Role of State in Partially Observable Reinforcement Learning

Doctoral Consortium

Andrea Baisero
Northeastern University
Boston, USA
baisero.a@northeastern.edu

ABSTRACT

Reinforcement learning agents that act under partial observability lack access to the environment state, and may need to account for the observable history to select actions optimally. However, offline training paradigms (e.g., training via simulator) are able to exploit state information during the training phase to improve learning performance. The literature contains a number of such methods that exploit state information during training, and empirically demonstrate superior performance during evaluation, e.g., asymmetric actor-critic methods that use state critics. However, such methods tend to be poorly motivated and lack a theoretically sound justification. In this work, we focus on the theoretical and practical consequences of using states to train partially observable agents, and propose interpretations to explain the role of state.

KEYWORDS

Reinforcement Learning; Partial Observability; Asymmetric Reinforcement Learning; Privileged Information

ACM Reference Format:
Andrea Baisero. 2025. Role of State in Partially Observable Reinforcement Learning: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Partially observable control is characterized by agents that are limited by acting upon indirect, partial, stochastic, and/or noisy observations of the environment state. This "simple" limitation has significant implications on problem complexity, the agent's learning process, and the resulting optimal behaviors. Optimal control under partial observability is characterized by two fundamental emergent properties compared to its fully-observable counterpart: (a) *information-gathering*, which refers to behaviors that reveal new information to the agent, e.g., opening a drawer, turning around a corner, etc. (b) *memorization*, which refers to the general necessity of acting based on past observations (a.k.a. the *observable history* h) rather than only the most recent observation [6, 14].

Single-agent partially observable control is commonly formalized as a *partially observable Markov decision process* (POMDP) [4, 15], and model-free agents are commonly modeled by policy functions that map an agent's *observable history* into an action distribution, $\pi \colon \mathcal{H} \to \Delta\mathcal{A}$. As expected by the problem setting, the policy interface lacks any direct notion of state. It is even possible to convert a POMDP into an equivalent MDP that completely integrates out the notion of state, e.g., Belief-MDPs [4]. Belief-MDPs are defined in terms of *beliefs*, i.e., distributions over the state space corresponding to the agent's uncertainty. Belief-MDPs are a clear demonstration of a fundamental property of partially observable control: the state itself does not matter; only the belief-state does.

Nonetheless, state plays a clear role in the underlying dynamics of any partially observable control problem: it is directly affected by agent actions via the state dynamics $T \colon \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$, and it directly determines rewards via the reward function $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. In that regard, state represents a form of *privileged* information of clear importance even for partially observable control, and methods that are able to exploit state while adhering to the stateless policy interface are able to achieve significant boosts in learning performance. Exploiting state information is possible in certain *offline* training paradigms[1] where the state is available during the training phrase but not during the execution phase, e.g., when training is performed via a simulated environment. In multi-agent control, this training paradigm is well-known as *centralized training for decentralized execution* (CTDE)[2] [1, 10]. In single-agent control, this has also been called *offline training for online execution* (OTOE) [3].

It is no surprise that a significant amount of effort has been spent on developing methods that exploit privileged state information, including single-agent methods [2, 3, 11, 17], multi-agent gradient-based methods [5, 7, 8, 19], and multi-agent value decomposition methods [9, 12, 13, 18]. However, many such methods often propose practical algorithms that are validated empirically by their performance, but often lack an adequate theoretical motivation for how state is used. Even though the latent state is an important component of partially observable control, and even though offline training paradigms provide a simple way to access state to train partially observable agents, misuses may lead to unintended or even catastrophic consequences [3, 8, 9], and properly considering the consequences of how state is used remains critically important.

2 ASYMMETRIC ACTOR-CRITIC FOR PORL

In this section, we provide a very brief overview of asymmetric actor-critic for single-agent partially-observable control.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1] Not to be confused with *offline* RL, a.k.a. *batch* RL, where training is performed on a static dataset rather than dynamic interactions with the environment.
[2] Although CTDE primarily focuses on sharing observations among all agents during training (another form of privileged information), the use of state is also common.

(Symmetric) Actor-Critic. In vanilla actor-critic, a history critic model $\hat{V}(h)$ is trained to evaluate the history policy $\pi(h)$, and used to obtain low-variance estimates of the policy gradient [16].

(Biased) Asymmetric Actor-Critic. Pinto et al. [11] develop a form of asymmetric actor-critic where the history critic $\hat{V}(h)$ is replaced with a state critic $\hat{V}(s)$. Aside from the change in input, the critic model is fundamentally trained and used in same way to obtain policy gradient estimates. For the better *and* for the worse, such a small change can have great implications on the resulting algorithm.

Unbiased Asymmetric Actor-Critic. Baisero and Amato [2] expose a number of theoretical and practical issues with the use of state values $\hat{V}(s)$ to evaluate history policies in partially-observable control problems. Such issues range from the state being inadequate to evaluate history-dependent behaviors, to the state value function of history policies $V^{\pi}(s)$ being not well-defined in partially-observable control. To resolve these issues, Baisero and Amato propose the use of *history-state* values $\hat{V}(h, s)$, another minor modification of previous methods. Such a small change resolves all the theoretical and practical issues with state values, and result in superior performance in environments with significant partial observability.

3 INTERPRETATIONS OF STATE

The mechanisms through which state values improve learning performance remain not well understood. We underscore yet again that partially observable control *should* be invariant to latent states, and that only the corresponding belief-states should matter. How is it, then, that asymmetric methods are able to outperform non-asymmetric counterparts so strongly? Finding the answer to this question will result in a better understanding of asymmetric methods, and help us design better asymmetric methods in the future.

We formulate four (non mutually exclusive) mechanisms through which state affects the training process and may hypothetically benefit asymmetric methods. The first two ("state as information" and "state as a feature") examine the state in terms of its information content compared to history. The second two ("state as exploration" and "state as bootstrapping") examine the state in terms of its direct effects on the optimization and learning processes.

State as Information. Interpreting the state in terms of information content is possibly the simplest interpretation to consider. It is natural to consider that the state may be useful to the training process of the agent by virtue of the additional information that it provides compared to the observable history. This information scales with state uncertainty, which implies that it is more effective in problems with high amounts of partial observability. According to this interpretation, a learning agent that evaluates its own actions based on the additional context of the groundtruth state is better equipped to determine the quality of its actions.

State as a Feature of History. It is also plausible to interpret the state information not as extraneous to the history, but as a different representation of the information that is intrinsic to the history itself. According to this interpretation, state can be viewed as a stochastic feature of history. A history h is implicitly associated with a belief $b(h) \in \Delta\mathcal{S}$ that represents a sufficient statistic for the purposes of control. At the same time, the state appears to be a sample from the belief $s \sim b(h)$, and therefore a stochastic realization of this sufficient statistic. It is therefore possible to reinterpret state as a feature of history in partially observable control. In other words, state represents information that is *intrinsic* to the history.

State as Exploration. State critics inject uncertainty-dependent variance into the policy gradient estimates [8], consequently affecting the convergence properties of the agent's optimization process. Though estimation variance is generally considered an issue to be mitigated, it can also provide some advantages, e.g., by overcoming plateaus or shallow local optima. Therefore, state values may be interpretable as a form of uncertainty-driven exploration.

State as Bootstrapping. Due to their sequential nature, histories are intrinsically more complex than states, and require sequence models that are significantly harder to train than the simpler models required by non-sequential states. Although state does not represent the correct type of information for partially observable control, it still provides information that is adjacent and relevant enough to bootstrap the training of better critic values. Therefore, is it possible that state values are helpful simply by virtue of useful state features $\phi(s)$ being significantly easier to learn than useful history features $\phi(h)$, especially in the earlier stages of training. If so, state values could cause a bootstraping effect that allows the agent to start learning before appropriate history features are learned.

Evaluation Methodology. We emphasize that the proposed interpretations of state are not necessarily mutually exclusive. Further, partial observability is a wide spectrum, and control problems may differ greatly, which may affect the importance of each interpretation. For example, "state as information" is more justified in low observability problems characterized by higher state uncertainty, whereas "state as a feature" is more justified in high observability problems characterized by lower state uncertainty. On the other hand, "state as exploration" may be adaptive to the amount of observabilty, given that state value variance is directly related to belief entropy. Similarly, "state as bootstrapping" may be equally justified both in high and low observability problems, since state features are easier to learn than history features in both cases.

It seems unlikely that a single evaluation could result in a single clear answer to our primary question. To evaluate these hypothesized roles of state, we consider a variety of empirical experiments that examine state values from different angles: (a) Evaluation of a variant of asymmetric actor-critic that employs partial state values rather than full states; (b) Evaluation of a variant of asymmetric actor-critic that employs counter-factual state values sampled independently from the current belief $\tilde{s} \sim b(h)$; (c) Evaluation of a variant of actor-critic that employs history values injected with artificial noise such that the resulting variance profile is comparable to that of state values; (d) Evaluation of a tabular variant of asymmetric actor-critic that does not employ value-function approximation models, and performs no generalization across histories and states; (e) Evaluation of the learning performance of the critics themselves; (f) Evaluation of the relative importance of state and history features in determining the critic's output.

ACKNOWLEDGMENTS

This work was partially funded by the NSF award number 2044993.

REFERENCES

[1] Christopher Amato. 2024. An Introduction to Centralized Training for Decentralized Execution in Cooperative Multi-Agent Reinforcement Learning. https://doi.org/10.48550/arXiv.2409.03052 arXiv:2409.03052 [cs].

[2] Andrea Baisero and Christopher Amato. 2022. Unbiased Asymmetric Reinforcement Learning under Partial Observability. In *Proceedings of the Conference on Autonomous Agents and Multiagent Systems*.

[3] Andrea Baisero, Brett Daley, and Christopher Amato. 2022. Asymmetric DQN for Partially Observable Reinforcement Learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

[4] Anthony R. Cassandra, Leslie Pack Kaelbling, and Michael L. Littman. 1994. Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence (AAAI'94)*. AAAI Press, Seattle, Washington, 1023–1028.

[5] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[6] Matthew Hausknecht and Peter Stone. 2015. Deep Recurrent Q-Learning for Partially Observable MDPs. In *2015 aaai fall symposium series*.

[7] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html

[8] Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. 2023. On Centralized Critics in Multi-Agent Reinforcement Learning. *Journal of Artificial Intelligence Research* (2023).

[9] Enrico Marchesini, Andrea Baisero, Rupali Bhati, and Christopher Amato. 2025. On Stateful Value Factorization in Multi-Agent Reinforcement Learning. (2025). Proceedings of the Conference on Autonomous Agents and Multiagent Systems.

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

[11] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. 2018. Asymmetric Actor Critic for Image-Based Robot Learning, Vol. 14. https://roboticsproceedings.org/rss14/p08.html

[12] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 10199–10210. https://proceedings.neurips.cc/paper_files/paper/2020/hash/73a427badebe0e32caa2e1fc7530b7f3-Abstract.html

[13] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51. http://jmlr.org/papers/v21/20-081.html

[14] Satinder P. Singh, Tommi S. Jaakkola, and Michael I. Jordan. 1994. Learning without state-estimation in partially observable Markovian decision processes. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning (ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 284–292.

[15] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning, second edition: An Introduction*. MIT Press. Google-Books-ID: uWV0DwAAQBAJ.

[16] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, Vol. 12. MIT Press. https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html

[17] Andrew Wang, Andrew C. Li, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. Mcilraith. 2023. Learning Belief Representations for Partially Observable Deep RL. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 35970–35988. https://proceedings.mlr.press/v202/wang23p.html ISSN: 2640-3498.

[18] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2020. QPLEX: Duplex Dueling Multi-Agent Q-Learning. https://openreview.net/forum?id=Rcmk0xxIQV

[19] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 24611–24624. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html