LP-based Algorithms for Reinforcement Learning

Peter Bartlett, Ofir Nachum, Aldo Pacchiano
peter@berkeley.edu,ofirnachum@google.com, pacchiano@berkeley.edu

Overview and Related Work The study of reinforcement learning (RL) has been traditionally dominated by dynamic programming (DP) approaches (2). An alternative approach exists known as the $\mathrm{V-LP}$, based on linear programming (LP) (6, 10), and it has recently received renewed interest for its potential ability to circumvent the optimization challenges of DP-based approaches in exchange for more mature and well-suited techniques associated with convex optimization (1, 4, 5, 15, 16). While these LP-based algorithms provide theoretical guarantees, they have thus far not shown good practical performance competitive with DP-based algorithms.

In contrast, a variant of the LP approach known as the $\mathrm{Q-LP}$ (11, 12, 13) has demonstrated impressive practical performance, but has not yet been shown to enjoy the same theoretical guarantees as the $\mathrm{V-LP}$. A fundamental feature of the $\mathrm{Q-LP}$ approach is its use of a *regularized* form of the standard LP objective. This seems to hint at a connection to state-of-the-art solutions in deep RL, which often rely on the use of *entropy regularization* for better training stability (8, 9). Moreover, the use of entropy regularization has recently been identified as a key element for providing convergence guarantees of $\mathrm{DP}-$based policy optimization algorithms (3, 7, 14) although the analyzed settings are typically far removed from the deep function approximators used in practice. Whether LP-based approaches can provide a better combination of good practical performance and guarantees in realistic settings is an open question of great interest.

Contributions The preprint resulting from our work can be accessed at `https://arxiv.org/abs/2103.09756`. In our work we provide the first finite time convergence rates for the REPS algorithm in the literature. We derive Accelerated Gradient Descent rates for REPS in the setting where the model is known to the learner. Similarly and under an explorability assumption we are able to derive convergence rates for optimization of the REPS objective based on Stochastic Gradient descent Steps.

References

[1] J. Bas-Serrano and G. Neu. Faster saddle-point optimization for solving large-scale markov decision processes. *arXiv preprint arXiv:1909.10904*, 2019.

[2] R. Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[3] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

[4] Y. Chen and M. Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

[5] D. P. De Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003.

[6] E. V. Denardo. On linear programming in a markov decision problem. *Management Science*, 16(5):281–288, 1970.

[7] Y. Efroni, L. Shani, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

[8] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.

[9] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[10] A. S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[11] O. Nachum and B. Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[12] O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019.

[13] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. Algaedice: Policy gradient from arbitrary experience, 2019.

[14] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist. Leverage the average: an analysis of regularization in rl. *arXiv preprint arXiv:2003.14089*, 2020.

[15] M. Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

[16] T. Wang, M. Bowling, D. Lizotte, and D. Schuurmans. Dual representations for dynamic programming. 2007.