CLEANing the Reward: Counterfactual Actions to Remove Exploratory Action Noise in Multiagent Learning (Extended Abstract)

Chris HolmesParker
Parflux LLC
chris@parflux.com

Mathew E. Taylor
Washington State University
taylorm@eecs.wsu.edu

Adrian Agogino
UCSC at NASA Ames
adrian.k.agogino@nasa.gov

Kagan Tumer
Oregon State University
ktumer@oregonstate.edu

ABSTRACT

Learning in multiagent systems can be slow because agents must learn both how to behave in a complex environment and how to account for the actions of other agents. The inability of an agent to distinguish between the true environmental dynamics and those caused by the stochastic exploratory actions of other agents creates noise in each agent's reward signal. This learning noise can have unforeseen and often undesirable effects on the resultant system performance. We define such noise as *exploratory action noise*, demonstrate the critical impact it can have on the learning process in multiagent settings, and introduce a reward structure to effectively remove such noise from each agent's reward signal. In particular, we introduce Coordinated Learning without Exploratory Action Noise (CLEAN) rewards and empirically demonstrate their benefits.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Miscellaneous

General Terms

Algorithms, Experimentation, Performance

Keywords

Coordination, Scalability, Exploration vs. Exploitation

1. INTRODUCTION

In multiagent systems, agents provide a constantly changing background in which each agent needs to learn its task [3, 4, 8]. As a consequence, agents need to extract the underlying reward signal from the noise of other agents acting within the environment. Issues arise when agents are treated as a part of the environment, and their exploratory actions are seen by other agents as stochastic environmental dynamics. 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). Additionally, methods of slowly turning down exploration (e.g., annealing) or intelligently modifying exploration (e.g., Win-Or-Learn-Fast WOLF [2]) fail to fully address this issue. This is because these techniques still rely upon agents taking explicit exploratory actions within the environment. CLEAN rewards address this issue using implicit exploration via counterfactual action based reward shaping techniques, such that all explicit exploration is removed from the learning process.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. EXPLORATORY ACTION NOISE

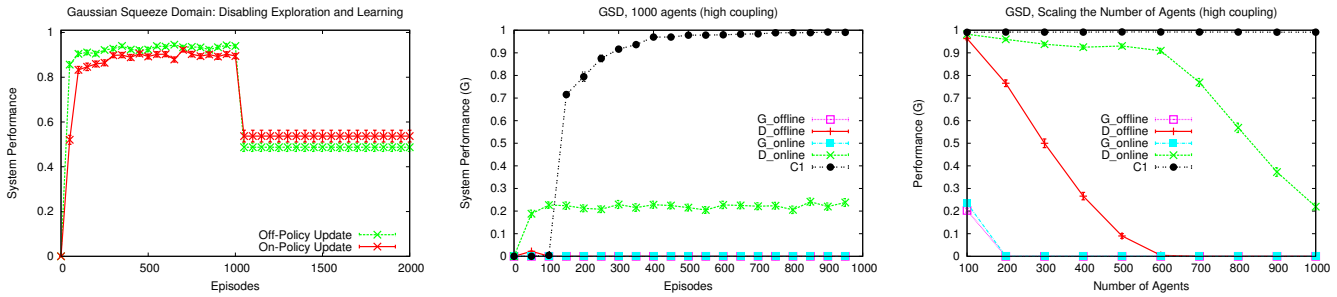
Agents often treat other agents as part of the environment — the exploratory actions of other agents become stochastic environmental noise [5, 6, 7, 10]. Here, agents are then unable to distinguish when their peers are taking purposeful actions or are exploring. This may cause agents bias their policies such that they actually depend upon the exploratory actions of other agents to perform well. Agents learning in the presence of exploration may not be learning optimal policies (Figure 1a) because agents cannot distinguish between true environmental dynamics and dynamics caused by the exploratory actions of other agents. Here, agents (the solution) actually become part of the problem (adding stochastic noise to the environment). This holds for both off-line and on-line learning methods.

3. CLEAN REWARDS

Coordinated Learning without Exploratory Action Noise (CLEAN) rewards address the structural credit assignment problem and issues arising from learning noise caused by exploration to promote learning, coordination, and scalability. CLEAN rewards separate explicit from implicit exploration. Agents behave greedily outwardly (explicitly) and explore internally (implicitly) via counterfactual exploratory actions. Agents use Equation 2 to perform counterfactual reward calculations:

$$\mathcal{C}_{1,i}(\mathbf{a}) \equiv G(\mathbf{a}_{a_i \leftarrow a'_i}) - G(\mathbf{a}) \quad (1)$$

$\mathbf{a}$ is the system action vector and a_i is agent i's action. This gives the agent a reward that represents how the system would have performed had it not followed its best policy, but instead had taken some counterfactual action, a'_i. CLEAN rewards use implicit counterfactual exploration to eliminate explicit exploratory action noise within the environment. **The Gaussian Squeeze Domain** A set of agents



(a) (b) (c)

Figure 1: **Gaussian Squeeze Domain Results**: a)When agents stop exploring in the GSD domain (episode 1000), system performance decreases due to exploratory action noise. b) CLEAN rewards outperform global and difference rewards. c) CLEAN rewards maintain superior performance as scaling increases.

in attempt to learn to optimize the capacity of the following system objective:

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

where x is the sum of the actions of agents ($x = \sum_{i=0}^{n} a_i$), μ is the mean and σ is the standard deviation of the system objective. Here, μ and σ define the target capacity, x, that the agents must coordinate their actions to achieve.

4. RESULTS

There were four types of experiments: random agents (baseline) and three types of Q-learning agents (global (G), difference (D) [1], and CLEAN ($\mathcal{C}_{1,i}$). Figure 1b shows the results for 1000 agents learning in the Gaussian Squeeze Domain with $\mu = 175$, $\sigma = 175$. The performance of agents using global rewards G is poor in both the online and the offline settings because global rewards do not provide individual agents with specific feedback on how their individual actions impacted the system performance compared to the actions of all of the other agents in the system (i.e., each agent's reward signal gets lost in the "noise" of the rest of the system). Agents using difference rewards D outperformed agents using G because difference rewards provide each agent with a reward that is reflective of it's own individual impact on the system performance. Unfortunately, difference rewards do not address the issues associated with exploratory action noise. The disparity in performance between CLEAN rewards and difference rewards can be directly attributed to the impact of exploratory action noise on the learning process. As seen, exploratory action noise can have a massive impact on learning performance, especially in large tightly coupled multiagent systems. Agents using CLEAN rewards all converge to nearly optimal performance, maintaining 5 times the performance of the next best technique (i.e., D) with scaling up to 1000 agents.

The GSD experiment in Figure 1c considers how performance changes as complexity increases. Figure 1c shows the results of scaling the number of agents with a fixed mean and variance ($\mu = 175$ and $\sigma = 175$). CLEAN rewards are more robust to scaling (increased congestion) than G and D because agents receive a cleaner learning signal.

5. DISCUSSION AND CONCLUSION

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. We first showed the potential impact of 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. We then introduced CLEAN rewards, which are shaped rewards that promote coordination and scalability in multiagent systems by addressing *exploratory action noise* caused by agent exploration.

Acknowledgements This work was partially supported by the National Science Foundation under grant NSF IIS-1149917 and the National Energy Technology Laboratory under grants DE-FE0012302 - DE-FE0011403.

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] M. Bowling. Convergence and no-regret in multiagent learning. *NIPS*, 2005.

[3] G. Chalkiadakis and C. Boutilier. Sequentially optimal repeated coalition formation under uncertainty. *JAAMAS*, 2012.

[4] M. Colby, C. HolmesParker, and K. Tumer. Coordination and control for large distributed sensor networks. In *FIIW*, 2012.

[5] A. Eck, L. Soh, S. Devlin, and D. Kudenko. Potential-based reward shaping for POMDPs. In *AAMAS*, 2013.

[6] C. HolmesParker, A. Agogino, and K. Tumer. Clean rewards for improving multiagent coordination in the presence of exploration. In *AAMAS*, 2013.

[7] H. Liu, E. Howley, and J. Duggan. Coevolutionary analysis - a policy exploration method for system dynamics models. In *System Dynamics Review*, 2012.

[8] M. Taylor, M. Jain, Y. Jin, M. Yooko, and M. Tambe. When should there be a "me" in "team"? Distributed multi-agent optimization under uncertainty. In *AAMAS*, 2010.

[9] L. Torrey and M. Taylor. Teaching on a budget - agents advising agents in reinforcement learning. In *AAMAS*, 2013.

[10] Y. Zhan, J. Wu, C. Wang, M. Liu, and J. Xie. On the complexity of undominated core and farsighted solution concepts in coalition games. In *AAMAS*, 2013.