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 robots present a significant challenge, due to the requirements for robots to model their own state, the states and actions of other robots, and any communication available. 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. They also cannot adapt to unknown or changing environments, like those that would be encountered in search and rescue after a natural disaster, where maps may be inaccurate and communication unreliable.

In this work, we study *shape formation*, where a group of agents in a 2-D space must arrange themselves into a desired shape. Agents can move within the space with limited velocity and observe other agents nearby, but cannot uniquely identify or explicitly communicate with them. The shape may be located anywhere in the space, as only agents' relative positions are important for task completion. Previous work on shape formation has used methods that require agents to be aware of their positions or use a fixed hierarchy, which are unrealistic requirements when sensing is limited and the environment is changing [3].

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, and these policies and these policies perform better than Q-learning or model-based methods in known spaces and in spaces with unknown obstacles. 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 grid world has a set of goal states, containing states where all agents form the goal shape. The group is rewarded only when it is in a goal state. Each agent maintains its own parameterized stochastic policy, and uses gradient ascent to find parameters for a policy with maximum expected reward.

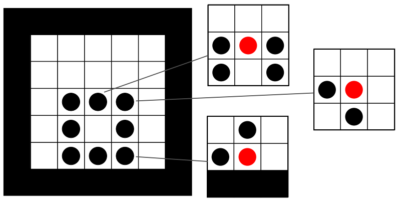

Fig. 1. Example of 8 agents forming a square in the discrete grid world, with examples of some agents' states on the right.

Estimating the gradient at a parameter value is done by running a simulation; each agent iteratively computes its own gradient as it moves around the grid world [4]. A conjugate gradient algorithm with line search is used to find parameters where the gradient is approximately zero [5]. The best choice of parameters for one agent depends on all other agents' policies, thus each agent cannot find an optimal policy alone. Instead, 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

A. Simulations

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. Each group's choice of corner is random, due to random behavior during the group's initial exploration. Within a group, each agent learns a similar policy and takes all positions in the goal shape with equal probability. Groups had good performance even for goal shapes where each agent could not see the whole shape within its local view, such as 4 agents forming a line, indicating that the learned coordination is implicit and does not depend on the exact number of agents present. 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.

B. Comparison to Hysteretic Q-Learners

We compared our approach to Hysteretic Q-Learning (HQL), a model-free Q-learning method in which each agent learns a Q-value function describing the expected reward for each observation-action pair. HQL typically works well for

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.

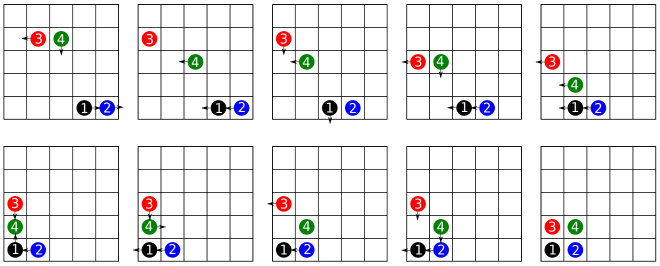

Fig. 2. A policy learned by 4 agents to form a square, beginning with random initialization in the top left and continuing from left to right. At each step, agents take the actions indicated in order. All agents have learned to move toward the lower left corner first, then form the final shape.

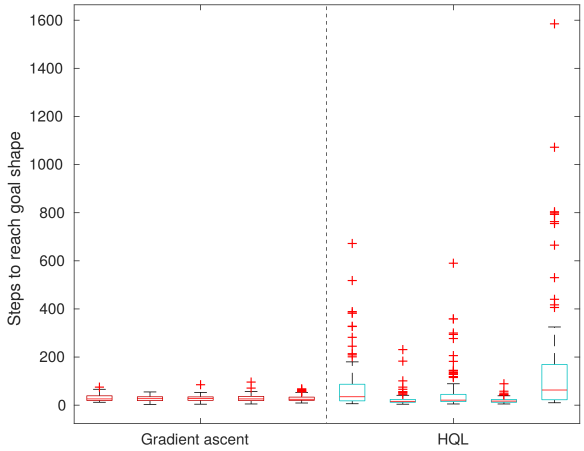

Fig. 3. For each algorithm, 5 different groups were trained and evaluated on 100 random initializations. This plot shows the performance of these groups, with each column showing the results from one group. Though HQL groups have median performance similar to gradient ascent groups, their worst-case performance tends to be much worse.

multiagent learning because it allows agents to adapt quickly to other agents learning better policies, but react slowly to other agents' exploratory behavior. The agents move around the grid world together and each compute their own Q-values, then follow an ε-greedy policy.

We evaluated groups of 4 agent forming squares trained with policy gradients and with HQL. In general, HQL produced a policy with similar median performance to gradient ascent agents, but with much more variance in individual evaluations of the same policy and in the quality of learned policies between different groups, as shown in Fig. 3.

We are interested in policies that not only complete shape formation well, but that can adapt to changes in the environment. We created test environments with obstacles, as shown in Fig. 4, and used them to evaluate groups of agents that had trained with no obstacles.

The two approaches performed similarly for most obstacles, but hysteretic Q-learners performed extremely poorly in spaces with an obstacle intersecting the wall they tended to converge on. Two of the groups of hysteretic Q-learners averaged over 100,000 steps to reach a goal state in this environment, with one taking over 500,000. In contrast, the obstacle that interfered most with the gradient ascent agents' policies was a block one grid square away from the corner where the group converged. However, all gradient ascent groups successfully completed the shape formation in about 13,000 steps, which is comparable to the 9,290 steps they took before learning any policy. While this is not optimal, it demonstrates that gradient ascent agents are able to overcome foreign obstacles more consistently than the HQL agents.

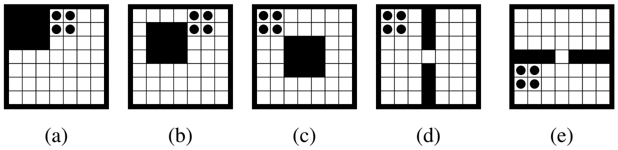

Fig. 4. Environments with obstacles, and common goal states that groups found. (a) Block in corner. (b) Block near corner. (c) Block in center. (d) Vertical wall. (e) Horizontal wall.

C. *Comparison to Model-Based Methods*

A common alternative to our model-free approach would have agents use an explicit model of the world and possible joint actions and state transitions to select the action with the highest expected reward [2]. With this method, exact solutions require large policy trees; while approximate solutions can be found by sampling the policy space, these solutions may be far from optimal. We used the MultiAgent Decision Process (MADP) toolbox [6] to compare the two methods. For two agents in a 3×3 grid, which requires a policy with only two steps, the MADP toolbox produced some approximate solutions, but no exact ones. For two agents in a 5×5 grid, which is the smallest problem evaluated with policy gradients, gradient ascent reached a good policy in about an hour, but none of the exact or approximate algorithms in the MADP toolbox could reach a solution.

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. We showed that policies learned using this method perform better with new obstacles in the environment than those learned using HQL, another model-free approach, and that they performed significantly better than model-based methods.

In the future, we would like to explore using the policy gradient approach on larger groups. Furthermore, we plan to work toward quantifying the requirements for our policies, which will be necessary to understand how those policies can be used in real applications.

REFERENCES

[1] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans Systems Man and Cybernetics Part C Applications and Reviews*, vol. 38, no. 2, p. 156, 2008.

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

[3] E. Bahçeci, O. Soysal, and E. Şahin, "A review: Pattern formation and adaptation in multi-robot systems," *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-03-43*, 2003.

[4] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *J Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[5] J. Baxter, P. L. Bartlett, and L. Weaver, "Experiments with infinite-horizon, policy-gradient estimation," *J Artificial Intelligence Research*, vol. 15, pp. 351–381, 2001.

[6] F. A. Oliehoek, M. T. J. Spaan, P. Robbel, and J. V. Messias, "The MADP toolbox: An open-source library for planning and learning in (multi-)agent systems," in *Sequential Decision Making for Intelligent Agents—Papers from the AAAI 2015 Fall Symposium*, Nov. 2015, pp. 59–62.