Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes

Alekh Agarwal ALEKHA@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052.

Sham M. Kakade SHAM@CS.WASHINGTON.EDU
University of Washington, Seattle, WA 98195.

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University, Princeton, NJ 08540.

Gaurav Mahajan GMAHAJAN@ENG.UCSD.EDU
University of California San Diego, La Jolla, CA 92093.

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Policy gradient (PG) methods are among the most effective methods in challenging reinforcement learning problems with large state and/or action spaces. However, little is known about even their most basic theoretical convergence properties, including: if and how fast they converge to a globally optimal solution (say with a sufficiently rich policy class); how they cope with approximation error due to using a restricted class of parametric policies; or their finite sample behavior. Such characterizations are important not only to compare these methods to their approximate value function counterparts (where such issues are relatively well understood, at least in the worst case), but also to help with more principled approaches to algorithm design.

This work provides provable and comprehensive characterizations of computational, approximation, and sample size issues with regards to policy gradient methods in the context of discounted Markov Decision Processes (MDPs). We focus on both: 1) "tabular" policy parameterizations, where the optimal policy is contained in the class and where we show global convergence to the optimal policy, and 2) restricted policy classes, which may not contain the optimal policy and where we provide agnostic learning results. In the *tabular setting* with exact gradients, our main results for the softmax policy parameterization show: asymptotic convergence to the global optimum; a polynomial convergence rate provided an additional KL-based entropy regularizer is used; and a dimension-free, "fast-rate" convergence to the global optimum using the Natural Policy Gradient (NPG). With regards to *function approximation*, we further analyze NPG with inexact gradients, where the policy parameterization may not contain the optimal policy. Under certain smoothness assumptions on the policy parameterization, we establish rates of convergence in terms of the quality of the initial state distribution and the quality of a certain compatible function approximation. One insight of this work is in formalizing how a favorable initial state distribution provides a means to circumvent worst-case exploration issues. Overall, these results place PG methods under a solid theoretical footing, analogous to the global convergence guarantees of iterative value function based algorithms[1].

[1] Extended abstract. Full version appears as [arXiv:1908.00261, v2].

References

Understanding the impact of entropy on policy optimization, 2019. URL https://arxiv.org/abs/1811.11214.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR, 2019.

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Mohammad Gheshlaghi Azar, Vicenç Gómez, and Hilbert J. Kappen. Dynamic policy programming. *J. Mach. Learn. Res.*, 13(1), November 2012. ISSN 1532-4435.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement lear ning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *NIPS 26*, 2013.

J. A. Bagnell, Sham M Kakade, Jeff G. Schneider, and Andrew Y. Ng. Policy search by dynamic programming. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 831–838. MIT Press, 2004.

J. Andrew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1019–1024, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997.

Richard Bellman and Stuart Dreyfus. Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13(68):247–251, 1959.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *CoRR*, abs/1906.01786, 2019. URL http://arxiv.org/abs/1906.01786.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Justin A Boyan. Least-squares temporal difference learning. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 49–56. Morgan Kaufmann Publishers Inc., 1999.

Qi Cai, Zhuoran Yang, Jason D. Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *CoRR*, abs/1905.10027, 2019. URL `http://arxiv.org/abs/1905.10027`.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051, 2019.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*, pages 1422–1432, 2018.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomput.*, 71(7-9):1180–1190, 2008. ISSN 0925-2312.