CLEAN Rewards for Improving Multiagent Coordination in the Presence of Exploration

(Extended Abstract)

Chris HolmesParker
Oregon State University
442 Rogers Hall
Corvallis, OR 97331
holmespc@onid.orst.edu

Adrian Agogino
UCSC at NASA Ames
Mail Stop 269-3
Moffett Field, CA 94035
adrian.k.agogino@nasa.gov

Kagan Tumer
Oregon State University
204 Rogers Hall
Corvallis, OR 97331
kagan.tumer@oregonstate.edu

ABSTRACT

In cooperative multiagent systems, coordinating the joint-actions of agents is difficult. One of the fundamental difficulties in such multiagent systems is the slow learning process where an agent may not only need to learn how to behave in a complex environment, but may also need to account for the actions of the other learning agents. Here, the inability of agents to distinguish the true environmental dynamics from those caused by the stochastic exploratory actions of other agents creates noise on each agent's reward signal. To address this, we introduce Coordinated Learning without Exploratory Action Noise (CLEAN) rewards, which are agent-specific shaped rewards that effectively remove such learning noise from each agent's reward signal. We demonstrate their performance with up to 1000 agents in a standard congestion problem.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Performance

Keywords

Learning, Reward Shaping, Multiagent Exploration

1. INTRODUCTION

Learning in large multiagent systems is a critical area of research with applications including controlling teams of autonomous vehicles [1], managing distributed sensor networks [4], and robot coordination [2]. A key difficulty of learning in such systems is that the agents in the system provide a constantly changing background in which each agent needs to learn its task. As a consequence, agents need to extract the underlying reward signal from the noise of other agents acting within the environment. This learning noise can have a significant and often detrimental impact on the resultant system performance.

When agents treat each other as part of the environment, each agent's exploratory actions are seen by other agents as stochastic environmental dynamics. Here, the inability of agents to distinguish the true environmental dynamics from those caused by the stochastic exploratory actions of other agents creates noise on each agent's reward signal. This problem cannot simply be addressed by turning off exploration and acting greedily (this has been repeatedly shown to result in poor performance as agents always exploit their current knowledge which is frequently incomplete or inaccurate [3]). We address this by introducing Coordinated Learning without Exploratory Action Noise (CLEAN) rewards, which are designed to effectively remove much of the learning noise caused by agents taking exploratory actions.

2. BACKGROUND

It is common for agents treat each other as part of the environment such that the exploratory actions of other agents are treated as stochastic environmental noise. However, under such assumptions, the agents are unable to distinguish when their peers are taking purposeful actions, from when they are taking random exploratory actions. Here, agents are frequently adapting their policies to better coordinate with the random exploratory actions of other agents, meaning that agents will end up learning to bias their policies such that they actually depend upon the exploratory actions of other agents in order to perform well. This means that agents learning optimal policies in the presence of exploration may not be optimal once learning is complete and exploration is turned off. In this setting, the agents' inability to distinguish between true environmental dynamics and dynamics caused by the exploratory actions of other agents means that the agents themselves (the solution) actually end up becoming part of the problem (added complexity due to stochastic learning noise).

3. CLEAN REWARDS

We develop Coordinated Learning without Exploratory Action Noise (CLEAN) rewards to simultaneously address the structural credit assignment problem and issues arising from learning noise caused by exploration in order to promote learning, coordination, and scalability in multiagent systems. These rewards utilize "off-policy" counterfactual actions which allow agents to approximate rewards associated with actions that were not actually taken. The key requirements of CLEAN rewards is that agents have some

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

approximation of the underlying system objective, G, and that the agents in the system follow their current target policies. Traditionally, following target policies has been shown to perform poorly due to a lack of exploration, however, CLEAN rewards address this shortcoming via off-policy counterfactual action exploration (Equation 1).

CLEAN rewards is defined as follows:

$$\mathcal{C}_{1,i} \equiv G(z_T - z_{T,i} + c_i) - G(z_T - z_{T,i} + c_i') \qquad (1)$$

where $\mathcal{C}_{1,i}$ is the CLEAN reward of agent i, z_T is the system state vector that results from the agents following their current target policies, $z_{T,i}$ is the action of agent i, c_i and c_i' are two counterfactual actions of agent i (i.e. alternative actions agent i could have taken instead of following its target policy), and G is the system objective. These CLEAN rewards replace the contribution of the agent's target action $z_{T,i}$ with two different counterfactual actions c_i and c_i', in the first and second terms, respectively. Here, the agent approximates the reward it would have received if it would have taken actions c_i and c_i'. Then, the agent compares the counterfactual rewards associated with each action, and provides the agent with a reward for action c_i based upon the difference between the two approximations (Equation 1).

3.1 The Gaussian Squeeze Domain

This domain assumes that there exists a set of agents which each contribute to a system objective, and the agents are attempting to learn to optimize the capacity of the system objective. The objective function for the domain is as follows:

$$G = xe^{\frac{-(x-\mu)^2}{\sigma^2}} \qquad (2)$$

where x is the cumulative sum of the actions of agents, μ is the mean of the system objective's Gaussian (effectively the target "x" that the agents are aiming for), σ is the standard deviation of the system objective's Gaussian. Here, the goal of the agents is to choose their individual actions x_i in such a way that the sum of their individual actions is to optimize Equation 2. Here, each agent has 10 actions ranging in participation value from zero to nine. The GSD is a congestion domain, where adjusting the variance changes the coordination complexity for agents within the system. The lower the variance, the higher the coupling of agents' joint actions.

4. RESULTS

As seen in Figure 1, the online performance (i.e., performance while learning when exploration is "on") of agents using difference and global (i.e., D and G) reward structures maintain better performance then their offline counterparts (i.e., D_{OFF} and G_{OFF}), even though the offline agents are greedily following their learned policies (i.e., exploration and learning is turned off and agents follow their fixed policies). This is because during learning, each agent's exploratory actions are "public", meaning that they can implicitly be observed by all other agents. In this setting, agents are learning to bias their individual policies in order to account for the random exploratory actions of other agents. In the end, these agents learn policies that actually depend upon the stochastic exploratory actions of other agents that are present during the online learning process, and when learning is "completed" and the other agents no longer take these exploratory actions, the performance actually decreases. Here, the solution (agents) actually become a part of the problem. Learning with CLEAN rewards address this shortcoming and effectively "filter off" the exploratory actions of other agents by "privatizing" each agent's exploration. As seen, CLEAN rewards significantly outperform other techniques with scaling up to 1000 agents.

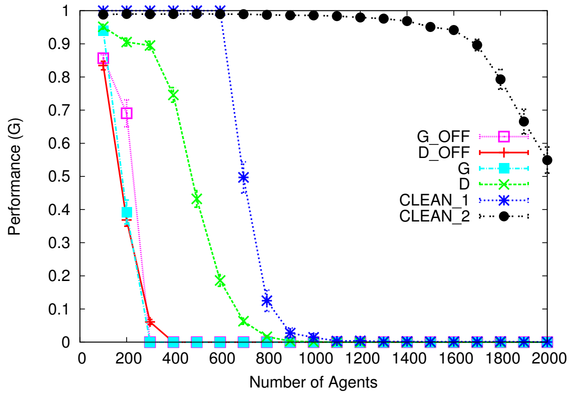

Figure 1: Scaling the number of agents with $\mu = 100$ and $\sigma = 100$. As the number of agents increases, the coupling between agents increases and the problem becomes more difficult and the *exploratory action noise* has more of an effect on system performance. As seen, CLEAN rewards are more robust than other rewards.

5. DISCUSSION

There has been a lot of research involving the exploration-exploitation tradeoff within the multiagent learning literature. However, relatively little work has been done to directly address the impact of learning noise caused by the exploratory actions of agents. In this work, we first showed the potential impact of such exploratory action noise on learning, demonstrating that exploratory actions can cause agents to bias their policies to depend upon the exploratory actions of others, which can lead to suboptimal learning. To address this, we introduced CLEAN rewards, which are shaped rewards designed specifically to promote coordination and scalability in multiagent systems by addressing both the structural credit assignment problem, as well as the *exploratory action noise* caused by agent exploration.

6. REFERENCES

[1] A. Agogino, C. HolmesParker, and K. Tumer. Evolving large scale uav communication system. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, Philadelphia, PA, July 2012.

[2] L. Panait and S. Luke. Cooperative multi-agent learning - the state of the art. *In the Journal of Autonomous Agents and MultiAgent Systems*, 2005.

[3] R. Sutton and A. Barto. *Reinforcement Learning An Introduction.* MIT Press, Cambridge, MA, 1998.

[4] S. Williamson, E. Gerding, and N. Jennings. Reward shaping for valuing communications during multi-agent coordination. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2009.