Leveraging Fully-Observable Solutions for Improved Partially-Observable Offline Reinforcement Learning

Extended Abstract

Chulabhaya Wijesundara
Northeastern University, STR
Boston, United States
wijesundara.c@northeastern.edu

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

Gregory Castañón
STR
Boston, United States
gregory.castanon@str.us

Alan Carlin
STR
Boston, United States
alan.carlin@str.us

Robert Platt
Northeastern University
Boston, United States
rplatt@ccs.neu.edu

Christopher Amato
Northeastern University
Boston, United States
c.amato@northeastern.edu

ABSTRACT

Offline reinforcement learning (RL) is valuable in settings where online interactions with an environment are impractical. While such settings are often partially-observable, existing offline RL methods typically focus on fully-observable Markov decision processes (MDPs) rather than partially-observable MDPs (POMDPs). To help close that gap, we present an offline RL algorithm for POMDPs that leverages expert policies from simpler, fully-observable versions of environments in an asymmetric learning setting. We provide theoretical grounding for how overlap between MDPs and POMDPs can be exploited to improve learning in the partially-observable setting, and our experiments empirically demonstrate that our method significantly improves performance compared to existing state-of-the-art MDP offline RL algorithms.

KEYWORDS

Reinforcement learning, partial observability, offline RL

ACM Reference Format:
Chulabhaya Wijesundara, Andrea Baisero, Gregory Castañón, Alan Carlin, Robert Platt, and Christopher Amato. 2025. Leveraging Fully-Observable Solutions for Improved Partially-Observable Offline Reinforcement Learning: Extended Abstract. 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

Many state-of-the-art offline reinforcement learning (RL) methods [8–10] are evaluated on fully-observable Markov decision process (MDP) data, whereas real-world problems are often characterized by partial observability due to sensor limitations and noise.

We approach this challenge through the lens of asymmetric RL [2, 3, 12], where the agent has access *during training* to privileged information such as the state and a fully-observable expert policy [11, 13–15], that may be exploited to train a partially-observable policy. To this end, we assume that the offline dataset $\mathcal{D}$ contains state information (as can be gathered, e.g., by a simulator). We propose *Cross-Observability Conservative Q-Learning* (CO-CQL), a new offline RL algorithm that exploits asymmetric learning from a fully-observable expert for partially-observable control.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

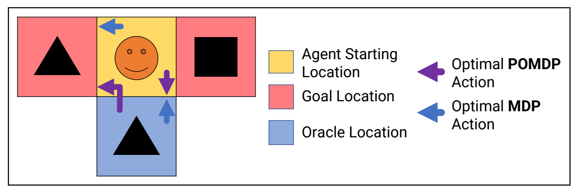

Figure 1: Simplified *Heaven-Hell* environment. An optimal partially-observable agent must visit an *oracle* (bottom blue location) while an optimal fully-observable agent does not.

Related Work. CO-CQL is closely related to Consevative Q-learning (CQL) [10], Conservative Soft Actor-Critic (CSAC) [10] and Cross-Observability Soft Imitation Learning (COSIL) [11]. CQL and CSAC combine value-based and actor-critic methods with a conservative regularizer $\mathcal{R}(Q) \doteq \mathbb{E}_{s\sim\mathcal{D}}\left[\max_a Q(s,a)\right] - \mathbb{E}_{s,a\sim\mathcal{D}}\left[Q(s,a)\right]$ that minimizes the gap between maximal and in-distribution values. COSIL augments the RL rewards with divergence-based pseudo-rewards $R(s_t,a_t) - \alpha D\left(\mu(s_t), \pi(h_t)\right)$ that promote similarity between partially-observable policy and fully-observable expert.

2 CROSS-OBSERVABILITY CONSERVATIVE Q-LEARNING

CO-CQL exploits the demonstrations of expert fully-observable policy to help guide the training of a partially-observable policy.

Consider a simplified variant of *Heaven-Hell* [5] shown in Figure 1. The agent must identify and reach the *good* exit while avoiding the *bad* exit. As a fully-observable problem, the agent directly observes the identity of the *good* exit. As a partially-observable problem, the agent must first visit an *oracle* to identify the exits and reduce its state uncertainty, and then backtrack to reach the *good* exit. Although the optimal partially-policy and the optimal fully-observable policies differ, there are several history/state contexts that overlap in terms of optimal behaviors. In these contexts, the fully-observable agent is able to provide relevant guidance to the partially-observable agent.

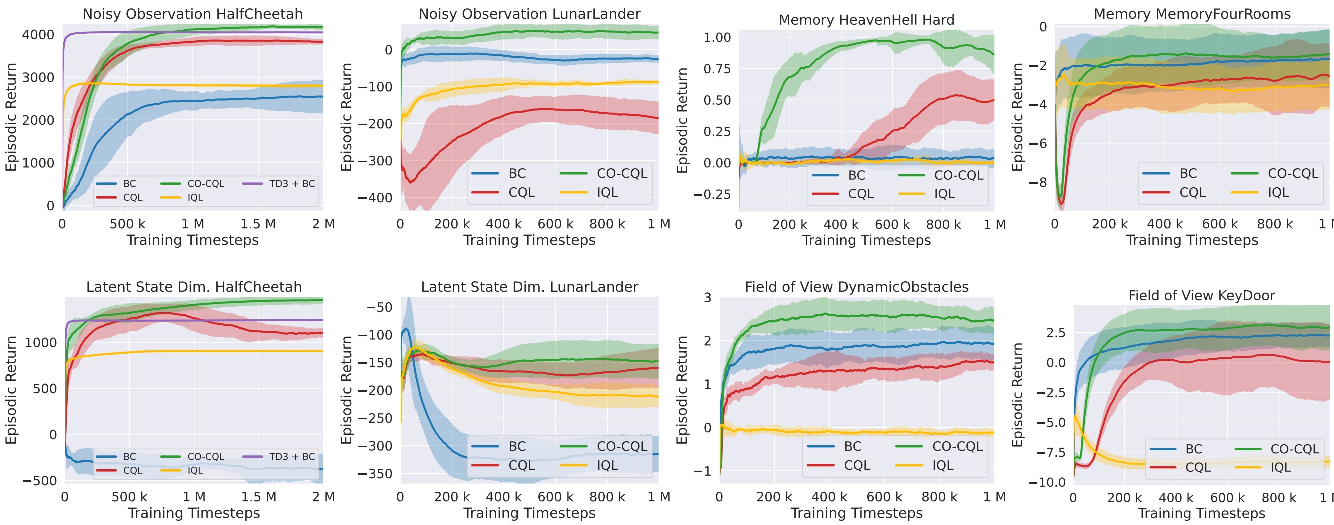

Figure 2: Mean and standard deviation of learning performance measured over 5 independent runs.

CO-CQL makes several extensions to vanilla fully-observable CQL, CSAC, and COSIL: (a) To exploit the fully-observable expert, we add a behavior cloning auxiliary objective; (b) to adequately handle the partially-observable data, state-based models are replaced with history-based models (e.g., $\pi(s)$ becomes $\pi(h)$, $Q(s,a)$ becomes $Q(h,a)$), using recurrent networks to process sequential data; (c) to additionally handle discrete control problems, we replace the underlying continuous SAC algorithm with discrete SAC [7]. The critic model is trained on an augmented conservative loss,

$$\begin{aligned} J^{Q}_{\text{CO-CQL}} &= \frac{1}{2}\,\mathbb{E}_{h,a,r,o\sim\mathcal{D}}\left[(y - Q(h,a))^2\right] + \lambda\mathcal{R}(Q)\,, \text{where} \\ y &= r + \gamma\,\mathbb{E}_{a'\sim\pi(hao)}\left[Q(hao,a') - \alpha\log\pi(a' \mid hao)\right], \end{aligned}$$

and the agent policy is trained on an augmented policy loss,

$$\begin{aligned} J^{\pi}_{\text{CO-CQL}} &= \mathbb{E}_{h,s\sim\mathcal{D},a\sim\pi(h)}\left[\alpha\log\pi(a \mid h) - Q(h,a)\right] \\ &\quad + \beta\,\mathbb{E}_{h,s\sim\mathcal{D}}\left[D\left(\mu(s)\,||\,\pi(h)\right)\right]. \end{aligned}$$

The behavior cloning term is interpretable as a form of imitation learning that projects the fully-observable expert behavior in partially-observable behavior space. In an online setting such as the one used in COSIL [11], the behavior cloning term can save exploration time, as the fully-observable expert already has knowledge about the fully-observable dynamics. In the offline setting of CO-CQL, this contribution by the fully-observable expert is particularly useful as exploration of new interactions is not allowed.

3 EVALUATION

We evaluate CO-CQL on discrete and continuous partially-observable control problems exhibiting a variety of challenges related to sensor noise, latent state variables, limited first-person POV, and memorization of the past. In discrete environments, we compare CO-CQL to recurrent CQL [10], recurrent IQL [9], and naive behavior cloning (BC) [1] from the fully-observable expert policy. In discrete environments, we additionally compare CO-CQL to recurrent TD3+BC [8].

The results in Figure 2 show that the performance of CO-CQL either exceeds or matches that of other baselines, demonstrating the efficacy of using state information in an asymmetric learning setting to inform the training of a partially-observable policy.

These results also demonstrate that CO-CQL generalizes well across a wide variety of partially-observable tasks. For example, *Half-Cheetah* and *Lunar-Lander* [6] require learning to handle complex continuous controls, *Heaven-Hell* [5] requires long-term memorization of the past, whereas *Memory-Four-Rooms*, *Dynamic-Obstacles*, and *Key-Door* [4] require processing discrete image observations with small fields of view that result in state aliasing. CO-CQL performs consistently well across environments, whereas the other baselines exhibit a trade-off by performing well in some environments and less so in others. This is particularly demonstrated by the relative performances of BC, CQL, and CO-CQL in the various environments; even when BC or CQL perform sub-optimally, CO-CQL is able to exploit the benefits both approaches to consistently obtain better performances.

4 CONCLUSION

In this work, we demonstrated that fully-observable expert policies can be used to address partially-observable offline RL. We created multiple novel partially-observable offline RL datasets from a variety of challenging environments, implemented recurrent versions of existing offline RL algorithms, and developed CO-CQL, a novel algorithm that exploits a fully-observable expert policy by combining RL with behavior cloning component and conservative value regularization. Our method primarily requires a dataset that also contains state information, e.g., as provided by a simulator. Access to a fully-observable expert policy appears as an additional requirement; however, is in principle obtainable by running fully-observable offline RL methods on the same dataset. Our approach performs better than state-of-the-art algorithms across a broad range of partially-observable environments.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Michael Bain and Claude Sammut. 1999. A Framework for Behavioural Cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*. Oxford University, GBR, 103–129.

[2] Andrea Baisero and Christopher Amato. 2022. Unbiased Asymmetric Reinforcement Learning under Partial Observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 44–52.

[3] Andrea Baisero, Brett Daley, and Christopher Amato. 2022. Asymmetric DQN for partially observable reinforcement learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 180)*, James Cussens and Kun Zhang (Eds.). PMLR, 107–117. https://proceedings.mlr.press/v180/baisero22a.html

[4] Andrea Baisero and Sammie Katt. 2021. gym-gridverse: Gridworld domains for fully and partially observable reinforcement learning. https://github.com/abaisero/gym-gridverse.

[5] Bonet Blai and Hector Geffner. 1998. Solving large POMDPs using real time dynamic programming. In *Working Notes Fall AAAI Symposium on POMDPs*, Vol. 218. https://bonetblai.github.io/reports/fall98-pomdp.pdf

[6] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. https://doi.org/10.48550/arXiv.1606.01540 arXiv:1606.01540 [cs].

[7] Petros Christodoulou. 2019. Soft Actor-Critic for Discrete Action Settings. https://doi.org/10.48550/arXiv.1910.07207 arXiv:1910.07207 [cs, stat].

[8] Scott Fujimoto and Shixiang Gu. 2021. A Minimalist Approach to Offline Reinforcement Learning. https://openreview.net/forum?id=Q32U7dzWXpc

[9] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline Reinforcement Learning with Implicit Q-Learning. https://openreview.net/forum?id=68n2s9ZJWF8

[10] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1179–1191. https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html

[11] Hai Huu Nguyen, Andrea Baisero, Dian Wang, Christopher Amato, and Robert Platt. 2022. Leveraging Fully Observable Policies for Learning under Partial Observability. https://openreview.net/forum?id=pn-HOPBioUE

[12] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. 2018. Asymmetric Actor Critic for Image-Based Robot Learning. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov (Eds.). https://doi.org/10.15607/RSS.2018.XIV.008

[13] Idan Shenfeld, Zhang-Wei Hong, Aviv Tamar, and Pulkit Agrawal. 2023. TGRL: An Algorithm for Teacher Guided Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1287, 17 pages.

[14] Andrew Warrington, Jonathan W. Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. 2021. Robust Asymmetric Learning in POMDPs. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 11013–11023. https://proceedings.mlr.press/v139/warrington21a.html ISSN: 2640-3498.

[15] Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. 2021. Bridging the Imitation Gap by Adaptive Insubordination. https://openreview.net/forum?id=Wlx0DqiUTD_