_1|s) \mathbb{P}(s'|s, a_1, a_2). \tag{14}$$

In other words, we average the rewards and the probability transition function for the original game with respect to the policy of the leader. The MDP defined as $\langle S, \mathbb{R}_2^{\pi_{\theta_1}}, A_2, \mathbb{P}^{\pi_{\theta_1}}, \gamma, \rho \rangle$ is then called an Averaged MDP for the follower [32].

The advantage of defining this averaged MDP is that we can compute $\nabla_{\theta_2} V_2^\lambda(\theta_1^t, \theta_2^t)$ in (9) at each time t by considering the follower as the only decision maker in this averaged MDP. This is simply because once we fix the leader policy, the gradient computation for the follower is identical in both these cases. We will use this fact to show that for any fixed policy followed by the leader, there exists a unique best response policy for the follower.

To this end, consider the averaged MDP $\langle S, \mathbb{R}_2^{\pi_{\theta_1}}, A_2, \mathbb{P}^{\pi_{\theta_1}}, \gamma, \rho \rangle$ with the follower seeking to optimize an entropy regularized version of its value function

over the choice of its policy $\bar{\pi}_{\bar{\theta}_2}$ through a gradient ascent algorithm of the form

$$\bar{\theta}_2^{t+1} = \bar{\theta}_2^t + \beta_t \nabla_{\bar{\theta}_2} \bar{V}_2^{\lambda}(\theta_1^t, \bar{\theta}_2^t), \tag{15}$$

where $\bar{V}_2^{\lambda}(\theta_1^t, \bar{\theta}_2^t)$ is the entropy regularized value function for this averaged MDP as defined by

$$\bar{V}_2^{\lambda}(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2})(\rho) = \bar{V}_2(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2})(\rho) + \lambda \bar{H}(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2}, \rho),$$
$$\bar{H}(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2}, \rho) = \mathbb{E}_{s_0 \sim \rho, \bar{a}_2^{\tau} \sim \bar{\pi}_{\bar{\theta}_2}(.|s_\tau), s_{\tau+1} \sim \mathbb{P}^{\pi_{\theta_1}}(.|s_\tau, \bar{a}_2^{\tau})} \left[\sum_{\tau=0}^{\infty} -\gamma^{\tau} \log \bar{\pi}_{\bar{\theta}_2}(a_2^{\tau}|s_\tau) \right].$$

The following result follows from those in the existing literature [16], [21] and aids in the study of the properties such as uniqueness and continuity of the best response of the follower.

Proposition 1: Consider the averaged MDP defined above with the follower utilizing a gradient ascent algorithm of the form (15) for a given π_{θ_1}. The following statements are true:

1) There is a unique optimal policy $\bar{\pi}_2^{\star}$ for the follower [16, Proposition 1].
2) The gradient ascent algorithm asymptotically converges to $\bar{\pi}_2^{\star}$. In particular, let $\mu(s) > 0, \forall s \in S$ and $\beta_t = \frac{(1-\gamma)^3}{8+\lambda(4+8\log A)} \forall t$, then $\exists\ C > 0$ such that $\forall\ t \geq 1$ [21, Theorem 6],

$$\bar{V}_2^{\lambda}(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2^{\star}})(\rho) - \bar{V}_2^{\lambda}(\pi_{\theta_1}, \bar{\pi}_{\bar{\theta}_2^t})(\rho) \leq \| \frac{1}{\mu} \|_{\infty} \frac{1 + \lambda \log A}{(1-\gamma)^2} e^{-C(t-1)}. \tag{16}$$

Remark 1: Since the follower policy converges to a unique optimal policy for the averaged MDP, this implies that in the Stackelberg game defined in Section II, for a fixed policy of the leader, the follower has a unique optimal response. In other words, the best response map in the game is a function.

Remark 2: While the optimal policy $\bar{\pi}_2^{\star}$ is unique as shown in Proposition 1, there might be multiple parameters $\{\bar{\theta}_2(s, a_{2,j})\}_{s \in S, a_{2,j} \in A_2}$ that yield the same optimal policy. The uniqueness of $\{\bar{\theta}_2(s, a_{2,j})\}_{s \in S, a_{2,j} \in A_2}$ can be ensured by following the procedure of [8, Example 5]. We assume that this procedure is followed and hence the parameter is also unique.

We now make the following assumption for further development.

Assumption 2: There exists a value $0 < \lambda < \infty$ such that $\nabla^2_{\theta_2} V_2^{\lambda}(\theta_1, B(\theta_1))^{-1}$ exists $\forall \theta_1 \in \mathbb{R}^{|S||A_1|}$.

Under assumption 2, the next result proves the continuity and differentiability of the best response.

Proposition 2: The best response function $B(\theta_1)$ is continuous and differentiable in θ_1.

Proof: The unique optimal policy of the follower (say $\theta_2^{'}$) for fixed leader policy θ_1 is a stationary point $\nabla_{\theta_2} V_2^{\lambda}(\theta_1, \theta_2^{'}) = 0$. Under Assumption 2, the implicit function theorem [24] implies that the best response function $B : X \subset \mathbb{R}^{|S||A_1|} \rightarrow \mathbb{R}^{|S||A_2|}$ is continuous and differentiable in θ_1. ∎

Remark 3: The implicit function theorem also provides an explicit formula to compute the derivative $\nabla B(\theta_1)$. Specifically, applying the implicit function theorem to the implicit function $\nabla_{\theta_2} V_2^{\lambda}(\theta_1, \theta_2) = 0$, the derivative of the best response function with respect to the parameter of the leader is given by

$$\nabla B(\theta_1) = -\nabla_{\theta_1, \theta_2} V_2^{\lambda}(\theta_1, B(\theta_1)) \left[\nabla^2_{\theta_2} V_2^{\lambda}(\theta_1, B(\theta_1)) \right]^{-1}. \tag{17}$$

We also consider the following regularity assumption on the derivative of the best response function.

Assumption 3: There exists $0 < \lambda < \infty$ and $0 < K < \infty$ such that

$$\| \nabla_{\theta_1, \theta_2} V_2^{\lambda}(\theta_1, \theta_2) \nabla^2_{\theta_2} V_2^{\lambda}(\theta_1, \theta_2)^{-1} \|_2 \leq K, \forall (\theta_1, \theta_2).$$

We can now state the following result.

Lemma 1: Under Assumption 3, $B(\theta_1)$ is Lipschitz.

Proof: The proof follows from the mean value theorem for vector valued functions and the fact that $\|\nabla B(\theta_1)\|_2$ is bounded by Assumption 3. ∎

The leader is optimizing its policy for the objective function $V_1(\theta_1, B(\theta_1))$. We next present properties of the function $V_1(\theta_1, B(\theta_1))$ that will aid in our convergence analysis.

Proposition 3: The function $V_1(\theta_1, B(\theta_1))$ is a continuous function and differentiable in θ_1. Further, its derivative is given by

$$\nabla V_1(\theta_1, B(\theta_1)) = \nabla_{\theta_1} V_1(\theta_1, B(\theta_1)) - \nabla_{\theta_2} V_1(\theta_1, B(\theta_1)) \nabla_{\theta_1, \theta_2} V_2^{\lambda}(\theta_1, B(\theta_1)) \left[\nabla^2_{\theta_2} V_2^{\lambda}(\theta_1, B(\theta_1)) \right]^{-1}. \tag{18}$$

Proof: From Proposition 2, the map $g(\theta_1) = (\theta_1, B(\theta_1))$ is a continuous and differentiable function in θ_1. The value function $V_1(\theta_1, \theta_2)$ is continuous [1] and differentiable in (θ_1, θ_2). Thus, taking the composition of two differentiable functions, we obtain that $V_1(g(\theta_1)) = V_1(\theta_1, B(\theta_1))$ is a continuous and differentiable function in θ_1. ∎

At the n^{th} iteration of the two-timescale algorithm, (say (θ_1^n, θ_2^n)), as the follower has not converged to their best response $B(\theta_1^n)$, the leader updates its policy according to (8) where the gradient is given by (19)

$$\nabla V_1(\theta_1^n, \theta_2^n) = \nabla_{\theta_1} V_1(\theta_1^n, \theta_2^n) - \nabla_{\theta_2} V_1(\theta_1^n, \theta_2^n) \nabla_{\theta_1, \theta_2} V_2^{\lambda}(\theta_1^n, \theta_2^n) \left[\nabla^2_{\theta_2} V_2^{\lambda}(\theta_1^n, \theta_2^n) \right]^{-1}. \tag{19}$$

We will use the ODE method as presented in [9], [10] to prove convergence of Stackelberg policy gradient. To this extend, we will make two additional assumptions (similar to [14]) before presenting our main result.

Assumption 4: The iterates of the Stackelberg policy gradient are bounded i.e. $\sup_t (\|\theta_1^t\| + \|\theta_2^t\|) < \infty$.

Assumption 5: The only internally chain transitive invariant sets of the differential equation $\dot{\theta}_1(t) = \nabla V_1(\theta_1(t), B(\theta_1(t)))$ are isolated equilibrium points.

V. Convergence of Algorithm

In this section, we introduce the main result of this paper that provides the conditions guaranteeing the convergence of

the proposed two-timescale policy gradient algorithm given by the parameter iterations (8) and (9) to a stationary point of the objective function of the leader. The convergence result is now stated.

Theorem 1: Consider the problem formulated in Section II. Under Assumptions 1-5, the Stackelberg policy gradient algorithm (8,9) converges to policies (θ_1^*, θ_2^*) where $\theta_2^* = B(\theta_1^*)$ and θ_1^* is a stationary point of $V_1(\theta_1, B(\theta_1))$.

Proof: The Stackelberg policy gradient algorithm can be re-written as

$$\begin{pmatrix}\theta_1^{t+1}\\ \theta_2^{t+1}\end{pmatrix} = \begin{pmatrix}\theta_1^{t}\\ \theta_2^{t}\end{pmatrix} + \begin{pmatrix}\alpha_t \nabla V_1(\theta_1^t, \theta_2^t)(\mu)\\ \beta_t \nabla_{\theta_2} V_2^{\lambda}(\theta_1^t, \theta_2^t)(\mu)\end{pmatrix} \tag{20}$$

$$\implies \begin{pmatrix}\theta_1^{t+1}\\ \theta_2^{t+1}\end{pmatrix} = \begin{pmatrix}\theta_1^{t}\\ \theta_2^{t}\end{pmatrix} + \beta_t \begin{pmatrix}\frac{\alpha_t}{\beta_t} \nabla V_1(\theta_1^t, \theta_2^t)(\mu)\\ \nabla_{\theta_2} V_2^{\lambda}(\theta_1^t, \theta_2^t)(\mu)\end{pmatrix}. \tag{21}$$

The condition $\lim_{t\to\infty} \frac{\alpha_t}{\beta_t} = 0$ induces a time-scale separation between the dynamics of the leader and follower. Under Assumption 4, [9, Corollary 2.1] the iterates of the algorithm $(\theta_1^t, \theta_2^t) \to H = \{(\theta_1, B(\theta_1)) : \theta_1 \in \mathbb{R}^{|S||A_1|)}\}$ as $t \to \infty$. Under Assumption 4, [10, Chapter 6, Theorem 2] the iterates θ_1^t of the leader track the solutions of the differential equation given by

$$\dot{\theta}_1(t) = \nabla V_1(\theta_1(t), B(\theta_1(t))). \tag{22}$$

Furthermore, under Assumption 5, [10, Chapter 2, Corollary 4] the iterates θ_1^t converges to a sample path dependent equilibrium point of the differential equation (22) Thus, the iterates (θ_1^t, θ_2^t) converge to policies (θ_1^*, θ_2^*) where $\theta_2^* = B(\theta_1^*)$ and θ_1^* is a stationary point of $V_1(\theta_1, B(\theta_1))$ i.e.$\nabla V_1(\theta_1^*, B(\theta_1^*)) = 0$. ∎

Remark 4: The condition $\lim_{t\to\infty} \frac{\alpha_t}{\beta_t} = 0$ induces a time-scale separation. Asymptotically, from the view of the leader, the iterates of the leader appear static and from the view of the leader, the iterates of the follower have converged to their best response given the leader policy.

Remark 5: We have shown convergence of the Stackelberg policy gradient converge on the set $\{\sup_t(\|\theta_1^t\| + \|\theta_2^t\|) < \infty\}$. In other words, the algorithm will converge when all the stationary points of $V_1(\theta_1, B(\theta_1))$ are stochastic policies.

APPENDIX I
SMOOTHNESS OF THE VALUE FUNCTIONS

The gradient $\nabla_{\theta_1} V_1(\theta_1, \theta_2)$ is Lipschitz continuous in θ_1 and the gradient $\nabla_{\theta_2} V_2^{\lambda}(\theta_1, \theta_2)$ is Lipschitz continuous in θ_2 [1], [21]. We will prove the smoothness of $V_1(\theta_1, \theta_2)$ in θ_2. Fix θ_1. Let $\pi_\alpha := \pi_{\theta_2 + \alpha u}$ and

$$\tilde{V}(\alpha) = V_1^{\pi_{\theta_1}, \pi_\alpha}(s_0) = \sum_{a\in A} \pi_{\theta_1}(a|s) Q_1^{\pi_{\theta_1}, \pi_\alpha}(s_0, a).$$

We proceed as follows:

$$\begin{aligned}\frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\Big|_{\alpha=0} &= \sum_{a_1\in A} \mathbb{P}(s'|s,a_1,a) \frac{d\pi_\alpha(a_1|s)}{d\alpha}\Big|_{\alpha=0}\\ \implies \Big|\frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\Big|_{\alpha=0}\Big| &\le \sum_{a_1\in A}\Big|\frac{d\pi_\alpha(a_1|s)}{d\alpha}\Big|_{\alpha=0}\Big|\\ &\le c_1 = 2,\end{aligned}$$

where the last equality follows from [1]. Similarly, observe that

$$\begin{aligned}\frac{d^2\mathbb{P}^\alpha(s'|s,a)}{(d\alpha)^2}\Big|_{\alpha=0} &= \sum_{a_1\in A} \mathbb{P}(s'|s,a_1,a) \frac{d^2\pi_\alpha(a_1|s)}{(d\alpha)^2}\Big|_{\alpha=0}\\ \implies \Big|\frac{d^2\mathbb{P}^\alpha(s'|s,a)}{(d\alpha)^2}\Big|_{\alpha=0}\Big| &\le \sum_{a_1\in A}\Big|\frac{d^2\pi_\alpha(a_1|s)}{(d\alpha)^2}\Big|_{\alpha=0}\Big|\\ &\le c_2 = 6,\end{aligned}$$

where again the last equality follows from [1]. Let $\tilde{P}(\alpha)_{(s,a)\to(s',a')} = \pi_{\theta_1}(a'|s')\mathbb{P}^\alpha(s'|s,a)$ and $\frac{d\tilde{P}(\alpha)}{d\alpha}\big|_{\alpha=0} = \pi_{\theta_1}(a'|s')\frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\big|_{\alpha=0}$. For any arbitrary vector x, we conclude that

$$\begin{aligned}\Big[\frac{d\tilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} x\Big] &= \sum_{(s',a')} \frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\Big|_{\alpha=0} x_{a',s'} \pi_{\theta_1}(a'|s')\\ &\implies \max_{\|u\|_2=1}\Big|\Big[\frac{d\tilde{P}(\alpha)}{d\alpha}\Big|_{\alpha=0} x\Big]_{s,a}\Big|\\ &= \max_{\|u\|_2=1}\Big|\sum_{(s',a')} \frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\Big|_{\alpha=0} x_{a',s'} \pi_{\theta_1}(a'|s')\Big|\\ &\le \|x\|_\infty \sum_{s'}\Big|\frac{d\mathbb{P}^\alpha(s'|s,a)}{d\alpha}\Big|_{\alpha=0}\Big| \le c_3\|x\|_\infty. \end{aligned} \tag{23}$$

Similarly, we have that

$$\max_{\|u\|_2=1}\Big|\Big[\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0} x\Big]_{s,a}\Big| \le c_4\|x\|_\infty. \tag{24}$$

Let $M(\alpha) = (I - \gamma\tilde{P}(\alpha))^{-1} = \sum_{n=1}^{\infty} \gamma^n \tilde{P}(\alpha)^n$. The above equation is obtained from the power series expansion where $\tilde{P}(\alpha)^n$ is a stochastic matrix for each n so that $M(\alpha)\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$. Hence, we deduce that

$$\max_{\|u\|_2=1} \|M(\alpha)x\|_\infty \le \frac{1}{1-\gamma}\|x\|_\infty. \tag{25}$$

Moreover, we have $R(\alpha)_{s_0,a_0} = \sum_{a\in A_2} \pi_\alpha(a_0|s) R(s_0, a_0, a)$ implies that

$$\|\frac{dR(\alpha)}{d\alpha}\|_\infty \le c_4 \quad \text{and} \quad \|\frac{d^2R(\alpha)}{(d\alpha)^2}\|_\infty \le \quad c_5.$$

Let $Q^\alpha(s,a)$ be the Q-function of player one when they follow policy π_{θ_1} and player two follows policy π_α. Then, we write the $Q^\alpha(s_0,a_0)$ as

$$\begin{aligned}Q^\alpha(s_0,a_0) &= e_{(s_0,a_0)}^T (I - \gamma\tilde{P}(\alpha))^{-1} R(\alpha)\\ &= e_{(s_0,a_0)}^T M(\alpha) R(\alpha)\end{aligned}$$

using the Bellman equation. This implies that

$$\frac{d^2Q^\alpha(s_0,a_0)}{(d\alpha)^2} = 2\gamma^2 e^T_{(s_0,a_0)}M(\alpha)\frac{d\tilde{P}(\alpha)}{d\alpha}M(\alpha)\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}M(\alpha)R(\alpha) + \gamma e^T_{(s_0,a_0)}M(\alpha)\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}M(\alpha)R(\alpha) + \gamma e^T_{(s_0,a_0)}M(\alpha)\frac{d\tilde{P(\alpha)}}{d\alpha}M(\alpha)\frac{dR(\alpha)}{d\alpha} + e^T_{(s_0,a_0)}M(\alpha)\frac{d^2R(\alpha)}{(d\alpha)^2}. \quad (26)$$

Since the infinity norm of each of the matrices are bounded, $\max_{\|u\|_2=1}\left|\frac{d^2Q^\alpha(s_0,a_0)}{(d\alpha)^2}\right|\alpha=0\Big|$ is bounded.

The value function for the player 1 is given by

$$\tilde{V}(\alpha) = \sum_{a\in A}\pi_{\theta_1}(a|s)Q^\alpha(s_0,a)$$

$$\implies \frac{d^2V(\alpha)}{d\alpha^2} = \sum_{a\in A}\pi_{\theta_1}(a|s)\frac{d^2Q^\alpha(s_0,a)}{d\alpha}$$

$$\implies \max_{\|u\|_2=1}\left|\frac{d^2V(\alpha)}{d\alpha^2}\right|\alpha=0 \leq |A|\max_{\|u\|_2=1}\left|\frac{d^2Q^\alpha(s_0,a)}{d\alpha}\right|\alpha=0$$

Since $\max_{\|u\|_2=1}\left|\frac{d^2Q^\alpha(s_0,a_0)}{(d\alpha)^2}\right|\alpha=0\Big|$ is bounded, $\max_{\|u\|_2=1}\left|\frac{d^2V(\alpha)}{d\alpha^2}\right|\alpha=0$ is also bounded. This completes the proof.

REFERENCES

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

[2] Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[3] Nicola Basilico and Nicola Gatti. Automated abstractions for patrolling security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1096–1101, 2011.

[4] Nicola Basilico, Nicola Gatti, Francesco Amigoni, et al. Leader-follower strategies for robotic patrolling in environments with arbitrary topologies. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pages 57–64, 2009.

[5] Nicola Basilico, Nicola Gatti, and Federico Villa. Asynchronous multi-robot patrolling against intrusions in arbitrary topologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1224–1229, 2010.

[6] Nicola Basilico, Davide Rossignoli, Nicola Gatti, and Francesco Amigoni. A game-theoretical model applied to an active patrolling camera. In *2010 International Conference on Emerging Security Technologies*, pages 130–135. IEEE, 2010.

[7] Michel Benaım and Morris W Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2):36–72, 1999.

[8] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[9] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[10] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[11] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

[12] Stephan Dempe. Bilevel optimization: theory, algorithms, applications and a bibliography. *Bilevel Optimization: Advances and Next Challenges*, pages 581–672, 2020.

[13] Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*, pages 3133–3144. PMLR, 2020.

[14] Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

[15] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[16] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

[17] Denizalp Goktas, Sadie Zhao, and Amy Greenwald. Zero-sum stochastic stackelberg games. *Advances in Neural Information Processing Systems*, 35:11658–11672, 2022.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[19] Debarun Kar, Thanh H Nguyen, Fei Fang, Matthew Brown, Arunesh Sinha, Milind Tambe, and Albert Xin Jiang. Trends and applications in stackelberg security games. *Handbook of dynamic game theory*, pages 1–47, 2017.

[20] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

[21] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[22] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms*, pages 2703–2717. SIAM, 2018.

[23] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[24] Jet Nestruev, AV Bocharov, and S Duzhin. *Smooth manifolds and observables*. Springer, 2003.

[25] Christos Papadimitriou and Georgios Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*, 16(2):53–63, 2019.

[26] Georgios Piliouras and Jeff S Shamma. Optimization despite chaos: Convex relaxations to complex limit sets via poincaré recurrence. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 861–873. SIAM, 2014.

[27] Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. In *International conference on machine learning*, pages 7953–7963. PMLR, 2020.

[28] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[29] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security games: Looking beyond a decade of success. IJCAI, 2018.

[30] Yevgeniy Vorobeychik and Satinder Singh. Computing stackelberg equilibria in discounted stochastic games (corrected version). In *Twenty-Sixth Conference on Artificial Intelligence (AAAI)*, 2012.

[31] Quoc-Liem Vu, Zane Alumbaugh, Ryan Ching, Quanchen Ding, Arnav Mahajan, Benjamin Chasnov, Sam Burden, and Lillian J Ratliff. Stackelberg policy gradient: Evaluating the performance of leaders and followers. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

[32] Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021.

[33] Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J Ratliff. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9217–9224, 2022.