CSCI 670 - Theoretical Thinking Assignment 2

Varun Bhatt
8461808238
vsbhatt@usc.edu

December 6, 2021

In this report, I will be writing about the work that I did for my master's thesis on communication between agents in a multi-agent cooperative task [1]. The problem is modeled as a decentralized partially observable Markov decision process (dec-POMDP) and the proposed solutions were inspired by works in Game Theory and Reinforcement Learning (RL).

1. **What is the practical background of your problem?**

   Tasks such as driving on the road or moving a heavy table involve multiple people and are inherently cooperative. In such cases, humans make use of communication to efficiently finish the task. For example, during driving, indicators are used to signal a potential lane change. While the indicator itself doesn't directly effect the result of the lane change attempt, communicating the intent helps the driver in the rear to respond appropriately and increases the chance of a successful lane change.

   With robots slowly being integrated into similar real-world tasks, learning to communicate with humans and other intelligent agents becomes necessary.

2. **What is the mathematical definition of your problem?**

   The multi-agent communication problem is modeled as a dec-POMDP with cheap-talk channel for communication.

   Dec-POMDP is a special case of a Markov game [2]. A Markov game contains multiple agents and an environment. The environment has a set of possible states and each agent has a set of possible actions. At each time-step, all agents take an action based on the current state and receive a corresponding reward from the environment. Hence, it is a natural way to define problems involving multiple agents acting in an environment. Since agents are expected to only see a small portion of the environment in real-world problems, they only receive a private observation from the environment instead of the complete state. For cooperative problems, all agents would be receiving the same reward which depends on the state and the joint action of all agents. Combined, this formulation results in a dec-POMDP.

   The goal of all the agents is to find a mapping from their observations to actions, called a policy, that maximizes the cumulative (discounted) long term reward resulting from their joint actions.

   To help the agents in this task, communication between agents is added through a cheap-talk channel. The agents can send a message to each other at each time-step which doesn't directly affect the reward obtained but can be used by other agents in the subsequent time-steps to take a more informed action.

3. **What are the interesting algorithmic questions concerning your problem?**

   Consider a dec-POMDP in a tabular setting with known transition and reward function, i.e, transitions and rewards corresponding to each state and action can be represented by a finite number of bits. Planning to find an optimal joint policy for decentralized control in this setting requires NEXP-time even when there are only two agents [3].

   In reality, the problems can be even more complex. The state and/or action spaces can be infinitely large (e.g. when they are represented as real numbers). Representing the policy in such a setting requires function approximation which currently has no theoretical guarantees. Additionally, the model

of the world is generally unknown and needs to be learnt online. Finally, the agents may be required to learn independently instead of being allowed to plan in a centralized manner which adds game theoretic questions such as convergence to a Nash equilibrium or convergence to the Pareto-optimal Nash equilibrium. The communication channel might help with independent learning but the current results are only empirical.

Scalability of the algorithm in terms of the number of agents is also an interesting question to ask since in cases like self-driving cars, there could be hundreds of interacting agents.

People are also interested in an extension to this problem known as ad-hoc teamwork [4], in which cooperation and communication is between agents that have not interacted previously. In such settings, optimality can be defined based on how easy it is to adapt to an agent policy.

4. **What is the state of the art for modeling and solving this problem?**

   Since solving a dec-POMDP requires exponential time, the current methods train the agents incrementally till a good solution is reached. Two major paradigms are used for training.

   The first one is called centralized training for decentralized execution that allows the agents to train together but requires them to act independently after training. Centralized training allows methods such as backpropagating losses through the communication channel [5].

   The second paradigm uses decentralized training so that agents don't need to know the internal workings of the other agents. Methods using this paradigm generally add additional losses to incentivize communication [6] or, as in our work [1], post-process the learnt policy to improve the performance.

5. **What are interesting open questions inspired by your formulation?**

   Since the current works on multi-agent communication are empirical, theoretical guarantee for those algorithms remains an open question.

   A slightly different but related question that people are interested in is understanding the signals being communicated. Papers related to emergent communication [7] set up the problem in the same way, but once the agents are trained, they study if the communicated signals have a meaning that is related to how the environment works.

References

[1] V. Bhatt and M. Buro, "Inference-based deterministic messaging for multi-agent communication," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[2] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, 1953.

[3] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of Operations Research*, 2002.

[4] P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein, "Ad hoc autonomous agent teams: Collaboration without pre-coordination," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2010.

[5] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, 2016.

[6] T. Eccles, Y. Bachrach, G. Lever, A. Lazaridou, and T. Graepel, "Biases for emergent communication in multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019.

[7] K. Wagner, J. A. Reggia, J. Uriagereka, and G. Wilkinson, "Progress in the simulation of emergent communication and language," *Adaptive Behaviour*, 2003.