Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes

Delgrange, Florent; Nowe, Ann; Pérez, Guillermo A.

Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes

Extended Abstract

Florent Delgrange[1,2], Ann Nowé[1], and Guillermo A. Pérez[2]

[1] AI Lab, Vrije Universiteit Brussel
[2] University of Antwerp – Flanders Make

While *reinforcement learning* (RL) has been applied to a wide range of challenging domains, from game playing [9] to real-world applications such as effective canal control [11], more widespread deployment in the real world is hampered by the lack of guarantees provided with the learned policies. Although there are RL algorithms which have limit-convergence guarantees in the discrete setting [12] (and even in some continuous settings with function approximation, e.g., [10]), these are lost when applying more advanced techniques which make use of general nonlinear function approximators [13] to deal with continuous *Markov decision processes* (MDPs) such as *deep*-RL (e.g., [9]). In this work, we apply such advanced RL algorithms to *unknown continuous* MDPs with (safety constrained) reachability or discounted-reward objectives, and we consider the challenge of simplifying and verifying RL policies. Our goal is to *enable model checking* [2] by learning an accurate, tractable model of the environment.

Bisimulation Guarantees. To recover the formal guarantees, we thus seek a verifiable *discrete latent model* that approximates the unknown environment. Given the original (continuous, possibly unknown) environment model $\mathcal{M}$, a *latent space model* is another (smaller, explicit) MDP $\overline{\mathcal{M}}$ with state-action space linked to the original one via state and action *embedding functions* ϕ and ψ. Intuitively, an agent can execute a latent policy $\bar{\pi}$ (i.e., a policy defined over the latent spaces) in $\mathcal{M}$ as follows: at each step of the interaction, the current state s of $\mathcal{M}$ is embedded to the latent space via $\phi(s) = \bar{s}$, then the agent executes the latent action $\bar{a}$ prescribed by the policy $\bar{\pi}$ by embedding it back to the original model via ψ. Then, $\mathcal{M}$ transitions to the next state s' according to its transition function $\mathbf{P}$, the original state s, and this resulting action. The guarantees rely on (i) the *bisimulation pseudometric* $\widetilde{d}_{\bar{\pi}}$ [5, 6], and (ii) two *local losses* $L_{\mathbf{P}}^{\xi_{\bar{\pi}}}$ and $L_{\mathcal{R}}^{\xi_{\bar{\pi}}}$ [7]. The former can be interpreted as the *largest behavioral difference* between $\mathcal{M}$ and $\overline{\mathcal{M}}$ when $\bar{\pi}$ is executed. In particular, a zero distance means that the agent behaves the same way in both models. The latter intuitively quantify respectively the expected distance between the origi-

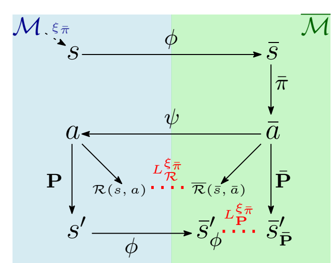

Fig. 1: Execution of $\bar{\pi}$.

nal and latent reward functions, $\mathcal{R}$ and $\overline{\mathcal{R}}$, as well as their transition functions, $\mathbf{P}$ and $\overline{\mathbf{P}}$. We show that these two losses bound $\widetilde{d}_{\overline{\pi}}$:

$$\mathop{\mathbb{E}}_{s\sim\xi_{\overline{\pi}}} \widetilde{d}_{\overline{\pi}}(s,\phi(s)) \leq \frac{L_{\mathcal{R}}^{\xi_{\overline{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\overline{\pi}}}}{1-\gamma};\ \widetilde{d}_{\overline{\pi}}(s_1,s_2) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\overline{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\overline{\pi}}}}{1-\gamma}\right)\left(\xi_{\overline{\pi}}^{-1}(s_1) + \xi_{\overline{\pi}}^{-1}(s_2)\right)$$

where $\xi_{\overline{\pi}}$ is a suitable distribution over states-actions likely to be seen under $\overline{\pi}$, γ is a discount, and s_1, s_2 have the same embedding $\phi(s_1) = \phi(s_2)$. These inequalities guarantee the *quality of the abstraction* and *representation*: when local losses are small, (i) in average, states and their embedding, and (ii) all states sharing the same discrete representation, are bisimilarly close. We give PAC approximation schemes to compute both the losses and said bounds. Next, we learn a distillation $\overline{\pi}$ of the RL policy along with $\overline{\mathcal{M}}$, where the behaviors of the agent can be formally verified. The bounds offer a confidence metric allowing to lift the guarantees obtained this way back to $\mathcal{M}$, when it operates under $\overline{\pi}$.

Variational MDP. We learn $\overline{\mathcal{M}}$ via a *variational autoencoder* (VAE) by maximizing a lower bound on the likelihood of traces generated by executing the original RL policy in $\mathcal{M}$. We derive a loss function incorporating variational versions of the local losses that enables learning (i) a discrete latent model, (ii) state-action embedding functions, and (iii) a distillation $\overline{\pi}$ of the RL policy. Our algorithm allows training this VAE in an efficient way and avoiding the so-called *mode collapse* problem, often occurring in variational models [1].

Experiments. We trained deep-RL policies [9, 8] for various benchmarks [3], which we then distill via our approach. The results reveal that optimizing the VAE-MDP (Fig. 2a) allows minimizing the local losses (Fig 2b). Furthermore, this enables the distillation of RL policies into $\overline{\pi}$, for which the formal guarantees apply: its performances in the original model are eventually recovered (Fig 2c).

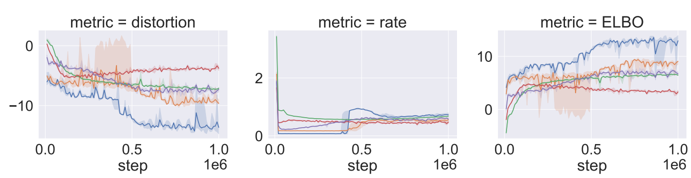

Fig. 2a. Variational metrics (VAE-MDP optimization)

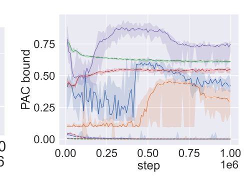

Fig. 2b. Local losses

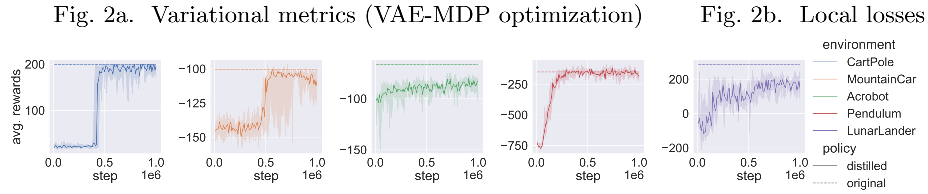

Fig. 2c. Distilled policy evaluation

This work has been published in the proceedings of the 36th AAAI Conference on Artificial Intelligence [4]. Ongoing work includes the extension of the approach to *Wasserstein autoencoders*, to provide additional *learning guarantees*.

Acknowledgments

This research received funding from the Flemish Government (AI Research Program) and was supported by the DESCARTES iBOF project. G.A. Perez is also supported by the Belgian FWO "SAILor" project (G030020N).

References

1. Alemi, A.A., Poole, B., Fischer, I., Dillon, J.V., Saurous, R.A., Murphy, K.: Fixing a broken ELBO. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 159–168. PMLR (2018)
2. Baier, C., Katoen, J.: Principles of model checking. MIT Press (2008)
3. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. CoRR **abs/1606.01540** (2016)
4. Delgrange, F., Nowé, A., Pérez, G.A.: Distillation of rl policies with formal guarantees via variational abstraction of markov decision processes. Proceedings of the AAAI Conference on Artificial Intelligence **36**(6), 6497–6505 (Jun 2022)
5. Desharnais, J., Gupta, V., Jagadeesan, R., Panangaden, P.: Metrics for labelled markov processes. Theor. Comput. Sci. **318**(3), 323–354 (2004)
6. Ferns, N., Panangaden, P., Precup, D.: Metrics for markov decision processes with infinite state spaces. In: UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005. pp. 201–208. AUAI Press (2005)
7. Gelada, C., Kumar, S., Buckman, J., Nachum, O., Bellemare, M.G.: Deepmdp: Learning continuous latent space models for representation learning. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 2170–2179. PMLR (2019)
8. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 1856–1865. PMLR (2018)
9. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nat. **518**(7540), 529–533 (2015)
10. Nowe, A.: Synthesis of "safe" fuzzy controllers based on reinforcement learning. Ph.D. thesis, Vrije Universiteit Brussel (1994)
11. Ren, T., Niu, J., Cui, J., Ouyang, Z., Liu, X.: An application of multi-objective reinforcement learning for efficient model-free control of canals deployed with iot networks. Journal of Network and Computer Applications **182**, 103049 (2021)
12. Tsitsiklis, J.N.: Asynchronous stochastic approximation and q-learning. Mach. Learn. **16**(3), 185–202 (1994)
13. Tsitsiklis, J.N., Roy, B.V.: An analysis of temporal-difference learning with function approximation. IEEE Trans. Autom. Control. **42**(5), 674–690 (1997)