Proceedings of Machine Learning Research vol 134:1–4, 2021 34th Annual Conference on Learning Theory

Softmax Policy Gradient Methods Can Take Exponential Time to Converge

Gen Li G-LI16@MAILS.TSINGHUA.EDU.CN
Tsinghua University, Beijing 100084, China.

Yuting Wei YTWEI@CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Yuejie Chi YUEJIECHI@CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Yuantao Gu GYT@TSINGHUA.EDU.CN
Tsinghua University, Beijing 100084, China.

Yuxin Chen YUXIN.CHEN@PRINCETON.EDU
Princeton University, Princeton, NJ 08544, USA.

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

The softmax policy gradient (PG) method, which performs gradient ascent under softmax policy parameterization, is arguably one of the de facto implementations of policy optimization in modern reinforcement learning. For γ-discounted infinite-horizon tabular Markov decision processes (MDPs), remarkable progress has recently been achieved towards establishing global convergence of softmax PG methods in finding a near-optimal policy. However, prior results fall short of delineating clear dependencies of convergence rates on salient parameters such as the cardinality of the state space $\mathcal{S}$ and the effective horizon $\frac{1}{1-\gamma}$, both of which could be excessively large. In this paper, we deliver a pessimistic message regarding the iteration complexity of softmax PG methods, despite assuming access to exact gradient computation. Specifically, we demonstrate that the softmax PG method with stepsize η can take

$$\frac{1}{\eta}|\mathcal{S}|^{2^{\Omega\left(\frac{1}{1-\gamma}\right)}} \text{ iterations}$$

to converge, even in the presence of a benign policy initialization and an initial state distribution amenable to exploration (so that the distribution mismatch coefficient is not exceedingly large). This is accomplished by characterizing the algorithmic dynamics over a carefully-constructed MDP containing only three actions. Our exponential lower bound hints at the necessity of carefully adjusting update rules or enforcing proper regularization in accelerating PG methods. [1]

Acknowledgments

G. Li and Y. Gu are supported by NSFC-61971266. Y. Wei, Y. Chi and Y. Chen have been supported in part by NSF CCF-2106778, DMS-2015447, CCF-1907661, CCF-2106739, CCF-2007911 and IIS-1900140, ONR N00014-18-1-2142, N00014-19-1-2404 and N00014-19-1-2120, AFOSR FA9550-19-1-0030, ARO W911NF-20-1-0097 and W911NF-18-1-0303.

1. Extended abstract. Full version appears as [arXiv:2102.11270, v2].

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*, 2019.

Andrea Agazzi and Jianfeng Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *International Conference on Learning Representations (ICLR)*, 2021.

Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pages 243–252. PMLR, 2017.

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

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Jalaj Bhandari. *Optimization Foundations of Reinforcement Learning*. PhD thesis, Columbia University, 2020.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Jalaj Bhandari and Daniel Russo. A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv:2007.06558, accepted to Operations Research*, 2020.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *arXiv preprint arXiv:2105.15186*, 2021.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

SS Du, C Jin, MI Jordan, B Póczos, A Singh, and JD Lee. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1068–1078, 2017.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Joao Paulo Jansch-Porto, Bin Hu, and Geir Dullerud. Convergence guarantees of policy optimization methods for Markovian jump linear systems. *arXiv preprint arXiv:2002.04090*, 2020.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*, 2021.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*, 2021.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33, 2020a.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*, 2019.

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

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*, 2019.

Richard S. Sutton. Temporal credit assignment in reinforcement learning. *PhD thesis, University of Massachusetts*, 1984.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Provable fictitious play for general mean-field games. *arXiv preprint arXiv:2010.04211*, 2020.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020b.

Wenhao Yang, Xiang Li, Guangzeng Xie, and Zhihua Zhang. Finding the near optimal policy via adaptive reduced regularization in mdps. *arXiv preprint arXiv:2011.00213*, 2020.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *arXiv preprint arXiv:2105.11066*, 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33, 2020a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. *arXiv preprint arXiv:2010.11364*, 2020b.

Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*, 2019.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020c.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*, 2021.