Capstone Plan

[1] and Aguirre Max[1]

[1] *PSTB, Paris, France*

March 3, 2025

Abstract

We propose a self-contained, detailed, description of a scalable standardized kernel (RKHS) approach to popular reinforcement learning algorithms, where agents interact off-line with environments having continuous states and discrete actions spaces, dealing with possibly unstructured datas. These algorithms, namely Q-Learning, Actor Critic, Policy Gradient, Hamilton-Jacobi-Bellman (HJB) and Heuristic Controls, are implemented with a RKHS library [10] using default settings. We show that this approach to reinforcement learning is accurate, robust, efficient and versatile, as we benchmark our algorithms in this paper on simple games, and use them as a baseline for our applications.

1 Introduction

We look at the application of Kernel (RKHS) methods to Reinforcement Learning, with potential application in numerous fields, for instance, finance.

2 Background

In here we do a litterature review and give the main background ideas for Reinforcement learning, Kernel methods, main algorithms and their limitations

2.1 Reinforcement Learning

2.1.1 General Framework for RL

2.1.2 Bellman Equations

2.2 Algorithms

2.2.1 Q-learning

2.2.2 Policy Gradient

2.2.3 Hamilton Jacobi Bellman

2.2.4 Heuristic-Controlled Learning

2.3 Kernel Reminder

3 Kernel RL Algorithms

Here we describe the algorithms from a numerical point of view.

3.1 Kernel RL framework

3.2 Kernel Q-Learning

3.3 Kernel-Based Q-Value Gradient Estimation

3.4 Kernel Actor-Critic with Bellman Residual Advantage

3.5 Kernelized Hamilton Jacobi Bellman

3.6 Heuristic-controlled Learning

4 Experiments

4.1 Cartpole

4.2 Lunar-Lander

5 Conclusion

References

[1] Sayak Ray Chowdhury and Aditya Gopalan, *On Kernelized Multi-armed Bandits.* Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research - 70 , pp 844-853

[2] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani, and K.Q. Weinberger, editors, Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pp. 2292–2300.

[3] Mania, H., Guy, A., and Recht, B., Simple random search of static linear policies is competitive for reinforcement learning. In Advances in Neural Information Processing Systems, 1800–1809.

[4] Cheng, Ching-An and Kolobov, Andrey and Swaminathan, Adith, *Heuristic-guided reinforcement learning.* Proceedings of the 35th International Conference on Neural Information Processing Systems - 1038, pp 13550 - 13563

[5] Mnih et al. (2015)"Human-level control through deep reinforcement learning"

[6] Van Hasselt, Guez, & Silver (2016) "Deep Reinforcement Learning with Double Q-learning"

[7] Williams, Ronald J., *Simple statistical gradient-following algorithms for connectionist reinforcement learning* Machine Learning, 8(3-4): pp 229-256

[8] Sing-Yuan Yeh, Fu-Chieh Chang, Chang-Wei Yueh, Pei-Yuan Wu, Alberto Bernacchia, Sattar VakiliSample Complexity of Kernel-Based Q-Learning

[9] P.G. LeFloch and J.-M. Mercier, Extrapolation and generative algorithms for three applications in finance, Wilmott, vol. 2024, iss. 133, 2024.

[10] P.G. LeFloch, J.-M. Mercier, and S. Miryusupov, CodPy: a Python library for numerics, machine learning, and statistics. arXiv:2402.07084

[11] P.G. LeFloch, J.-M. Mercier, and S. Miryusupov, A class of kernel-based scalable algorithms for data science. arXiv:2410.14323

[12] P.G. LeFloch, J.-M. Mercier, A new method for solving Kolmogorov equations in mathematical finance, DOI : 10.1016/j.crma.2017.05.003

[13] Watkins, C. J. C. H., & Dayan, P. , *Q-learning.* Machine Learning, 8(3-4), pp 279–292.

[14] Watkins, C. J. C. H., & Dayan, P. , Why Should I Trust You, Bellman? The Bellman Error is a Poor Replacement for Value Error, arXiv:2201.12417 [cs.LG] , https://doi.org/10.48550/arXiv.2201.12417

[15] Volodymyr Mnih, Koray Kavukcuoglu,David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra and Martin A. Riedmiller , *Playing Atari with Deep Reinforcement Learning.*

[16] Sutton, R. S., & Barto, A. G., *Reinforcement learning: An introduction (2nd ed.)* MIT Press.

[17] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics,* Springer Science, Business Media, LLC, 2004.

[18] Silver et al. (2014) "Deterministic Policy Gradient Algorithms"

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford and Oleg Klimov, *Proximal Policy Optimization Algorithms* arXiv:1707.06347, https://doi.org/10.48550/arXiv.1707.06347

[20] Ormoneit, Sen Kernel-Based Value Function Approximation (2002)

[21] Engel et al.Gaussian Process Temporal Difference Learning (GPTD) (2003)

[22] X. Xu, D. Hu and X. Lu, *Kernel-Based Least Squares Policy Iteration for Reinforcement Learning,* in IEEE Transactions on Neural Networks, vol. 18, no. 4, pp. 973-992, July 2007, doi: 10.1109/TNN.2007.899161.

6 APPENDIX