Model-Free Policy Gradients for Multi-Agent Shape Formation

Elizabeth Boroson, Fei Sha, and Nora Ayanian

I. INTRODUCTION

Distributed solutions for tasks that require tight coordination between multiple agents present a significant challenge. Common approaches for multi-agent coordination either depend on full observability and deterministic policies, like Q-learning algorithms [1], or use models that cannot scale above a few agents [2]. These techniques cannot be applied in larger groups or scenarios with only partial observability.

In this work, we study *shape formation*, where a group of agents must arrange themselves into a desired shape.Agents can move within the space and observe other agents nearby, but cannot identify or explicitly communicate with them.

We introduce a model-free approach for a multi-agent system to learn distributed policies. The agents use gradient ascent to jointly reach policies to complete shape formation efficiently. While this problem provides a working example, the technique is much more general and can be applied to any cooperative problem and scaled to any size group.

II. METHODS

We simulate the shape formation task using a grid world, as shown in Fig. 1, with a set of goal states. Agents occupy one grid cell and can move to an adjacent empty cell. Each agent observes only the surrounding 8 grid cells, and cannot communicate with or identify other agents. The group is rewarded only when it is in a goal state. Each agent uses gradient ascent on a parameterized stochastic policy to find parameters for a policy with maximum expected reward.

Estimating the gradient at a parameter value is done by running a simulation; the gradient is computed iteratively by each agent as it moves around the grid world. A conjugate gradient algorithm with line search is used to find parameters where the gradient is approximately zero [3]. The best choice of parameters for one agent depends on all other agents' policies, so we iterate through the set of agents, and each agent completes one step of conjugate gradient ascent until all agents reach parameters with gradients near zero.

III. EXPERIMENTAL RESULTS

We performed a series of experiments in the simulation. Groups of 2 to 8 agents trained to reach a set of goal states, then the groups' performance was evaluated. All groups reached policies significantly better than random exploration.

All groups learned to converge in a corner of the grid. In doing this, the agents implicitly agree on a meeting point, which is much more efficient than searching the space to find each other. Fig. 2 shows the behavior of all agents during one evaluation of shape formation. Larger groups required longer training times to reach similar results; however, the training time scales only linearly with the number of agents.

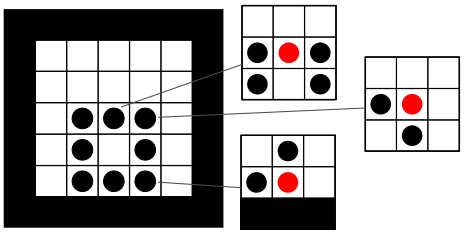

Fig. 1. Discrete gridworld, with examples of some agents' states

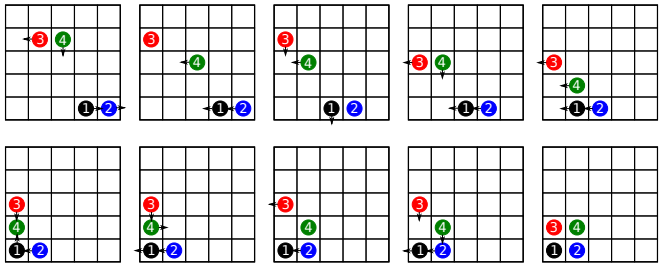

Fig. 2. A policy learned by 4 agents to form a square. All agents have learned to move toward the lower left corner first, then form the final shape.

A common alternative to our model-free approach would have agents use an explicit model of the world to select the action with the highest expected reward [2]. We used the MultiAgent Decision Process (MADP) toolbox [4] to compare the two methods. For the smallest problem with 2 agents, gradient ascent reached a good policy in about an hour, but none of the exact or approximate algorithms in the MADP toolbox could solve a problem of this size.

IV. CONCLUSION AND FUTURE WORK

We have developed a method for learning policies for coordination in a group of agents and introduced shape formation, a multi-agent problem requiring tight coordination, to demonstrate its performance. In the future, we plan to explore using this approach on larger groups, for example, by parallelizing the training process.

Authors are with the Dept. of Computer Science, University of Southern California, USA {boroson, feisha, ayanian}@usc.edu.
E. Boroson gratefully acknowledges support from the Annenberg and Viterbi Graduate School Fellowships. This work was supported by NSF CAREER IIS-1553726 and the Okawa Foundation.

REFERENCES

[1] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, 2008.

[2] F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.

[3] J. Baxter, P. L. Bartlett, and L. Weaver, "Experiments with infinite-horizon, policy-gradient estimation," *JAIR*, vol. 15, 2001.

[4] F. A. Oliehoek, M. T. J. Spaan, P. Robbel, and J. Messias, "The MADP toolbox: An open-source library for planning and learning in (multi-) agent systems," in *2015 AAAI Fall Symp. Series*, 2015, pp. 59–62.