Multi-Agent Reinforcement Learning with Epistemic Priors

Thayne T. Walker[1,2], Jaime S. Ide[1,3], Minkyu Choi[1,4], Michael Guarino[1] and Kevin Alcedo[1]

[1]Lockheed Martin Artificial Intelligence Center (LAIC), Applied AI Team, Shelton, CT, USA.
[2]University of Denver, Department of Computer Science, Denver, CO, USA.
[3]Yale University, Department of Psychiatry, New Haven, CT, USA.
[4]University of Texas at Austin, Department of Electrical and Computer Engineering, Austin, TX, USA.

Abstract

It is important for autonomous multi-agent teams to coordinate actions so collaborative goals can be achieved efficiently without conflicts. Without coordination, the goal may be achieved inefficiently, or in the worst case, not at all. Practical issues in multi-agent, real-time systems are limited sensing and communication capabilities. A significant number of multi-agent algorithms rely on accurate state information for all agents in order to effectively coordinate. In this paper, we propose an approach called Reinforcement Learning with Epistemic Priors (MARL-EP). MARL-EP uses epistemic estimation of the knowledge and actions of other agents from planners to infer portions of the observation space which are unobservable. We show that MARL-EP allows a very high level of coordination to be achieved with severely impaired sensing and zero communication between agents.

Introduction

In this paper, we seek solutions to multi-agent control problems, where autonomous agents must work together to achieve goals. It is important to coordinate actions between the agents in order to achieve a collaborative goal efficiently and without conflicts. Without coordination, in the best case a collective goal may be achieved inefficiently, in the worst case the collective goal may not be achieved at all. Multi-agent control has applications in warehouses, firefighting, surveillance, transportation, games and more.

Due to the actions of other agents or the workings of nature, the environment may change without an agent noticing. Under normal circumstances, perceived changes to the environment and intended actions may need to be communicated to other agents in order to ensure that the collective goal can still be achieved.

A practical issue which must be accounted for in the coordination of multiple agents in real-time, distributed systems is the possibility of communication loss. The most common communication paradigm is via radio signals. In this paradigm, loss of communication can occur for a variety of reasons such as: equipment failure, network failure or signal loss. Signal loss can occur for reasons such as signal range constraints, line of sight obstruction, signal jamming or noise. Communications links may also have limited bandwidth, reducing the amount of data that can be transmitted. Other constraints may limit the communications further, such as clandestine operations where agents intentionally attenuate or suppress the amplitude and/or volume (and hence probability of detection) of their radio emissions. Sensing may be impaired for a variety of similar reasons.

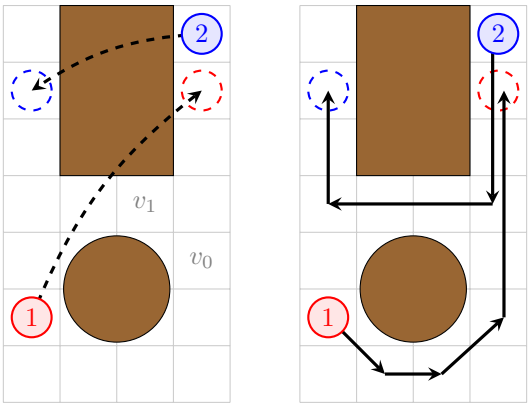
In all of these scenarios, a team of autonomous agents may need to coordinate with little to no sensing or communication. In this paper, we propose a system of coordination without communication called multi-agent reinforcement learning with epistemic priors (MARL-EP) that incorporates theory of mind (Premack and Woodruff 1978) in the form of epistemic logic (Bolander and Andersen 2011) embodied as a deterministic planner to create a cohesive multi-agent plan that incorporates the estimated knowledge of other agents. This information is then leveraged for reinforcement learning (RL) (Sutton and Barto 2018). When the state of teammates cannot be observed, observations are augmented using inferred state information from epistemic logic. We show that MARL-EP performs comparable to when agents have perfect information, even with severely impaired sensing and zero communication between agents.

Problem Definition

Figure 1 illustrates a motivating example in which agents must move from their respective start states (circles with solid lines) to their respective goal states (circles with dashed lines). This must be done without coming into conflict with other agents or obstacles (shown in brown).

This is similar to the cooperative navigation problem (Lowe et al. 2017), except everything in the environment is partially observable, and actions are stochastic. Agents have perfect knowledge of the start and goal states of all agents and obstacles, but (unlike the cooperative navigation problem) agents cannot communicate and they have a limited sensing range, meaning, during execution they may never have opportunity to directly perceive other agents.

In Figure 1, we can see that if each agent plans to reach its goal (using a shortest path) without taking the other agent into account, both agents will take paths that pass through v_1, shown in Figure 1(a) and hence will collide. Given the fact that both agents know the goal of the other agent, it is possible for the agents to independently conclude (without communication) that there is an (optimal) solution in which



(a) (b)

Figure 1: (a) An example instance of a cooperative navigation problem and (b) a solution for the problem instance.

agent 1 moves through label v_0 and agent 2 moves through label v_1 as shown in Figure 1(b). Making agents coordinate without any communication is the intent of MARL-EP.

Background

In this section, we cover the backgrounds of two disciplines: Multi-Agent Reinforcement Learning (MARL), and Epistemic Logic.

Multi-Agent Reinforcement Learning

The example from Figure 1 is a specific instance of a decentralized partially-observable Markov decision process (DEC-POMDP) (Bernstein et al. 2002), which can be expressed as a multi-agent version of POMDP (Kaelbling, Littman, and Cassandra 1998) represented by the tuple $\langle k, S, A, T, R, \Omega, O \rangle$ where:

- k is the number of agents,
- S: $\{S_1 \times ... \times S_k\}$ is a set of *joint* states,
- A: $\{A_1 \times ... \times A_k\}$ is a set of *joint* actions,
- T: $S \times A \rightarrow S$ is a stochastic transition function,
- R: $S \times A \rightarrow \mathbb{R}$ is the reward function.
- Ω is a set of observations and
- O: $S \times [1..k] \rightarrow \Omega$ is a stochastic observation function.

We seek a solution to the DEC-POMDP, a set of stochastic policies for each agent π_i: $S \times \Omega \rightarrow A_i$ which is a mapping from state-observation pairs to actions where $\pi_i(a_i|s_i, o_i)$ represents the probability of selecting an action $a_i \in A_i$ given the current joint state $s_i \in S_i$ and observation $o_i \in \Omega$. The goal of RL is to find π^*, an optimal policy that maximizes the sum of expected rewards.

For DEC-POMDPs, , independent Q-learning (IQL) (Tan 1993), the practice of decomposing a multi-agent problem into simultaneous single-agent problems, helps avoid issues related to the curse of dimensionality. Because of non-stationarity in simultaneously learned policies, distributed value functions (DVF) are commonly used in order to guarantee convergence (Sunehag et al. 2017). While DVFs can be used to train independent policies, even in limited communication scenarios (Matignon, Jeanpierre, and Mouaddib 2012), a significant number of algorithms solve DEC-POMDPs by training separate policies with a centralized DVF. This is done by counter-factually providing a complete action history to the centralized DVF at training time, though the observation space may be limited to locally-observable regions (Lowe et al. 2017; Rashid et al. 2018; Foerster et al. 2016). In (Lowe et al. 2017) the relative position of all agents is assumed to be known, allowing for effective global collaboration, but violating our assumptions about limited communication and sensing. In (Rashid et al. 2018), only the locally-observable portion of state is known to each agent, meaning macro-level collaboration between agents is not fostered. This is because the local state provides no clues of distinguishability from other global states which share the same local state. We seek to compensate for this shortcoming by using priors based on Epistemic Logic.

Epistemic Logic

Epistemic logic (Bolander and Andersen 2011), a type of modal logic (Zalta, Nodelman, and Allen 1995), deals with estimation of the knowledge of agents. This estimation of other's knowledge capitalizes on the notion of perspective shifts (Engesser et al. 2017). That is, one agent looking at the scenario from another agent's perspective. Collaboration without communication requires estimation of other agents' state and perception (Faulk and Frey 2021) and what actions they might take based on what they know (or what they do not know) (Engesser et al. 2017). This has a parallel with the theory of mind in which primates keep mental states of themselves and others (Premack and Woodruff 1978).

We propose that epistemic logic be used in POMDPs with limited sensing and communication, in order to allow agents make decisions in a richer, more informed observation space. Using epistemic priors with MARL allows agents to use what they know about other agents (e.g., their goals), and what they know that other agents know in order to foster collaboration in situations where parts of the state space cannot be observed and must be inferred.

Reinforcement Learning With Epistemic Priors

Per our assumption of zero communication and limited sensing, the content of an observation o_i would normally be limited to locally observable information. This limits the expressiveness of a policy because many observations are indistinguishable due to lack of information about the global state. To alleviate this, we propose a formulation of a policy π to include *epistemic priors*. Specifically, the action selection policy is now conditioned on epistemic information: $\pi_i(a_i|s_i, o_i, e_i)$ where e_i is the epistemic estimate. We refer to e_i generally as an *epistemic prior*. It was shown that estimation of other agent's actions results in improvement of the global effectiveness of the learned policies (Nagayuki et al. 2000). MARL-EP uses epistemic priors in order to fill in the gaps of local observations to more completely estimate global state, and increase the global optimality of policies.

Convention of Operation

Consider the following illustration: Before Alice left to her cottage for the week, she agreed to let Bob borrow her cottage for the weekend. Bob, having no further communication from Alice arrives at the cottage that weekend and finds the door locked. Bob realizes that Alice knows that she did not communicate anything about the key to the cottage. Bob also knows that Alice intends for him to use the key. Bob realizes that it is a common convention to leave a key under the door mat. Bob checks under the door mat and finds the key. We can now infer Alice's portion of the story. Alice locks up the cottage in preparation to leave. Alice knows that Bob intends to use the cottage. Alice realizes that she did not communicate about the key to Bob. Alice believes that Bob knows about the convention of leaving the key under the mat. Alice leaves the key under the mat and departs.

In both cases, the actors in the story understood the goal of the other actor (common knowledge). In both cases, the actor reasoned about what the other actor knows (epistemic logic). Both actors assumed a convention of operation to help determine their actions (epistemic planning).

The estimation of e_i is made possible by instantiating common knowledge of an a-priori *convention of operation*. Examples of such conventions are ubiquitous in the real world, for example protocols for transportation (e.g., driving on the right or left side of the road) collision avoidance in aviation (e.g., vertical separation) (Administration 2022) and playbooks in sports (Molineaux, Aha, and Sukthankar 2009). The conventions may differ depending on the specific problem setting. A convention of operation establishes common knowledge of a set of rules which help agents cooperate smoothly, with or without explicit communication.

Minimally, a convention of operation C: $S, G \rightarrow A$ is a function, which given a current state S, and goal state G, returns an action A. C must satisfy the following properties:

- **comprehensive** It must cover all possible states in the multi-agent state space.
- **deterministic** Given an input, it will always produce the same output.
- **chainable** It can be used to produce a trajectory.

For example, for the traffic convention "stay to the right", when two cars approach each other on a single-lane road (the state) moving in opposite directions (the goal), both cars move to the right (the action).

Manually developing a convention of operation (e.g., aviation protocols) can be a tedious and error-prone process. However, C is similar to an MDP-style (non-stochastic) agent, and can be embodied using a multi-agent RL policy or deterministic planner. In our experiments we use a search-based, multi-agent planner (Walker, Sturtevant, and Felner 2020) for C, but an RL policy (Rashid et al. 2018) could also be used to perform roll-outs in a deterministic way.

Assuming each agent has an identical copy of C, everything is in place for us to perform epistemic estimation. Because C is deterministic, each agent can safely assume that other agent's "interpretation" of C is identical.

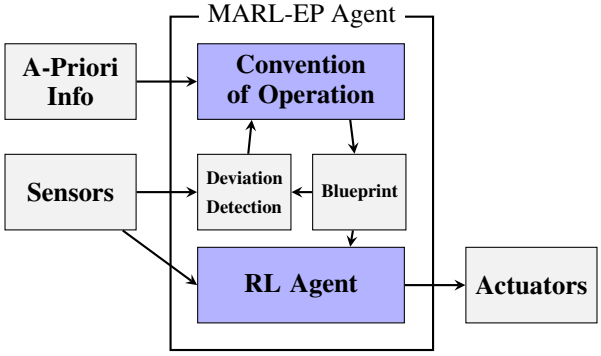


Figure 2: MARL-EP System Architecture.

Epistemic Blueprints

In the MARL-EP architecture (Figure 2), we endow epistemic capabilities to each individual agent by equipping them with a convention of operation in the form of a deterministic multi-agent planner. Assuming each agent invokes the planner on identical inputs, (by the definition of determinism) each agent will independently compute an identical multi-agent plan. We call this plan an *epistemic blueprint* (or just *blueprint*). Referring again to Figure 1, all agents compute the *full* multi-agent plan (i.e., blueprint) using their deterministic planner, then execute their own part of the multi-agent plan. Assuming agents' information about the world does not diverge, agents can safely execute their own portion of the plan from the blueprint and achieve implicit coordination without any communication.

Multi-Agent RL with Epistemic Priors

We use QMIX (Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning) (Rashid et al. 2018) as our learning algorithm. QMIX trains decentralized policies in a centralized end-to-end fashion. The key idea behind QMIX is to decompose the global value function (estimated over joint states and actions $\{S, A\}$) as a non-linear combination of local value functions (estimated over agent's local states and actions $\{o_i, a_i\}$). The central Q-network (Q_{tot}) is trained using the entire observation, while the policy networks (Q_i) are trained using only local observations. The mixing network is learned via a monotonic function that ensures that as the values of the local value functions increase, the global value function also increases. The monotonicity condition is represented by the expression: $\frac{\partial Q_{tot}}{\partial Q_i} \geq 0, \forall i \in I$, where I represents all agents. This factorized representation allows QMIX to capture interactions between agents while avoiding the exponential complexity growth that is typical in MARL. The QMIX algorithm has been shown to perform well in a variety of multi-agent environments, including StarCraft II micromanagement tasks and cooperative navigation tasks.

In this work, QMIX is used to train agents using solely local observations, but augmented with epistemic priors e_i from the blueprint generated by C, the deterministic planner. Pseudocode of our proposed approach is presented in

Algorithm 1: QMIX-EP, a modified version of the original method. Compared to the original algorithm, the main differences are in **Line 6**, in which we use the multi-agent planner (Walker, Sturtevant, and Felner 2020) to generate e_i for the particular episode, and in **Line 9** where these priors are added to the agents' local observation o_i. For brevity, we omit details of updating θ (**Line 17**).

Algorithm 1: QMIX with Epistemic Priors

1: Initialise θ: the parameters of mixing network, agent networks and hypernetwork;
2: Set the learning rate α and replay buffer $D = \{\}$;
3: $\text{step} = \theta, \theta^- = \theta$;
4: **while** $\text{step} < \text{step}_{max}$ **do**
5: $t = 0, s_0 = \text{initial state}$;
6: Estimate the epistemic priors e_i for each one of the agents i for the entire episode;
7: **while** $s_t \neq terminal$ and $t <$ episode limit **do**
8: **for** each agent i **do**
9: $\tau_t^i = \tau_{t-1}^i \cup \{(o_t, e_t, a_{t-1})\}$;
10: $\epsilon = \text{epsilon} - \text{schedule}(step)$;
11: $a_t^i = \arg\max_{a_t^i} Q_i(\tau_{t-1}^i, a_t^i)$ with $p = 1-\epsilon$;
12: **end for**
13: Get reward r_t and next state s_{t+1};
14: $D = D \cup \{(s_t, a_t, r_t, s_{t+1})\}$;
15: $t + 1$, step = step + 1;
16: **end while**
17: Update model parameters θ of mixing network, agent networks Q_i, and hypernetwork.
18: **end while**

Experimental Results

In order to validate the benefits of epistemic priors when training MARL agents, we use a popular task in the Multi-agent Particle Environment (MPE)[1], known as *Simple-Spread*. Given a particular number of agents and landmarks, the goal in the Simple-Spread task is to reach the landmarks as quickly as possible and avoid collision among the agents (Figure 3). At each time-step, agents receive a penalty and the episode is terminated when all landmarks are reached. The global reward is given by the sum of distances of each agent to all the landmarks. We train agents in the following scenarios with varying levels of information:

[1]https://github.com/openai/multiagent-particle-envs

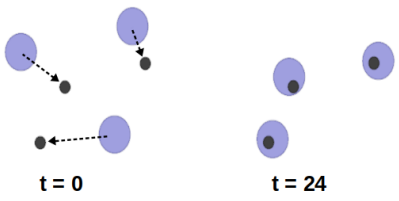


Figure 3: Simple-spread task. Agents cooperate to reach the landmarks quickly while avoiding collisions.

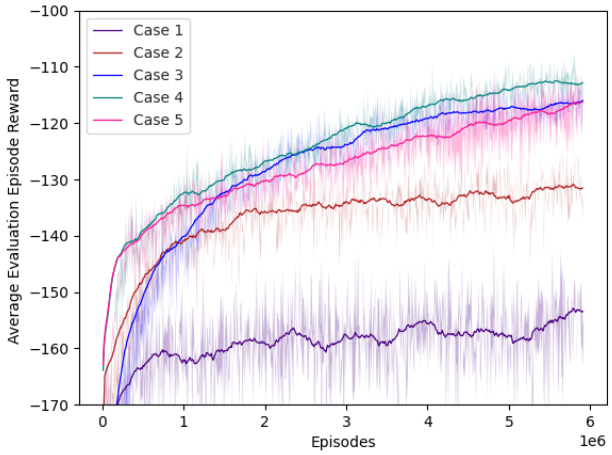


Figure 4: Evaluation of trained QMIX agents for different cases.

- **Case 1**: no sensing (baseline). Agents do not have access to other agents location;
- **Case 2**: limited sensing. Agents have access to other agents location when close;
- **Case 3**: perfect sensing. Agents know others' locations;
- **Case 4**: limited sensing, with priors (QMIX-EP). Agents have access to other agents location when close, and use estimated location otherwise.
- **Case 5**: no sensing, with priors (QMIX-EP). Similar to Case 1 but agents use estimated location of other agents;

In Figure 4, we depict the mean rewards for the different cases. With the use of epistemic priors in QMIX-EP (Cases 4 and 5), significant improvements are achieved even with no or limited sensing, compared to the standard QMIX (Cases 1 and 2). QMIX-EP (Cases 4 and 5) also have comparable performance to the perfect information Case 3 even with limited or no sensing. Finally, QMIX-EP also appears to increase the rate of training.

Conclusions

We have shown how the coordination of multiple agents with limited sensing and communication abilities can be done effectively by encoding a convention of operation for all agents and using it to produce comprehensive epistemic priors. We have shown that the use of epistemic priors is effective in fostering coordination among agents by allowing them to infer the actions of agents that they cannot observe. Using epistemic priors allows significant performance improvements versus without priors, and achieves performance levels similar to having perfect information. Future work includes online execution (real scenarios) and re-planning (real-time updating of blueprints).

References

Administration, F. A. 2022. FAA Order JO 7110.65Z - Air Traffic Control.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research*, 27(4): 819–840.

Bolander, T.; and Andersen, M. B. 2011. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1): 9–34.

Engesser, T.; Bolander, T.; Mattmüller, R.; and Nebel, B. 2017. Cooperative epistemic multi-agent planning for implicit coordination. *arXiv preprint arXiv:1703.02196*.

Faulk, D. K.; and Frey, T. L. 2021. Autonomous Collaboration in the Presence of Degraded Communication. In *AIAA Scitech 2021 Forum*, 1267.

Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Matignon, L.; Jeanpierre, L.; and Mouaddib, A.-I. 2012. Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes. In *Twenty-sixth AAAI conference on artificial intelligence*.

Molineaux, M.; Aha, D. W.; and Sukthankar, G. 2009. Beating the defense: Using plan recognition to inform learning agents. Technical report, KNEXUS RESEARCH CORP SPRINGFIELD VA.

Nagayuki, Y.; Ishii, S.; Ito, M.; Shimohara, K.; and Doya, K. 2000. A multi-agent reinforcement learning method with the estimation of the other agent's actions. In *Proceedings of the Fifth International Symposium on Artificial Life and Robotics*, volume 1, 255–259. Citeseer.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526.

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, 4295–4304. PMLR.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Walker, T. T.; Sturtevant, N. R.; and Felner, A. 2020. Generalized and Sub-Optimal Bipartite Constraints for Conflict-Based Search. In *AAAI Conference on Artificial Intelligence*.

Zalta, E. N.; Nodelman, U.; and Allen, C. 1995. Stanford encyclopedia of philosophy.