Policy Convergence Under the Influence of Antagonistic Agents in Markov Games

Chase P. Dowling*
Pacific Northwest National Laboratory
Seattle, WA 98109
chase.dowling@pnnl.gov

Ted Fujimoto
Pacific Northwest National Laboratory
Richland, WA 99354
ted.fujimoto@pnnl.gov

Nathan Hodas
Pacific Northwest National Laboratory
Richland, WA 99354
nathan.hodas@pnnl.gov

Abstract

We propose an empirical study of how an antagonistic agent can subvert centralized multi-agent learning and prevent policy convergence in the context of Markov games. We hypothesize that recent results in differential Nash equilibria are candidate conditions for policy convergence when agents learn via policy gradient methods, and motivate this line of inquiry by introducing antagonistic agents. We compare the performance of groups of agents with prescribed reward function forms with and without antagonistic opponents and measure policy convergence. This experiment aims to steer the research community towards a promising plan-of-attack in addressing open questions in Markov games.

1 Introduction

We will investigate policy convergence in a non-cooperative **m**ulti-**a**gent MDP utilizing **d**eep, **d**eterministic **p**olicy **g**radient (MADDPG) [8], a class of Markov game [7] where learning agents simultaneously perform policy gradient via a centralized actor-critic method. To motivate this line of inquiry and place the proposed work in a practical context, we introduce an *antagonistic agent rewarded by the negative of victim agent(s) rewards but does not share an identical action space to the victim agents*. Recent work in game theory has proven in multi-agent gradient learning environments with continuous action spaces agents can converge to policy limit cycles, or not converge at all, depending on the joint gradient flow of the agents' reward functions. Hence, the introduction of an antagonistic agent in a MADDPG setting could prevent the group from converging to individual or group optimal policies, or even any policy at all.

Open Questions For a Markov game where all agents learn via policy gradient, in the presence of an antagonistic agent whose objective function is to maximize reward penalties to some subset of target victims over a given time horizon, the following are open questions:

1. Under what conditions does the group's joint policy converge to a stable equilibrium?
2. What is the cumulative reward cost incurred by the presence of an antagonistic agent?
3. Under what conditions is the presence of an antagonistic agent detectable?

*cpatdowling.github.io

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

An initial experimental study is well-suited to orienting a theoretical plan-of-attack at addressing these open questions in subsequent work; question 1 is particularly challenging (see Remark 15 [9]). In section 2, we provide an overview of previous work in multi-agent RL and Markov games. In section 3, we define an antagonistic agent in a MADDPG setting. In section 4, we state our hypothesis and propose our experimental protocol that measures the impact of the antagonist on the group, and verify the extent to which victim agents converge to potentially sub-optimal behavior (such as limit cycles in agents' policies) in the presence of an antagonist, as well as provide the intuition for this line of questioning given some simple, initial results based on recent work.

2 Background

Markov games [7] occupy the intersection of multi-agent Markov decision processes (MDP) and matrix games [2, 13]. Game theory's *theoretical* overlap with more recent work in methods for both on- and off-policy training of multi-agent MDP's, however, remains open for further theoretical and empirical exploration [9, 8, 5, 10]. Laying the groundwork by formalizing the relationship between stochastic games and MDP's in [2], the work of Bowling has studied joint policy convergence for agents utilizing mixed learning strategies in MDP's [4] and convergence to mixed equilibria for special classes of games for which optimal strategies (read policies) are probabilistic [3]. What remains open for exploration, however, is general conditions for policy convergence in multi-agent MDP's (as opposed to the construction of policy learning algorithms that *must* converge) [4, 10].

Recent work by Mazumdar et al. [9] give necessary and sufficient conditions for convergence to a so-called differential Nash equilibrium in stateless games where competing agents improve their strategy according to a (stochastic) gradient update on their individual cost functions. The authors show that gradient updates yield a gradient flow in costs with respect to the joint policy space, computed as the game Jacobian; and policy divergence or convergence to a stable point or limit cycle according to the eigenvalues of the game Jacobian. Thus, conditions for convergence for off-policy gradient methods in multi-agent MDP's like MADDPG may be more appropriately generalized than originally in [4], where—intuitively—limit cycle behavior can be observed for certain learning algorithms and deterministic policies in games such as Rock-Paper-Scissors [4]. To wit, the authors of [8] state, "agents can derive a strong policy by overfitting to the behavior of their competitors. Such policies are undesirable as they are brittle and may fail when the competitors alter strategies". The concept of Nash equilibrium is particularly suited to the analysis of such cases.

We treat an "antagonistic" agent as distinct from "adversarial" [6], providing a specific definition in section 3. While sometimes referring to non-cooperative Markov games [10], usage of "adversarial" in RL varies, often referring to exogenous environmental factors that are outside the control of the agents [11, 1]. As learning-enabled systems become prevalent in decentralized and non-cooperative infrastructural systems in future built environments (e.g. autonomous vehicles or smart-grid participants like E.V. charging stations), analysis of reward function design to render agents robust to antagonistic participants *in addition to* adversarial environments will be critical in minimizing social costs incurred in decentralized infrastructure.

3 Methodology

In this work we focus on the off-policy version of multi-agent deep deterministic policy gradient (MADDPG) [8], an actor-critic method with a centralized critic. In MADDPG, N actors or agents share a set of states $\mathcal{S}$, with individual action sets $\mathcal{A}_1, \ldots, \mathcal{A}_N$ and observation sets $\mathcal{O}_1, \ldots, \mathcal{O}_N$. The observation sets track both the state of the system s_t as well as competing agent actions. An individual agent implements a stochastic policy $\pi_{\theta_i} : \mathcal{O}_i \times \mathcal{A}_i \to [0, 1]$ and obtains a reward $r_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$. Let $o_i \in \mathcal{O}_i$, $a_i \in \mathcal{A}_i$, and $\boldsymbol{o} = \langle o_1, \ldots, o_N \rangle$, $\boldsymbol{a} = \langle a_1, \ldots a_n \rangle$, and so on for $\boldsymbol{\theta}$, and $\boldsymbol{\pi}$. Each agent i aims to maximize its own total discounted expected return over time horizon T. In MADDPG, the centralized critic learns a action-value function $Q_i^{\boldsymbol{\pi}}(\boldsymbol{o}, \boldsymbol{a})$ for each agent $i \in 1, \ldots, N$. Each agent performs the actor update to the policy parameterization θ_i,

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{\boldsymbol{o}, a \sim \mathcal{D}} \left[\nabla_{a_i} Q_i^{\boldsymbol{\pi}}(\boldsymbol{o}, \boldsymbol{a}) \nabla_{\theta_i} \pi_{\theta_i}(a_i | o_i) \right] \tag{1}$$

where the set $\mathcal{D}$ records the observations, actions, and rewards received by the agents thus far. Concurrently the critic updates the estimate the action-value function according to the loss function

$$\mathcal{L}(\theta_i) = \mathbb{E}_{\boldsymbol{o},\boldsymbol{a}}\left[(Q_i^{\boldsymbol{\pi}}(\boldsymbol{o},\boldsymbol{a}) - y)^2\right] \tag{2}$$

where $y = r_i + \gamma Q_i^{\boldsymbol{\pi}}$. In off-policy MAADPG, $Q_i^{\boldsymbol{\pi}}$, the value function of all agents in the system must be estimated by every agent in addition to the critic. We make use the policy inference scheme in [8], where each agent maintains an approximation of competing agents' policies $\hat{\boldsymbol{\pi}}_{\phi_i^j}$ where ϕ_i^j is the ith agent's estimated policy parameterization for the jth agent, which we will denote $\hat{\boldsymbol{\pi}}$ in our results.

We now define an antagonistic agent. Let $V_i^{\pi_i} : \mathcal{S} \rightarrow \mathbb{R}$, be the state-value function, where $V_i^{\boldsymbol{\pi}}(s) = \Sigma_{a\in\mathcal{A}_i}\pi_{\theta_i}(a_i|s)Q_i^{\boldsymbol{\pi}}(\boldsymbol{o},\boldsymbol{a})$, the expectation of the action-value over the agent's policy. Additionally let $\Delta V_{i,t}^{\boldsymbol{\pi}} = V_i^{\boldsymbol{\pi}}(s_{t+1}) - V_i^{\boldsymbol{\pi}}(s_t)$.

Definition 3.1. The *Antagonistic Value Function* for antagonistic agent k with victim j at a state s_t under a policy $\boldsymbol{\pi}$ is defined as:

$$V_k^{\boldsymbol{\pi}}(s_t) = -\mathbb{E}_{\hat{\pi}_{\phi_k^j}}\left[\sum_t^T \gamma^t \Delta V_{j,t}^{\boldsymbol{\pi}} \middle| s_t\right] \tag{3}$$

where γ_k is the discount rate. In words, the antagonist, while maintaining an estimate of the victim agent's action-value function across states, aims to simply maximize the negative of the expected discounted reward over the remaining time horizon of the shared MDP according to the antagonist's estimate of the victim's policy. All other agents learn from predefined reward functions r_i (which we constrain in the experimental protocol) while maintaining estimates of competing agents' policies (in which an antagonistic agent may be present).

3.1 Experimental Intuition: Differential Nash equilibria

Recent work [9] has demonstrated necessary and sufficient conditions for the existence of so-called differential Nash equilibria when, compared to a canonical MADDPG, 1) multiple agents share a single fixed state, and each have a continuous action space, 2) act on continuously differentiable *cost* functions of the joint actions of all agents, and 3) improve their individual policy performance according to gradient descent[2].

Adapting the notation of [9], consider a set of competing gradient learners $\mathcal{I} = \{1, \dots N\}$, with reward functions $C_i(x_i, x_{\neg i})$ where $C : \mathcal{X} \rightarrow \mathbb{R}$ for $\mathcal{X} \subset \mathbb{R}^{d_i}$. In words, each agents cost function is at least a function of their own choice x_i and some number of other agents' choices $x_{\neg i}$ where $\neg i$ denotes the agents other than i. For our purposes here, we will assume that C is at least twice continuously differentiable, and that each player's action space $\mathcal{X}$ is a connected open set.

Each agent updates a deterministic policy according to the gradient along $-\eta\frac{\partial}{\partial x_i}C_i(x_{i,t}, x_{\neg i,t})$, specifically the derivative of their cost function along their own action space. Because each agent is only updated with respect to the derivatives of their own strategy, we can define the joint strategy (and thus the input to each agent's cost function) to be the vector $\boldsymbol{x} := \langle x_1, \dots x_N\rangle$.

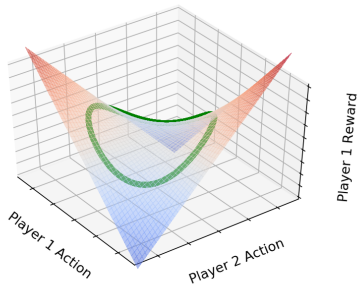
Definition 3.2. The gradient of the joint strategies of all agents costs is,

$$\nabla C = \left\langle \frac{\partial C_1}{\partial x_1}, \dots, \frac{\partial C_N}{\partial x_N} \right\rangle \tag{4}$$

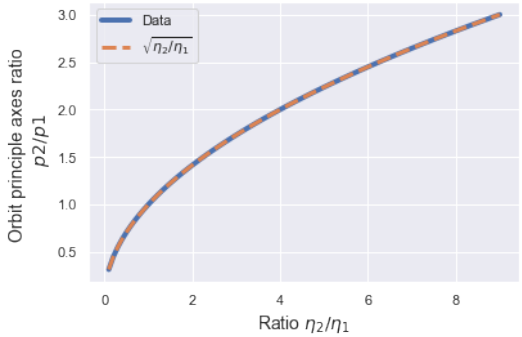
the joint strategies update (in the limit of the gradient descent learning rate) with each iteration t of each agent's policy gradient according to the differential equation,

$$\partial \boldsymbol{x}_{t+1} = -\nabla C(\boldsymbol{x}_t). \tag{5}$$

[2]The authors of [9] use cost, rather that reward functions, such that gradient improvements are descending rather than ascending—for consistency with theoretical results stated in [9] we utilize their notation, rather that reformatting for reward functions and ascending gradient improvement



(a) Periodic orbit induced by antagonistic agent influencing victim's gradient learning trajectory on agent 1's cost surface.



(b) Ratio of principle axes of orbit projected onto the action space $X \times Y$, as a function of ratio of agent learning rates.

By our differentiability assumptions, we can examine the critical points of (5) by computing the Jacobian $\nabla^2 C$. With these quantities, we can state a primary result of [9].

Definition 3.3. (Mazumdar et al. 2020) A point $x \in X$ is a locally asymptotically stable equilibirum of the continuous time dynamics (5) if $\nabla C(\boldsymbol{x}) = 0$ and $\text{Re}(\lambda) > 0 \; \forall \lambda \in \text{spec}(\nabla^2 C)$

Definition (3.3) are necessary and sufficient conditions for an attracting differential Nash equilibrium for both deterministic and stochastic gradient learners in what can be considered a gradient flow game. If any $\text{Re}(\lambda) < 0$, then a critical point is a saddle point. Under stochastic gradient descent, this is known to be a divergent critical point [12]. These results do not presuppose the presence of an antagonistic agent, however, it provides us with an analytical scaffold upon which to clearly define an antagonistic agent and observe the effect.

Definition 3.4. An antagonistic agent in a gradient flow game is an agent with a set of victims $V \subseteq I$ performing deterministic policy gradient according to the cost function, $C_{h(V)} = -\sum C_i$. Assign a single antagonistic agent the Nth index. We have that the $\nabla C = \left\langle \ldots, \frac{-\partial \Sigma C_{h(V)}}{\partial x_N} \right\rangle$, and similarly for the Jacobian.

Two agent case To illustrate the emergent dynamics we will hypothesize exist in *stateful* Markov games, we will focus on the 2 player case of a gradient flow game, with one hangry agent with action $y \in Y$, and one victim with action $x \in X$, and subsequent reward functions $f(x, y)$ and $C_h(x, y)$. In the 2 player case we can state a simple lemma on the stability of periodic equilibrium in gradient learning dynamics.

Lemma 1. If, $\forall x, y \in X, Y$, $\left[\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2}\right]^2 - 4\frac{\partial^2 f}{\partial x \partial x} < 0$, then the eigenvalues λ of the Jacobian are complex, the gradient learning dynamics may converge to a periodic orbit.

By definition of an antagonistic agent, $C_h(x, y) = -f(x, y)$. The inequality is recovered from the discriminant of the characteristic polynomial of the Jacobian, but depends on the dominance of $|\lambda|$ or η (updates to the joint policy spiral outwards if $|\lambda| >> \eta$). We can show empirically an elliptical limit cycle has major and minor axes proportional to the agents with larger and smaller learning rates, respectively. Consider the cost functions, $C_1 = x + xy$ and $C_2 = -x - xy$, where the antagonistic agent 2 attempts to maximize the costs of agent 1 with respect to their actions y. Figure 1a plots the trajectory of each player's deterministic policy on the cost surface of player 1. Figure 1b plots the ratio of the player's learning rates η_1 (victim) and η_2 (antagonist) against the ratio of the length of the orbit's principle axes p_1 (victim) and p_2 (antagonist). We conjecture, as evidenced by Fig. 1b, that independent of cost functions, $\frac{p_2}{p_1} = \sqrt{\frac{\eta_2}{\eta_1}}$.

Relationship to MADDPG Extending to MADDPG learning environments introduces several additional factors: Firstly, actions are discretized. Second, policies are state-dependent, so any analysis of a policy convergence must be conditioned on state transitions. For stochastic policies, we can measure the KL divergence of consecutive policy updates measured at each time t during

training epochs, $KL(\pi_{\theta_i,t}|\pi_{\theta_i i,t-1})$ *conditioned* on each possible state transition. Additionally, the time horizon is finite, thus limiting behavior may not be observed, particularly if agents are required to estimate opponent value functions and the state space is large. We address these technical leaps in our experimental plan.

4 Experimental Protocol

Hypothesis Can we observe convergent, limit cycle, and divergent policies, conditioned on a specific state, learned in a MADDPG environment if reward functions satisfy conditions in [9] for each critical point type on otherwise continuous action spaces?

Note the hypothesis does not depend on the presence of an antagonistic agent. Confirming or rejecting this hypothesis provides evidence and impetus for theoretical investigation for the three previously listed open questions, subject to the presence of a well-defined antagonistic agent.

Question 1: If the hypothesis is confirmed, then the conditions in [9] are likely a candidate special case of a more general sense of gradient flow in Markov games.

Question 2: Assume the hypothesis is confirmed. If policies conditioned on state converge to a limit cycle, our evidence suggest the limit cycle depends on hyperparameters that influence the gradient step-size; we will observe this. If policies conditioned on state converge to a fixed point, we can compare this the cumulative reward received by a victim agent not subject to an antagonist.

Question 3: Assume the hypothesis is confirmed. The answers to questions 1 and 2 will provide a basis for understanding when an antagonistic agent can be detected directly, if at all. We will explicitly test if an antagonistic agent can be detected in the affirmative by measuring 1) the impact to total reward of the victim and 2) the difference in joint policy, if convergent, where both 1 and 2 are a function of antagonistic agent parameterization (over multiple realizations).

4.1 Training Framework

Our experiment is practically motivated by studying the presence of an antagonistic agent. It will be composed of a treatment group and a control group both performing MADDPG, each comprised of 5 agents with linear stochastic policies, agent i's action, $a_i \sim \pi_{\theta_i}$. The treatment group contains 1 antagonistic agent with 1 victim, and the control class contains no antagonistic agents. Confirming our hypothesis requires positively observing the predicting behavior in either case.

Non antagonistic agents will have quadratic reward functions, where agent action dependency arises from mixed terms in choices of action. Specifically,

$$r_i(s, \boldsymbol{a}) = \sum_{j=1}^{N} \sum_{k=1}^{N} w_{i,j}(s) a_j a_k + \sum_{m=1}^{N} w_m(s) a_m \tag{6}$$

where products of actions ensures pair-wise agent reward dependency on competitors in an easily structured way, with weights $W_{i,j}(s) : S \to \mathbb{R}^{N \times \Sigma_i |\mathcal{A}_i|}$, for agent i, for the j degree of the reward polynomial. By choosing a population of 4 non-antagonistic agents, we will represent each possible combination of agent reward function dependencies: antagonist-victim, victim-neutral, neutral-neutral, and independent. The shared state-space S will contain 3 fully-connected states, and agents will perform actions discretized over an open, zero centered interval $(0, 1)$, forming identical action sets $\mathcal{A}_i$ of size $|\mathcal{A}_i| = 1,000$ (chosen large enough to approximate a continuous action space), for each agent at each state. Therefore, the policy changes will depend on the current state, parameterized per-agent by $W_{i,j}(s)$.

At each of the 3 states, *the continuous analogue of the game Jacobian conditioned on that state will satisfy the conditions of* [9], such that each state—for large enough time horizon T—is hypothesized to exhibit for gradient methods, policy convergence to a fixed point (positive definite), policy convergence to a limit cycle (skew symmetric, complex eigenvalues), and policy divergence (negative definite). This is achieved by choice of $W_{i,j}(s)$ for each like in the continuous action space examples C_1 and C_2 in the two agent example case, but over finely discretized actions in the $(0, 1)$ interval. We will consider three distinct transition kernels: 1) uniform probability of state transition independent

of joint action $\boldsymbol{a}$, 2) a 0.99 probability of remaining at the current state independent of joint action $\boldsymbol{a}$, and 3) a linear function of joint action $\boldsymbol{a}$ increasing from 0 to 1. By visualizing and measuring the KL divergence of stochastic policy updates over training epochs, $KL(\pi_{\theta_i,t}|\pi_{\theta_i,t-1})$ for a given state s, we will confirm or reject our hypothesis for each transition kernel.

Finally, in testing our hypothesis, we will further explore the effect of an antagonistic agent on small groups of victim agents while restricted to a space of polynomial reward functions with mixed terms by varying each agent's discount factor γ_i, randomly scaling $W_{i,j}(s)$, and observing differences with and without an antagonistic agent by measuring, in addition to stochastic policy KL divergence between time steps, agent state-entropy and cumulative reward. Furthermore, by holding all other environment parameters constant, we will investigate the effect of increasing the number of agents N and victims $M \subset N$ to be very large and scaling learning rates as a function of these agent demographics (i.e. antagonist, victim, or neutral) to explore the effect observed in Fig. 1b.

References

[1] Raman Arora, Ofer Dekel, and Ambuj Tewari. Deterministic mdps with adversarial rewards and bandit feedback. *arXiv preprint arXiv:1210.4843*, 2012.

[2] Michael Bowling and Manuela Veloso. An analysis of stochastic game theory for multiagent reinforcement learning. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2000.

[3] Michael Bowling and Manuela Veloso. Rational learning of mixed equilibria in stochastic games. In *UAI2000 Workshop entitled Beyond MDPs: Representations and Algorithms*, 2000.

[4] Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, pages 1021–1026. Lawrence Erlbaum Associates Ltd, 2001.

[5] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

[6] Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

[7] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[8] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pages 6379–6390, 2017.

[9] Eric Mazumdar, Lillian J Ratliff, and S Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

[10] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

[11] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

[12] Shankar Sastry. *Nonlinear systems: analysis, stability, and control*, volume 10. Springer Science & Business Media, 2013.

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