Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning

Extended Abstract

Pedro P. Santos
Instituto Superior Técnico, INESC-ID
Lisbon, Portugal
pedro.pinto.santos@tecnico.ulisboa.pt

Diogo S. Carvalho
Instituto Superior Técnico, INESC-ID
Lisbon, Portugal
diogo.s.carvalho@tecnico.ulisboa.pt

Miguel Vasco
KTH Royal Institute of Technology
Stockholm, Sweden
miguelsv@kth.se

Alberto Sardinha
Pontifical Catholic University of Rio de Janeiro, INESC-ID
Rio de Janeiro, Brazil
sardinha@inf.puc-rio.br

Pedro A. Santos
Instituto Superior Técnico, INESC-ID
Lisbon, Portugal
pedro.santos@tecnico.ulisboa.pt

Ana Paiva
Instituto Superior Técnico, INESC-ID
Lisbon, Portugal
paiva.a@gmail.com

Francisco S. Melo
Instituto Superior Técnico, INESC-ID
Lisbon, Portugal
fmelo@inesc-id.pt

ABSTRACT

We introduce *hybrid execution* in multi-agent reinforcement learning (MARL), a new paradigm in which agents aim to successfully complete cooperative tasks with arbitrary communication levels at execution time by taking advantage of information-sharing among the agents. Under hybrid execution, the communication level can range from a setting in which no communication is allowed between agents (fully decentralized), to a setting featuring full communication (fully centralized), but the agents do not know beforehand which communication level they will encounter at execution time. To formalize our setting, we define a new class of multi-agent partially observable Markov decision processes (POMDPs) that we name hybrid-POMDPs, which explicitly model a communication process between the agents.

KEYWORDS

Multi-Agent Reinforcement Learning; Reinforcement Learning; Multi-Agent Systems; Machine Learning.

ACM Reference Format:
Pedro P. Santos, Diogo S. Carvalho, Miguel Vasco, Alberto Sardinha, Pedro A. Santos, Ana Paiva, and Francisco S. Melo. 2024. Centralized Training with Hybrid Execution in Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) aims to learn utility-maximizing behavior in scenarios involving multiple agents. Deep MARL methods have been successfully applied to multi-agent tasks such as game-playing [8], traffic light control [11], or energy management [2]. Despite recent successes, the multi-agent setting is substantially harder than its single-agent counterpart [1].

As a way to deal with the exponential growth in the state/action space and with environmental constraints, both in perception and actuation, existing methods aim to learn decentralized policies that allow the agents to act based on local perceptions and partial information. The paradigm of *centralized training with decentralized execution* is at the core of recent research in the field [3, 7, 9]; such paradigm takes advantage of the fact that additional information, available only at training time, can be used to learn decentralized policies in a way that the need for communication is alleviated.

While in some settings partial observability and/or communication constraints require learning fully decentralized policies, the assumption that agents cannot communicate at execution time is often too strict for a great number of real-world scenarios [4, 12]. In such domains, learning fully decentralized policies should be deemed too restrictive since such policies do not take into account the possibility of communication between the agents. Other MARL strategies, which do take advantage of additional information shared among the agents, can surely be developed [13].

In this work, we propose RL agents that are able to exploit the benefits of centralized training while taking advantage of information-sharing at execution time. We introduce the paradigm of *hybrid execution*, in which agents act in scenarios with arbitrary (but unknown) communication levels that can range from no communication (fully decentralized) to full communication between the agents (fully centralized). In particular, we consider scenarios with faulty communication during execution, in which agents passively share their local observations to perform partially observable cooperative

tasks. We formalize the setting of hybrid execution in MARL by introducing *hybrid partially observable Markov decision processes* (H-POMDPs), a new class of multi-agent POMDPs.

2 HYBRID EXECUTION IN MULTI-AGENT RL

A fully cooperative multi-agent system with Markovian dynamics can be modeled as a decentralized partially observable Markov decision process (Dec-POMDP) [6]. A Dec-POMDP is a tuple

$$([n], \mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, r, \gamma, \mathcal{Z}, \mathcal{O}),$$

where $[n] = \{1, \ldots, n\}$ is the set of indexes of n agents, $\mathcal{X}$ is the set of states of the environment, $\mathcal{A} = \times_i \mathcal{A}_i$ is the set of joint actions, where $\mathcal{A}_i$ is the set of individual actions of agent i, $\mathcal{P}$ is the set of probability distributions over next states in $\mathcal{X}$, one for each state and action in $\mathcal{X} \times \mathcal{A}$, $\mathcal{P}_0$ is the probability distribution over initial states, $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ maps states and actions to expected rewards, $\gamma \in [0, 1[$ is a discount factor, $\mathcal{Z} = \times_i \mathcal{Z}_i$ is the set of joint observations, where $\mathcal{Z}_i$ is the set of local observations of agent i, and $\mathcal{O}$ is the set of probability distributions over joint observations in $\mathcal{Z}$, one for each state and action in $\mathcal{X} \times \mathcal{A}$. A decentralized policy for agent i is $\pi_i : \mathcal{Z}_i \to \mathcal{A}_i$ and the joint decentralized policy is $\pi : \mathcal{Z} \to \mathcal{A}$ such that $\pi(z_1, \ldots, z_n) = \big(\pi_1(z_1), \ldots, \pi_n(z_n)\big)$.

Fully decentralized approaches to MARL directly apply standard single-agent RL algorithms for learning each agent's policy π_i in a decentralized manner [10]. More recently, under the paradigm of centralized training with decentralized execution, methods such as QMIX [9] aim at learning decentralized policies with centralization at training time while fostering cooperation among the agents. Finally, if we know that all agents can share their local observations among themselves at execution time, we can use any of the approaches above to learn fully centralized policies.

None of the aforementioned classes of methods assumes, however, that agents may sometimes have access to other agents' observations and sometimes not. Therefore, decentralized agents are unable to take advantage of the additional information that they may receive from other agents at execution time, and centralized agents are unable to act when the sharing of information fails. In this work, we introduce hybrid execution in MARL, a setting in which agents act regardless of the communication process while taking advantage of additional information they may receive during execution. To formalize this setting, we define a new class of multi-agent POMDPs, named hybrid-POMDPs (H-POMDPs), which explicitly considers a communication process among the agents.

2.1 Hybrid Partially Observable Markov Decision Processes

We define a hybrid-POMDP (H-POMDP) as a tuple

$$([n], \mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, r, \gamma, \mathcal{Z}, \mathcal{O}, C)$$

where, in addition to the tuple that describes the Dec-POMDP, we consider a $n \times n$ communication matrix C such that $[C]_{i,j} = p_{i,j}$ is the probability that, at a certain time step, agent i has access to the local observation of agent j in $\mathcal{Z}_j$. H-POMDPs generalize both the notion of decentralized execution and centralized execution in MARL. Specifically, for a given Dec-POMDP, we can consider C as the identity matrix to capture fully decentralized execution or as a matrix of ones to capture fully centralized execution.

In our setting, we assume that at execution time agents will face an H-POMDP with an unknown communication matrix C, sampled from a set $\mathcal{C}$ according to an unknown probability distribution μ. The performance of the agent is measured as $J_\mu(\pi) = \mathbb{E}_{C \sim \mu}\left[J(\pi; C)\right]$, where $J(\pi; C)$ denotes the expected discounted cumulative reward of policy π under an H-POMDP with communication matrix C. At training time, the agents have access to the fully centralized H-POMDP. Thus, the setting we consider is one of centralized training with *hybrid* execution and an unknown communication process.

Connecting H-POMDPs and Dec-POMDPs: We can cast every H-POMDP as a Dec-POMDP as follows: (i) the observation space of each agent corresponds to the joint observation space, adequately tuned by including an additional token to encode missing observations; (ii) the emission function consists of a masking function that makes only a subset of observations visible to each agent at each timestep, as parameterized by C; and (iii) the remainder elements are the same. However, we seek to find a method that can act on H-POMDPs regardless of the matrix C. To accommodate such fact, we can extend the state of the Dec-POMDP to include matrix C and redefine the distribution over the initial states of the Dec-POMDP to also encode the uncertainty over the communication matrix C as parameterized by μ. Even though there exist connections between H-POMDPs and Dec-POMDPs, the introduction of the H-POMDPs formulation is still relevant as it succinctly encodes the various degrees of centralization that arise in recent MARL research.

3 CONCLUSION

In this work, we introduced the paradigm of centralized training with hybrid execution, which we formalized by introducing a new class of multi-agent POMDPs, named H-POMDPs.

Given the aforementioned connections between H-POMDPs and Dec-POMDPs, we expect the worst-case computational complexity of solving H-POMDPs to be similar to that of Dec-POMDPs, known to be NEXP-complete for finite-horizon problems [5]. Nevertheless, efficient methods to approximately solve H-POMDPs that take advantage of the specificities of the paradigm of hybrid execution can still surely be developed. We leave the study of such methods for future work.

ACKNOWLEDGMENTS

This work was supported by Portuguese national funds through the Portuguese Fundação para a Ciência e a Tecnologia (FCT) under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) (INESC-ID multi-annual funding), PTDC/CCI-COM/5060/2021 (RELEvaNT), and PTDC/CCI-COM/7203/2020 (HOTSPOT). In addition, this research was supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No. 952215, and by the Air Force Office of Scientific Research under award number FA9550-22-1-0475. Pedro P. Santos acknowledges the FCT PhD grant 2021.04684.BD, and Diogo S. Carvalho the FCT PhD grant 2020.05360.BD. This work has also been supported by the Swedish Research Council, Knut and Alice Wallenberg Foundation and the European Research Council (ERC-BIRD).

REFERENCES

[1] Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. 2021. Multi-Agent Reinforcement Learning: A Review of Challenges and Applications. *Applied Sciences* 11, 11 (2021).

[2] Xiaohan Fang, Jinkuan Wang, Guanru Song, Yinghua Han, Qiang Zhao, and Zhiao Cao. 2020. Multi-Agent Reinforcement Learning Approach for Residential Microgrid Energy Scheduling. *Energies* 13, 1 (2020).

[3] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. *CoRR* abs/1605.06676 (2016).

[4] Florence Ho, Ana Salta, Ruben Geraldes, Artur Goncalves, Marc Cavazza, and Helmut Prendinger. 2019. Multi-Agent Path Finding for UAV Traffic Management. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 131–139.

[5] Frans Oliehoek and Christopher Amato. 2016. A Concise Introduction to Decentralized POMDPs. (01 2016). https://doi.org/10.1007/978-3-319-28929-8

[6] Frans A. Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.

[7] Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos Vlassis. 2011. Optimal and Approximate Q-value Functions for Decentralized POMDPs. *CoRR* abs/1111.0062 (2011).

[8] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2020. Comparative Evaluation of Multi-Agent Deep Reinforcement Learning Algorithms. *CoRR* abs/2006.07869 (2020).

[9] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *CoRR* abs/1803.11485 (2018).

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

[11] Hua Wei, Guanjie Zheng, Vikash V. Gayah, and Zhenhui Li. 2019. A Survey on Traffic Signal Control Methods. *CoRR* abs/1904.08117 (2019).

[12] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access* 8 (2020), 58443–58469.

[13] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2022. A Survey of Multi-Agent Reinforcement Learning with Communication. https://doi.org/10.48550/ARXIV.2203.08975