A Theoretical Analysis of Deep Q-Learning

Jianqing Fan JQFAN@PRINCETON.EDU
Princeton University

Zhaoran Wang ZHAORAN.WANG@NORTHWESTERN.EDU
Northwestern University

Yuchen Xie YCXIE@U.NORTHWESTERN.EDU
Northwestern University

Zhuoran Yang ZY6@PRINCETON.EDU
Princeton University

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

[1] Abstract

Despite the great empirical success of deep reinforcement learning, its theoretical foundation is less well understood. In this work, we make the first attempt to theoretically understand the deep Q-network (DQN) algorithm (Mnih et al., 2015) from both algorithmic and statistical perspectives. In specific, we focus on a slight simplification of DQN that fully captures its key features. Under mild assumptions, we establish the algorithmic and statistical rates of convergence for the action-value functions of the iterative policy sequence obtained by DQN. In particular, the statistical error characterizes the bias and variance that arise from approximating the action-value function using deep neural network, while the algorithmic error converges to zero at a geometric rate. As a byproduct, our analysis provides justifications for the techniques of experience replay and target network, which are crucial to the empirical success of DQN. Furthermore, as a simple extension of DQN, we propose the Minimax-DQN algorithm for zero-sum Markov game with two players. Borrowing the analysis of DQN, we also quantify the difference between the policies obtained by Minimax-DQN and the Nash equilibrium of the Markov game in terms of both the algorithmic and statistical rates of convergence.

Keywords: Deep Q-Learning, Markov Decision Process, Zero-Sum Markov Game

Introduction. In this work, we aim to provide theoretical guarantees for DQN (Mnih et al., 2015), which can be cast as an extension of the classical Q-learning algorithm (Watkins and Dayan, 1992) that uses deep neural network to approximate the action-value function. Although the algorithmic and statistical properties of the classical Q-learning algorithm are well-studied, theoretical analysis of DQN is highly challenging due to its differences in the following two aspects.

First, in online gradient-based temporal-difference reinforcement learning algorithms, approximating the action-value function often leads to instability. Baird (1995) proves that this is the case even with linear function approximation. The key technique to achieve stability in DQN is experience replay (Lin, 1992; Mnih et al., 2015). In specific, a replay memory is used to store the

[1] Extended abstract. Full version appears as arXiv reference, arXiv:1901.00137.

trajectory of the Markov decision process (MDP). At each iteration of DQN, a mini-batch of states, actions, rewards, and next states are sampled from the replay memory as observations to train the Q-network, which approximates the action-value function. The intuition behind experience replay is to achieve stability by breaking the temporal dependency among the observations used in training the deep neural network.

Second, in addition to the aforementioned Q-network, DQN uses another neural network named the target network to obtain an unbiased estimator of the mean-squared Bellman error used in training the Q-network. The target network is synchronized with the Q-network after each period of iterations, which leads to a coupling between the two networks. Moreover, even if we fix the target network and focus on updating the Q-network, the subproblem of training a neural network still remains less well-understood in theory.

In this paper, we focus on a slight simplification of DQN, which is amenable to theoretical analysis while fully capturing the above two aspects. In specific, we simplify the technique of experience replay with an independence assumption, and focus on deep neural networks with rectified linear units (ReLU) (Nair and Hinton, 2010) and large batch size. Under this setting, DQN is reduced to the neural fitted Q-iteration (FQI) algorithm (Riedmiller, 2005) and the technique of target network can be cast as the value iteration. More importantly, by adapting the approximation results for ReLU networks to the analysis of Bellman operator, we establish the algorithmic and statistical rates of convergence for the iterative policy sequence obtained by DQN. As we will show in the main results, the statistical error characterizes the bias and variance that arise from approximating the action-value function using neural network, while the algorithmic error geometrically decays to zero as the number of iteration goes to infinity.

Furthermore, we extend DQN to two-player zero-sum Markov games (Shapley, 1953). The proposed algorithm, named Minimax-DQN, can be viewed as a combination of the Minimax-Q learning algorithm for tabular zero-sum Markov games (Littman, 1994) and deep neural networks for function approximation. Compared with DQN, the main difference lies in the approaches to compute the target values. In DQN, the target is computed via maximization over the action space. In contrast, the target obtained computed by solving the Nash equilibrium of a zero-sum matrix game in Minimax-DQN, which can be efficiently attained via linear programming. Despite such a difference, both these two methods can be viewed as approximately applying the Bellman operator to the Q-network. Thus, borrowing the analysis of DQN, we also establish theoretical results for Minimax-DQN. Specifically, we quantify the suboptimality of policy returned by the algorithm by the difference between the action-value functions associated with this policy and with the Nash equilibrium policy of the Markov game. For this notion of suboptimality, we establish the both algorithmic and statistical rates of convergence, which implies that the action-value function converges to the optimal counterpart up to an unimprovable statistical error in geometric rate.

Our contribution is three-fold. First, we establish the algorithmic and statistical errors of the neural FQI algorithm, which can be viewed as a slight simplification of DQN. Under mild assumptions, our results show that the proposed algorithm obtains a sequence of Q-networks that geometrically converges to the optimal action-value function up to an intrinsic statistical error induced by the approximation bias of ReLU network and finite sample size. Second, as a byproduct, our analysis justifies the techniques of experience replay and target network used in DQN, where the latter can be viewed as a single step of the value iteration. Third, we propose the Minimax-DQN algorithm that extends DQN to two-player zero-sum Markov games. Borrowing the analysis for DQN, we establish the algorithmic and statistical convergence rates of the action-value functions associated with the sequence of policies returned by the Minimax-DQN algorithm.

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *arXiv preprint arXiv:1711.01731*, 2017.

Jens Kober and Jan Peters. Reinforcement learning in robotics: A survey. In *Reinforcement Learning*, pages 579–610. Springer, 2012.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4):293–321, 1992.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pages 157–163. Elsevier, 1994.

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.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International conference on machine learning*, pages 807–814, 2010.

Martin Riedmiller. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pages 317–328. Springer, 2005.

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.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

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.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.