Pacific Northwest
NATIONAL LABORATORY

Policy Convergence Under the Influence of Antagonistic Agents in Markov Games

NeurIPS Pre-registration
Workshop 2020

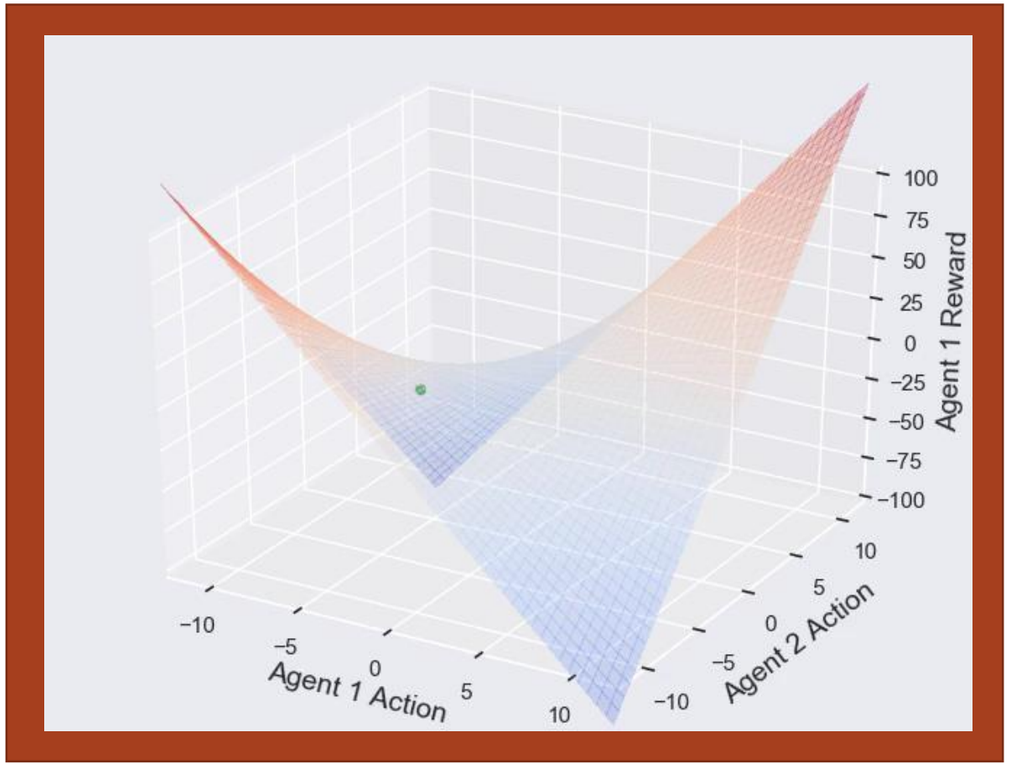
Chase Dowling, Ted Fujimoto, Nathan Hodas
Contact: chase.dowling@pnnl.gov

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Background & Intuition

- Recent results in game theory can be used to provide insight into policy convergence in multi-agent Markov decision processes
- For gradient-learning agents, differential Nash equilibria could be extrapolated to MDP settings
- In a state-free setting, can describe when joint policies converge to stable point, a limit cycle, or diverge altogether[1]
- An antagonistic agent is introduced to motivate the study. An antagonistic agent has a distinct action space, but their objective function is the negative of their victim's



Joint policy convergence to limit cycle for two agents with equal fixed gradient step sizes plotted on agent 1's reward surface

1. 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

Experimental Plan

Hypothesis: State-conditioned policy convergence depends on spectrum of the joint reward Jacobian

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

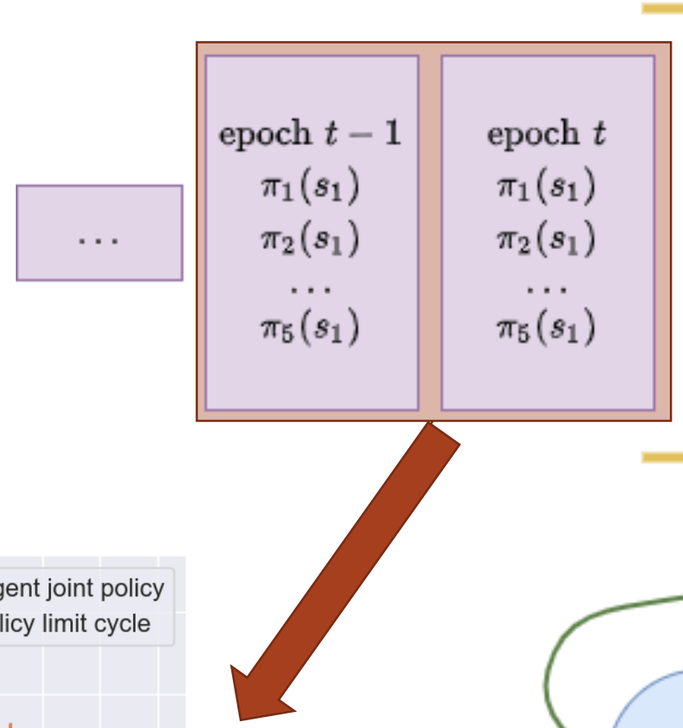
$$w_{\{.\}}(s)$$

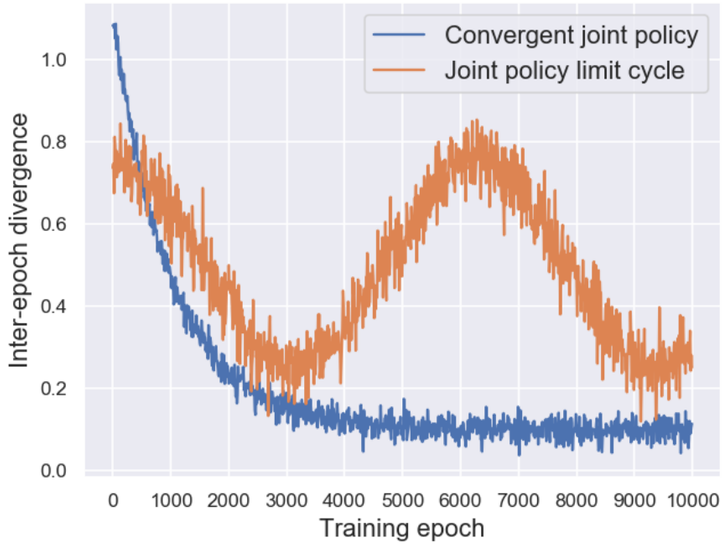
$$\nabla^2 R = \begin{bmatrix} \frac{\partial^2 r_1(s,\boldsymbol{a})}{\partial a_1^2} & \cdots & \frac{\partial^2 r_1(s,\boldsymbol{a})}{\partial a_1 a_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 r_N(s,\boldsymbol{a})}{\partial a_N a_1} & \cdots & \frac{\partial^2 r_N(s,\boldsymbol{a})}{\partial a_N^2} \end{bmatrix}$$

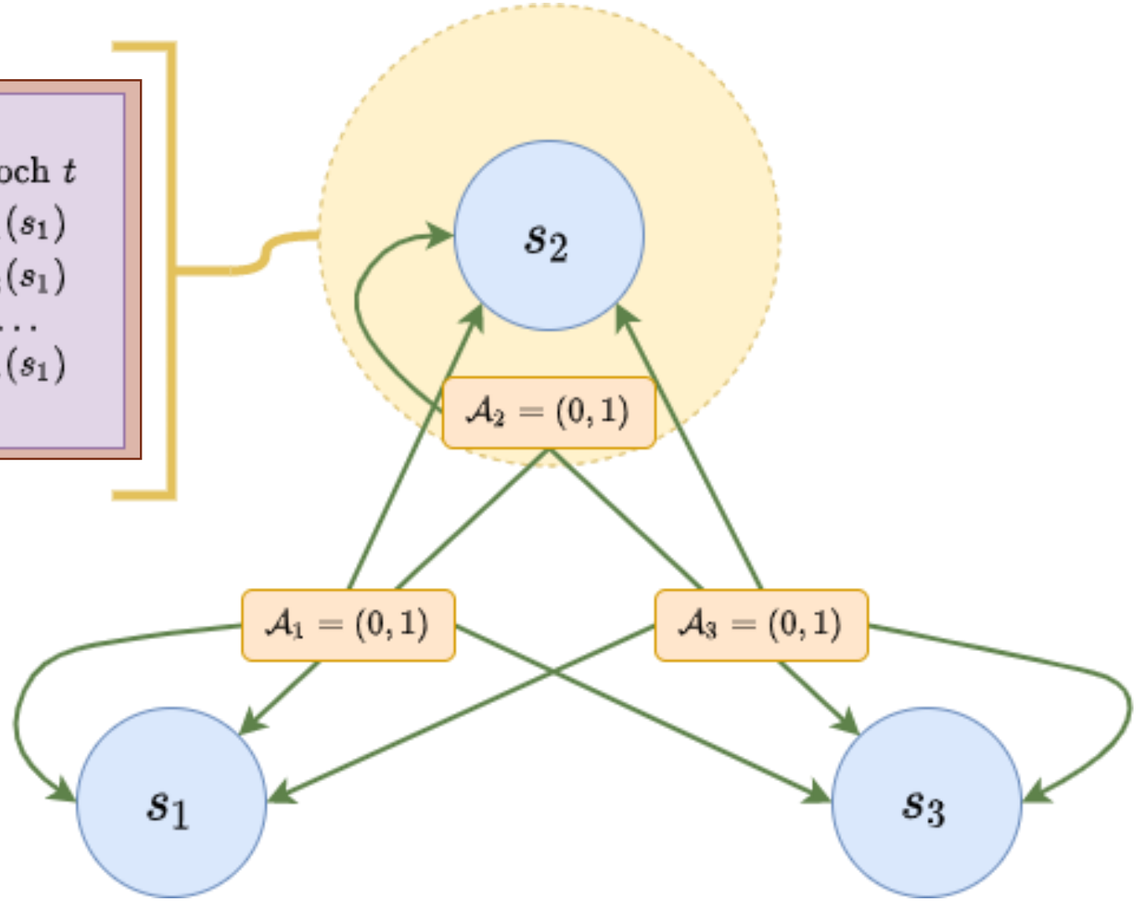
In our experiments, instantaneous reward functions will be polynomials of degree 2 as functions of all agents' choices. The choices of weights determines the spectrum

Experimental Plan

Hypothesis: State-conditioned policy convergence depends on spectrum of the joint reward Jacobian







Experiment plan MDP structure; state transitions represented in green for multiple kernels to be tested

Experimental Plan

- Five agents, trained off-policy with multi-agent deep deterministic policy gradient[2]
  - Five agents represent each combination of interaction: antagonist-victim, victim-victim, victim-neutral, and neutral-neutral.
  - The agents maintain an estimate of the other agents' policies, where in addition to their instantaneous reward, receiving 1) no information on, 2) noisy estimates of and 3) the exact values of the other agents' actions.
- Finely discretized action space with simple transition kernels
  - Transition kernels will be 1) high probability of remaining at the current state independent of action, 2) equal probability of remaining or transitioning, and 3) a linear function of the action
- Jacobian of instantaneous reward functions on continuous version of action space is 1) positive definite, 2) negative definite, or 3) complex spectrum
- Beyond our experimental plan, we will lay out avenues for further exploration

2. 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

Concluding Remarks

- Incentive alignment for a broader set of institutions.
- Incentive for scientists to risk negative results, and value gained in ensuring negative results *from reviewed experiments* have a venue to be shared
- Helpful suggestions from peers can be incorporated in advance

Thank you! Questions? Suggestions?