When Less is More: Reducing Agent Noise with Probabilistically Learning Agents

Extended Abstract

Jen Jen Chung
Oregon State University
Corvallis, OR
jenjen.chung@oregonstate.edu

Scott Chow
Oregon State University
Corvallis, OR
chows@oregonstate.edu

Kagan Tumer
Oregon State University
Corvallis, OR
kagan.tumer@oregonstate.edu

ABSTRACT

Distributed agents concurrently learning to coordinate in a multiagent system can suffer from considerable amounts of agent noise. This is the noise that arises from the non-stationarity of the learning environment for each individual agent since other agents in the system are also constantly updating their policies, thereby continually shifting the goal posts for successful coordination. In this work, we propose a method to reduce agent noise by allowing individual agents to probabilistically determine whether or not to undergo policy updates. We show that using this method to adapt the number of actively learning agents over time provides improvements in convergence speed of the team as a whole without affecting the final converged learning performance.

KEYWORDS

Multiagent Learning; Agent Noise; Reasoning about Action

ACM Reference Format:
Jen Jen Chung, Scott Chow, and Kagan Tumer. 2018. When Less is More: Reducing Agent Noise with Probabilistically Learning Agents. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 MOTIVATION & BACKGROUND

Distributed multiagent learning is a powerful method by which a team of cooperative agents search for optimal joint strategies for achieving a global objective such as package delivery, space exploration, and rescue missions. One basic assumption made in most learning algorithms is that the environment is stationary. Repeated interactions between the agent and the environment result in a reward signal that may be stochastic, but is centered about a mean that does not change over time. However, this premise is violated with the introduction of multiple learning agents, since the behavior of every agent changes over time as they learn, resulting in a complex web of interactions amongst the agents. The perturbations in the system caused by other agents concurrently learning results in a significant amount of noise in the global reward signal, thus the term *agent noise*.

Structural Credit Assignment. Methods designed to address the structural credit assignment problem (how much each agent contributed to the outcome) reduce *intra-epoch* noise in the reward signal by filtering out the credit/blame for other agents' behavior. Potential-based reward shaping [8] addresses this issue with an additional reward to the global reward which is provided to encourage certain behaviors and decrease convergence time. The difference reward is another example of a shaped reward [1], where agents compute the difference between the global reward and the hypothetical global reward that would have been received had the agent executed a *counterfactual* state-action.

Another approach that indirectly reduces intra-epoch agent noise is to model other agents' states and actions. The concept of fictitious play has been investigated as a technique for learning coordination [4, 7, 9]. Ad hoc teaming involves the transfer of policies for coordinating ad hoc teams formed online [2]. However, both of these methods rely on strong assumptions about other agents in the system to model their behavior. In our work, we aim to address agent noise without explicit models of other agents.

Inter-Epoch Agent Noise. The problem of *inter-epoch* agent noise in multiagent learning arises from the fact that individual agents are concurrently updating their executed control policies, thereby producing a learning environment in which the transition functions change over time.

Coordinated Learning without Exploratory Action Noise (CLEAN) rewards by HolmesParker et al. [6] address this issue by allowing agents to take "private" exploratory actions that are compared against the "public" on-policy action it takes as it appears to other agents. This allows agents to explore without disturbing other agents' learning. The "Win or Learn Fast" or WoLF principle is another approach addressing inter-epoch noise [3]. WoLF increases the learning rate of losing agents and decreases the learning rate of winning agents. In our work, we probabilistically increase the number of agents updating their policies over the course of learning in a cooperative domain to directly address inter-epoch noise.

The contribution of this paper is a direct method for mitigating agent noise by limiting the number of agents learning at a given time. Intuitively, because every learning agent contributes agent noise to the reward of every other agent, reducing the number of agents modifying their policies reduces the total amount of agent noise. By reducing agent noise, each agent receives a cleaner reward signal, allowing overall improvements in the learning performance of the team as a whole.

We test our algorithm on a multi-night variant of the El Farol Bar Problem [5] and show that our learning algorithm results in faster convergence while reaching similar converged performance compared to the case with all agents concurrently learning.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

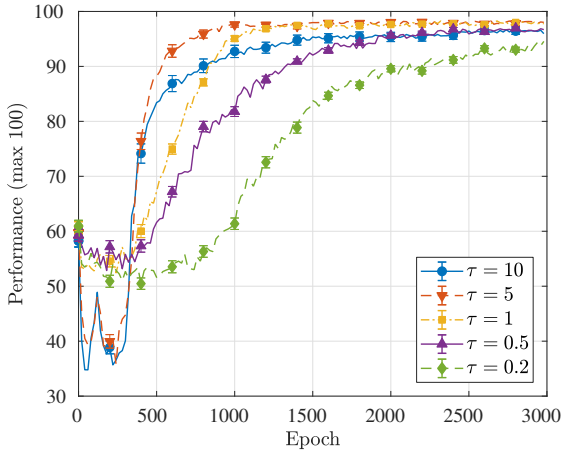

Figure 1: Global reward achieved versus number of learning epochs. Error bars represent error from the mean. Probabilistically determining the number of agents concurrently learning increases convergence rate compared to permitting all agents to learn from the start.

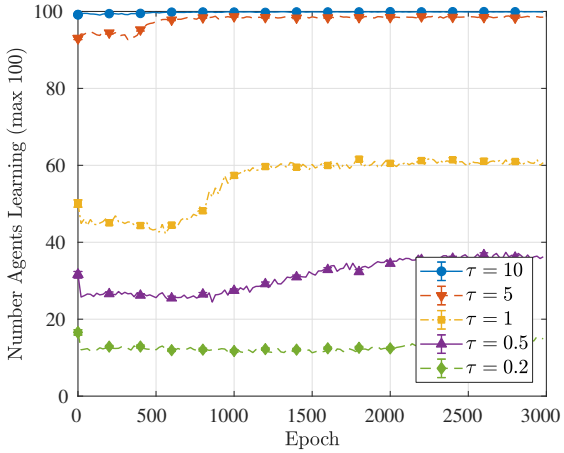

Figure 2: Number of agents learning versus number of learning epochs for the bar problem for the same experiment as Figure 1. Error bars represent error from the mean. The probability of learning (thus number of agents learning) is dependent on D and the τ parameter.

2 ADAPTING WHEN TO LEARN

We employ a probabilistic framework to determine when an agent should learn. The probability that an agent learns is given by:

$$P_{learn} = 1 - \exp\left(-\frac{\tau}{|D|}\right), \tag{1}$$

where D is the difference reward (Equation 3) and τ is a parameter that regulates the rate at which the probability of learning increases. In this case, we use $\frac{1}{|D|}$ as a metric for the impact an agent has on the system. When an agent gets a large difference reward, it is more likely to have reached a good policy and should therefore have a lower chance of learning to allow poorer performing agents to learn with less noise.

3 DOMAIN

The Multi-Night Bar Problem is an extension of the El Farol Bar Problem [5]. We use this problem not as a congestion problem but as a learning problem to evaluate speed of learning convergence. In the multi-night bar problem, there is a bar that is open k nights. Each night has an optimal capacity c such that maximum enjoyment is received if c agents attend that night. Each agent must choose to attend a night without any consultation with other agents.

Once each agent has chosen a night, the reward for each week is:

$$G = \sum_{n} c \times \exp\left(-\frac{(a_n - c)^2}{\gamma}\right) \tag{2}$$

where G is the global reward, c is the capacity, and a_n is the number of agents that chose to attend night n. γ is a parameter damping the reward, set to 10 for this experiment.

We consider the 10 night bar problem with 100 agents. Each night has an optimum capacity of 10. Each agent learns via Q-Learning with a learning rate of $\alpha = 0.1$. Results are based on 100 runs.

We instantiate a Q-learner for each agent and provide the difference reward D_i as the reward signal, which is defined as:

$$D_i = G - G_{-i} = c \times e^{-\frac{(a_n - c)^2}{\gamma}} - c \times e^{-\frac{((a_n - 1) - c)^2}{\gamma}} \tag{3}$$

where G_{-i} is the global reward were agent i not present.

4 RESULTS

We use our probabilistic framework to allow the number of agents learning concurrently to increase over time. Figures 1 and 2 shows learning speed is dependent on the number of agents learning. We run the same experiment with multiple τ values to show how this parameter affects the probability of learning. The τ value controls the sensitivity of the probabilistic learning to the individual's impact as measured by $\frac{1}{|D|}$. Regulating the number of concurrent learners at the start allows for faster convergence after an initial delay, even if only half of the agents are learning at any given time as shown for $\tau = 1$. When τ is large, the probability of learning increases until all agents are learning at all times, as shown for $\tau = 10$.

5 DISCUSSION

In this work, we showed that reducing the number of agents that are concurrently learning can be beneficial for reducing agent noise. We proposed a probabilistic framework for choosing impactful agents to learn. We applied our framework to the multi-night bar problem and demonstrated a speedup in the learning transience.

Future work would involve exploring more salient versions of impact. For this work, we chose to use a very simple metric, $\frac{1}{|D|}$, to evaluate the impact of an agent, which results in a learning similar to WoLF with poorer performing agents increasing its probability learning. The performance of our formulation can be improved with more informative measures of impact. Additionally, more complex, heterogeneous domains may be better suited for impact-based learning.

REFERENCES

[1] Adrian K. Agogino and Kagan Tumer. 2008. Efficient evaluation functions for evolving coordination. *Evolutionary Computation* 16, 2 (2008), 257–288.

[2] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242 (2017), 132–171.

[3] Michael Bowling and Manuela Veloso. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136, 2 (2002), 215–250.

[4] George W. Brown. 1951. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation* 13, 1 (1951), 374–376.

[5] Mitchell Colby, Theodore Duchow-Pressley, Jen Jen Chung, and Kagan Tumer. 2016. Local approximation of difference evaluation functions. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*. 521–529.

[6] Chris HolmesParker, Matthew E. Taylor, Adrian K. Agogino, and Kagan Tumer. 2014. CLEAN Rewards to Improve Coordination by Removing Exploratory Action Noise. In *International Conference on Intelligent Agent Technology*. Warsaw, Poland, 127–134.

[7] David S. Leslie and Edmund J. Collins. 2006. Generalised weakened fictitious play. *Games and Economic Behavior* 56, 2 (2006), 285–298.

[8] Andrew Y. Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*. 278–287.

[9] Xiaofeng Wang and Tuomas Sandholm. 2003. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Advances in Neural Information Processing Systems*. 1603–1610.