The Wasserstein Believer
Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models
Extended Abstract

Raphael Avalos[1 *] Florent Delgrange[1,2 *]
Ann Nowé[1 †] Guillermo A. Pérez[2,3 †] Diederik M. Roijers[1,4 †]

[1]Vrije Universiteit Brussel (Belgium) [2]University of Antwerp (Belgium)
[3]Flanders Make (Belgium) [4]City of Amsterdam (The Netherlands)

Reinforcement learning (RL) [16] is the set of machine learning algorithms that allow an agent to learn a *control policy* by solely interacting with an environment, in order to achieve its design objectives.

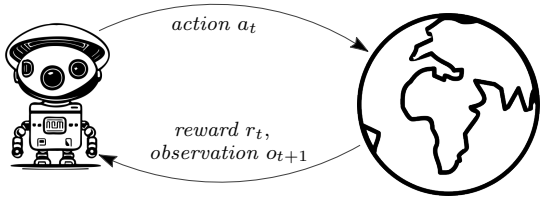

Fig. 1. RL interaction loop.

At each step $t \geq 0$ of this interaction, the agent performs an *action* a_t in the environment and, as outcome, it receives a *reward* r_t and an *observation* of the *next state* of the environment o_{t+1} (Fig. 1). The typical goal of an RL agent is to learn a *policy* π that prescribes which action to choose at each step and allows optimizing the *return*, i.e., the expected cumulative sum of rewards. The environment is generally assumed *unknown*, i.e., the agent does not have an explicit access to a *model* describing all its possible dynamics.

Partial observability. When the state of the environment is fully observable, policies solely based on the last observation $a_t \sim \pi(\cdot \mid o_t)$ are sufficient for optimal control. However, in real-world applications, this is rarely the case; the environment is often *partially observable*, meaning the observation perceived by the agent is not necessarily a *sufficient statistic* to optimize the return. This occurs in many scenarios such as robotics (e.g., using lidars or cameras) [11], recommendation systems [17], and autonomous vehicles [12]. As such, optimal policies must take the entire interaction *history* into account to take decisions accordingly: $a_t \sim \pi(\cdot \mid \text{history} = a_0, o_1, a_1, \ldots, o_t)$. Since the space of possible histories scales exponentially in the length of the episode, using histories to condition policies is intractable. To overcome those challenges, SOTA algorithms compress the history into a fixed-size vector with the help of *recurrent neural networks* (RNNs) [8,3,6,7,9,13,4,5]. However, *none of those techniques guarantee that the representation induced by RNNs is suitable for optimizing the return.*

Alternatively, one may maintain a *belief*, i.e., a probability distribution over the states of the environment in which the agent believes it is located. Beliefs are a sufficient statistic for control [10], so optimizing belief-conditioned policies

* Both authors contributed equally, alphabetic order; † Equal supervision.

$a_t \sim \pi(\cdot \mid \text{belief} = b_t)$ is sufficient. However, updating this belief (passing from b_t to b_{t+1}) is typically intractable as it requires (i) an environment model, (ii) performing computation over the full (complex, unobservable) state space.

Our contribution. To tackle those challenges, we work around the assumption that the agent has access to the true environment state *but only during its training* (as it is standard in multi-agent RL with the "centralized training, decentralized execution" framework [15,2]). This enables learning a *latent model* of the environment. We introduce *Wasserstein Belief Updater* (WBU), a model that learns to replicate the theoretical, intractable belief update rule based on the learned dynamics through feedforward neural networks. **The learned model is guaranteed to be an accurate abstraction of the original environment while WBU is guaranteed to induce a suitable representation to optimize the return.**

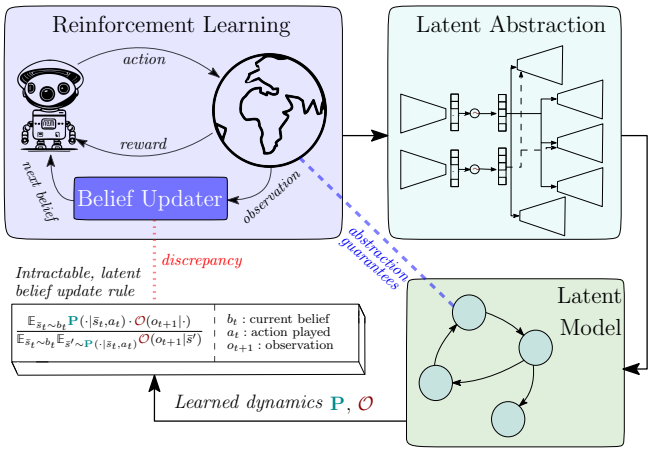

Fig. 2. Wasserstein Belief Updater (WBU) framework

Results. In contrast to RNN-based approaches and those relying on particle filtering [9], *our approach is principled and comes with abstraction and representation guarantees*, which explains its competitiveness in a variety of environments exhibiting different kinds of partial observability (Fig. 3).

This work has been published in the proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024) [1]

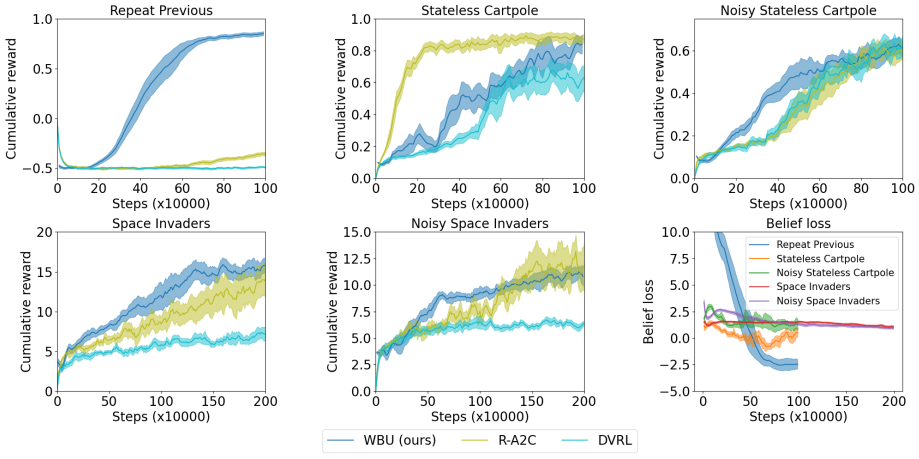

Fig. 3. Evolution of the (i) return for WBU, R-A2C (RNN + A2C [14]) and DVRL [9], and (ii) estimated belief loss during learning for WBU (mean and standard error).

Acknowledgements

This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program and was supported by the DESCARTES iBOF project. R. Avalos is supported by the Research Foundation – Flanders (FWO), under grant number 11F5721N. G.A. Perez is also supported by the Belgian FWO "SAILor" project (G030020N). We thank Mathieu Reymond, Denis Steckelmacher, and Mustafa Mert Çelikok for their valuable feedback.

References

1. Avalos, R., Delgrange, F., Nowe, A., Perez, G., Roijers, D.M.: The wasserstein believer: Learning belief updates for partially observable environments through reliable latent space models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=KrtGfTGaGe
2. Avalos, R., Reymond, M., Nowé, A., Roijers, D.M.: Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning. In: AAMAS '22: Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (Extended Abstract) (2022)
3. Chen, X., Mu, Y.M., Luo, P., Li, S., Chen, J.: Flow-based recurrent belief state learning for pomdps. In: International Conference on Machine Learning. pp. 3444–3468. PMLR (2022)
4. Gregor, K., Papamakarios, G., Besse, F., Buesing, L., Weber, T.: Temporal Difference Variational Auto-Encoder. 7th International Conference on Learning Representations, ICLR 2019 (6 2018). https://doi.org/10.48550/arxiv.1806.03107, https://arxiv.org/abs/1806.03107v3
5. Gregor, K., Rezende, D.J., Besse, F., Wu, Y., Merzic, H., van den Oord, A.: Shaping Belief States with Generative Environment Models for RL. Advances in Neural Information Processing Systems **32** (6 2019). https://doi.org/10.48550/arxiv.1906.09237, https://arxiv.org/abs/1906.09237v2
6. Hafner, D., Deepmind, T.L., Ba, J., Norouzi, M., Brain, G.: Dream to Control: Learning Behaviors by Latent Imagination (12 2019). https://doi.org/10.48550/arxiv.1912.01603, https://arxiv.org/abs/1912.01603v3
7. Hafner, D., Lillicrap, T.P., Norouzi, M., Ba, J.: Mastering atari with discrete world models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=0oabwyZbOu
8. Hausknecht, M., Stone, P.: Deep recurrent q-learning for partially observable MDPs. In: AAAI Fall Symposium - Technical Report. vol. FS-15-06, pp. 29–37. AI Access Foundation (2015)
9. Igl, M., Zintgraf, L., Le, T.A., Wood, F., Whiteson, S.: Deep variational reinforcement learning for POMDPs. In: 35th International Conference on Machine Learning, ICML 2018. vol. 5 (2018)
10. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial intelligence **101**(1-2), 99–134 (1998)
11. Lauri, M., Hsu, D., Pajarinen, J.: Partially observable markov decision processes in robotics: A survey. IEEE Transactions on Robotics **39**(1), 21–40 (2023). https://doi.org/10.1109/TRO.2022.3200138

12. Liu, W., Kim, S.W., Pendleton, S., Ang, M.H.: Situation-aware decision making for autonomous driving on urban road using online pomdp. In: 2015 IEEE Intelligent Vehicles Symposium (IV). pp. 1126–1133. IEEE (2015)
13. Ma, X., Karkus, P., Hsu, D., Lee, W.S.: Particle filter recurrent neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5101–5108 (2020)
14. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T.P., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous Methods for Deep Reinforcement Learning. 33rd International Conference on Machine Learning, ICML 2016 **4**, 2850–2869 (2 2016), http://arxiv.org/abs/1602.01783
15. Oliehoek, F.A., Spaan, M.T., Vlassis, N.: Optimal and approximate Q-value functions for decentralized POMDPs. Journal of Artificial Intelligence Research **32**, 289–353 (10 2008). https://doi.org/10.1613/jair.2447, http://dx.doi.org/10.1613/jair.2447
16. Szepesvári, C.: Algorithms for Reinforcement Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers (2010). https://doi.org/10.2200/S00268ED1V01Y201005AIM009, https://doi.org/10.2200/S00268ED1V01Y201005AIM009
17. Wu, Y., Macdonald, C., Ounis, I.: Partially observable reinforcement learning for dialog-based interactive recommendation. In: Proceedings of the 15th ACM Conference on Recommender Systems. pp. 241–251 (2021)