Final Project Proposal

K-learning Alteration of ‘RL-as-inference’ Framework: **casestudy**: Average-reward RL & Non-equilibrium Statistical Mechanics

Klimentina Krstevska & Ilija Nikolov & Jacob Makar-Limanov
CSCI 2951-F: Learning and Sequential Decision Making

March 19, 2024

1 Introduction: the RL-as-inference Controversy

The framework of reinforcement learning (RF) or optimal control provides a mathematical formalization of intelligent decision making that is powerful and broadly applicable. While the general form of the RF problem enables effective reasoning about uncertainty, the connection between RF and inference in probabilistic models is not immediately obvious. Such a connection has been made, a notable summary is given by Levine [3], and it has been shown to have considerable value when it comes to algorithm design: formalizing a problem as probabilistic inference in principle allow bringing to bear a wide array of approximate inference tools, extending the model in flexible and powerful ways, and reasoning about compositionality and partial observability. In fact, a generalization of the RF or optimal control problem, which is sometimes termed maximum entropy reinforcement learning, is equivalent to exact probabilistic inference in the case of deterministic dynamics, and variational inference in the case of stochastic dynamics. This research has been dubbed ‘RL-as-inference.’

On the other hand, a recent paper by O’Donoghue *et al.* [9] has surfaced a key shortcoming in that approach, and suggested the limits in which RL can be coherently cast as an inference problem. In particular, they state that an RL agentmust consider the effects of its actions upon future rewards and observations: The exploration-exploitation tradeoff. In all but the most simple settings, the resulting inference is computationally intractable so that practical RL algorithms must resort to approximation [7, 9, 11]. They have demonstrated that the popular ‘RL as inference’ approximation can perform poorly in even very basic problems [9]. However, they suggested that with a small modification the framework does yield algorithms that can provably perform well, and showed that the resulting algorithm is equivalent to the recently proposed K-learning, which we further connect with Thompson sampling [9].

1.1 Casestudy: Non-equilibrium-statistical-mechanics-inspired RL

In reinforcement learning (RL), an agent sequentially interacts with an environment in order to discover optimal behaviors. Most of the established frameworks for RL feature discounting, wherein future returns are discounted relative to current returns. Although this framework has the benefit that it leads to useful bounds and convergence properties, it is not appropriate for problems in which discounting is not a principled approach. To address this issue, Arriojas *et al.* [1] used an alternative framework based on the average-reward formalism, which optimizes for long-term average returns. Notably, when coupled with entropy regularization, the average-reward formulation establishes a connection with the well-studied problem of free energy minimization in non-equilibrium statistical mechanics (NESM). Utilizing this connection, they show how concepts and tools based on large deviation theory in NESM can be leveraged to develop novel algorithms for solving problems with stochastic dynamics in RL.

1.2 Motivation and Connection to Non-equilibrium Statistical Mechanics

Reinforcement learning (RL) is an important field of research in machine learning that is increasingly being applied to complex optimization problems [5]. In parallel, concepts from physics have contributed to important advances in RL with developments such as entropy-regularized RL [6, 5, 4, 8, 2, 10]. While these developments have led to advances in both fields, obtaining analytical solutions for optimization in entropy-regularized RL is an open problem [1]. Arriojas *et al.* [1] established a mapping between entropy-regularized RL and research in nonequilibrium statistical mechanics focusing on Markovian processes conditioned on rare events. In the long-time limit, they applied approaches from large deviation theory to derive exact analytical results for the optimal policy and optimal dynamics in Markov decision process (MDP) models of reinforcement learning. The results obtained had lead to an analytical and computational framework for entropy-regularized RL which is validated by simulations. The mapping connects current research in reinforcement learning and non-equilibrium statistical mechanics, thereby opening avenues for the application of analytical and computational approaches from one field to cutting-edge problems in the other.

1.3 Research question

Firstly, we start by exploring Arriojas *et al.*'s work [1] on entropy regularised RF that uses large deviation theory. There the advantages of using MDP-based, entropy regularized reinforcement learning and Markovian processes conditioned on rare event in the long-time limit are lead out. With this approach, we can make connections to large deviation theory and derive exact analogical expression that characterize the trajectory distributions conditioned on optimality. These results derived for the optimal dynamics would allow us to produce analytical expressions for optimal value functions.

Secondly, we verify these derived results using the grid-world mazes they suggest, which have complete dynamics model available, *i.e.*, all available states, actions and transition dynamics are known beforehand. Additionally, we will make our own tests that are based on complete-dynamics models.

Thirdly, given the aforementioned problems with the 'RL-as-inference' framework, clearly laid out in [9], we begin to explore models where the complete dynamics is not known. To be fair, the use cases of the entropy regularized RF by Arriojas *et al.* [1] is to be connected to physics problems, where this is usually not the case. Our goal is to put the model to the test to see how well it performs in this situation. We have not found that the authors did this, as they were not interested in this particular point.

Fourthly, as proposed by O'Donoghue *et al.* [9], we will use the K-Learning alteration of the 'RL-as-inference' framework and attempt to derive an analytical solution similar to the one in our main paper [1], which has not been done previously. The main difference between the two approaches is that the entropy regularized RL only approximates the Bayes-optimal policy due to the exponential lookahead, and it does do efficient exploration. To mitigate this fundamental shortcoming, O'Donoghue *et al.* [9] suggest using a heuristics for balancing exploration with exploitation, namely, the Thompson sampling, by using a K-learning algorithm. This algorithm includes an epistemic uncertainty explicitly and drives exploration better.

Fifthly, we use the examples proposed in both papers to test both algorithm and compare their efficiency. This has not been done previously and it is useful to see if the main arguments of O'Donoghue *et al.* [9] hold.

Sixty and lastly, we will combine the results of the previous parts to see whether the claims of O'Donoghue *et al.* [9] about the useful/less/ness of 'RL-as-inference' are indeed true. In fact, they claimed that problem formulation ignores the role of epistemic uncertainty, that algorithms derived from the 'RL-as-inference' can perform poorly on even simple tasks. We would like to evaluate the extension to which this statement is true and provide some clarity on why there is a lot of extensive research in 'RL-as-inference' methodologies.

To summarize, we have the following milestones

1. Follow through the derivation of Arriojas *et al.* [1] about mapping between entropy-regularized RL and research in non-equilibrium statistical mechanics focusing on Markovian processes conditioned on rare events.
2. Use Arriojas *et al.*'s tests and write our own about their algorithm [1].
3. Follow through the arguments against 'RL-as-inference' by O'Donoghue *et al.* [9]
4. Rederive Arriojas *et al.*'s results [1] using the K-learning alteration suggested by O'Donoghue *et al.* [9]
5. Tests both algorithms using the tests provided by both Arriojas *et al.* [1] & O'Donoghue *et al.* [9]
6. Summarize results and evaluate useful/less/ness of the 'RL-as-inference' framework.

References

[1] Argenis Arriojas, Jacob Adamczyk, Stas Tiomkin, and Rahul V. Kulkarni. Entropy regularized reinforcement learning using large deviation theory. *Physical Review Research*, 5(2):023085, May 2023. doi:10.1103/PhysRevResearch.5.023085.

[2] Ariel Barr, Willem Gispen, and Austen Lamacraft. Quantum Ground States from Reinforcement Learning. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, pages 635–653. PMLR, August 2020. doi:10.48550/arXiv.2006.09044.

[3] Sergey Levine. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. May 2018. doi:10.48550/arXiv.1805.00909.

[4] Henry W. Lin, Max Tegmark, and David Rolnick. Why Does Deep and Cheap Learning Work So Well? *Journal of Statistical Physics*, 168(6):1223–1247, September 2017. doi:10.1007/s10955-017-1836-5.

[5] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, 810:1–124, May 2019. doi:10.1016/j.physrep.2019.03.001.

[6] Pankaj Mehta and David J. Schwab. An exact mapping between the Variational Renormalization Group and Deep Learning, October 2014. doi:10.48550/arXiv.1410.3831.

[7] Brendan O'Donoghue. Variational Bayesian Reinforcement Learning with Regret Bounds, December 2022. doi:10.48550/arXiv.1807.09647.

[8] Mateusz Ostaszewski, Lea M. Trenkwalder, Wojciech Masarczyk, Eleanor Scerri, and Vedran Dunjko. Reinforcement learning for optimization of variational quantum circuit architectures. In *Advances in Neural Information Processing Systems*, volume 34, pages 18182–18194. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/9724412729185d53a2e3e7f889d9f057-Paper.pdf.

[9] Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=S1xitgHtvS.

[10] Dominic C Rose, Jamie F Mair, and Juan P Garrahan. A reinforcement learning approach to rare trajectory sampling. *New Journal of Physics*, 23(1):013013, January 2021. doi:10.1088/1367-2630/abd7bd.

[11] Jean Tarbouriech, Tor Lattimore, and Brendan O'Donoghue. Probabilistic Inference in Reinforcement Learning Done Right. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. URL: https://openreview.net/forum?id=_0ggHCj8cPx.