Deep Policy Gradient Methods, RKHS and Convergence Guarantees of Neural Network Parameterized Policies

Author: Riashat Islam

8th May 2016

1 Introduction

We consider the task of improving convergence guarantees of policy gradient methods. Policy gradient methods are one of the fundamental ideas of reinforcement learning, in which the parameterized policy can be directly optimized using its own function approximator. Previous work considered the optimization of stochastic policy gradients Sutton et al. ((1999)), while more recent work showed the existence of deterministic policy gradient methods Silver et al. ((2014)). Furthermore, recent work from Mnih et al. ((2013)) considered the use of non-linear function approximators for the action-value Q function to play Atari games, making Atari a new benchmark task for reinforcement learning algorithms.

In this work, we propose the use of using non-linear policy parameterisation and improve the convergence guarantees of policy gradient methods under neural network parameterised policies. Our goal is to show that using a non-linear parameterisation, the gradient of the cumulative reward objective function $J(\theta)$ can be efficiently optimized through the use of backpropagation algorithm. Furthermore, recent work from Lever and Stafford ((2015)) showed that non-parametric policies can also be learn by performing gradient based policy optimisation in the reproducing kernel Hilbert space (RKHS). Recently, Furmston and Lever ((2015)) also considered the use of approximate Newton methods for policy optimisation, and used the Hessian of the objective function for finding the optimal policy. We propose to combine these approaches, to investigate the convergence guarantees, and proof the existence of deep policy gradient methods in reinforcement learning, using the Atari framework or the MuJoCo simulator for control tasks as our benchmark simulation platform.

2 Background

2.1 Problem Statement

Policy gradient methods are known to have better convergence guarantees since they are independent of the convergence of the value function Q. Such approaches are typically effective for reasonably higher dimensional state and action spaces, allowing the use of controlled exploration in large environments or MDPs. However, it is well known that policy gradient methods typically convergence to the local optima instead of the global optima of the objective function $J(\theta)$. Policy gradient methods, both stochastic and deterministic policy gradients rely on the use of gradient ascent for finding an optimal policy. While such approaches can reasonably converge under small state spaces using a linear parameterisation, the use of policy gradients in large state spaces such as the Atari framework is hypothesized to perform poorly. When considering non-linear parameterisation to model complex policies, it becomes even more difficult for policy gradients to work in a framework like the Atari platform.

2.2 Related Work

Recently, Silver et al. ((2014)) showed that policy gradient methods can work under continuous action spaces through the use of deterministic policy gradients. DPG considers finding the gradient of the action-value

function. While some recent work from Google DeepMind also showed the use of Deep Deterministic Policy Gradients (DDPG) in the Atari framework Lillicrap et al. ((2015)), it is still an open problem to explicitly consider deep policy gradient methods in such domains. While some experiments from Silver et al. ((2014)) used a neural network parameterisation of the policy, but no proof of convergence were considered in earlier work for such methods.

Another interesting direction of work is the use of RKHS function spaces for operating in a non-parametric class for policy optimisation Lever and Stafford ((2015)). By using vector valued RKHS, and using the policy gradient as an entire function in the RKHS, we can consider the representation and optimisation of complex policies. It would be interesting to investigate both the stochastic and deterministic policy gradient methods in the RKHS, while also using a non-linear policy parameterisation. Related work also proposed the use of approxamte Newton methods for optimisation, by finding the approximate Hessian of the policy objective function.

While not directly related to our problem, some recent methods based on guided policy search has been shown to perform well on high dimensional control problems, experimented on the MuJoCo simulator, Levine et al. ((2015)), Levine and Koltun ((2013)). Such methods based on guided policy search directly considered the use of neural network policy parameterisation Levine and Abbeel ((2014)). Further related work also considered the use of Bayesian optimisation, considering $J(\theta)$ as a black-box function, and using Guassian processes for direct optimisation of $J(\theta)$ from which the optimal policy can be derived, Wilson et al. ((2014)).

3 Approach

In this section, we briefly outline some of the approaches we will consider for improving convergence of deep policy gradient methods, using a policy parameterisation with a large neural network, and analysing the difference in convergence guarantees to the global optima.

- **Stochastic and Deterministic Gradients:** Our first task would be to analyse the convergence guarantees of both stochastic and deterministic policy gradient methods, under non-linear policy parameterisation. We will consider both on-policy and off-policy actor-critic approaches, analysing exploration using an ϵ-greedy approach under such settings. For our first task, we will consider policy gradients using the Atari as a simulation platform.
- **Natural Gradients of Stochastic and Deterministic Policies:** Natural gradients in policy gradient methods are known to have better convergence guarantees than vanilla policy gradients. Under non-linear parameterisation, we will further investigate the use of natural deterministic gradients that has recently been further proved by Silver et al. ((2014)).
- **Policies in the RKHS:** Building from work from Lever and Stafford ((2015)), we will explore the use of the RKHS function class for non-parametric policy optimisation. Under neural network parametrisation of policies, we will consider how the RKHS-based optimisation can have better convergence gurantees.
- **Approximate Hessian of Policy Gradients:** Second-order optimisation, with the use of momentum based and Nesterov's accelerated gradient based approches have been shown to achieve good optimisation guarantees in deep learning. We want to investigate the use of approximate Hessians in policy gradients. While such approaches maybe computationally expensive, especially under large state and action spaces, we want to demonstrate whether second-order optimisation can achieve better convergence in deep policy gradient methods.
- **Bayesian Optimisation in RL:** Rather than finding the gradient of the objective function in policy gradients, we will analyse whether the use of Gaussian Processes and Bayesian optimisation framework can be used for direct optimisation of $J(\theta)$. Our hypothesis is that, using the Bayesian Optimisation framework, if we can directly find the global optima, then policy gradient methods can be scaled up efficiently, irrespective of the type of policy parameterisation being used.

References

T. Furmston and G. Lever. A gauss-newton method for markov decision processes. *CoRR*, abs/1507.08271, 2015. URL http://arxiv.org/abs/1507.08271.

G. Lever and R. Stafford. Modelling policies in mdps in reproducing kernel hilbert space. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015. URL http://jmlr.org/proceedings/papers/v38/lever15.html.

S. Levine and P. Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1071–1079, 2014. URL http://papers.nips.cc/paper/5444-learning-neural-network-policies-with-guided-policy-search-under-unknown-dynamics.

S. Levine and V. Koltun. Guided policy search. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1–9, 2013. URL http://jmlr.org/proceedings/papers/v28/levine13.html.

S. Levine, N. Wagener, and P. Abbeel. Learning contact-rich manipulation skills with guided policy search. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 156–163, 2015. doi: 10.1109/ICRA.2015.7138994. URL http://dx.doi.org/10.1109/ICRA.2015.7138994.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015. URL http://arxiv.org/abs/1509.02971.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 387–395, 2014. URL http://jmlr.org/proceedings/papers/v32/silver14.html.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063, 1999. URL http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.

A. Wilson, A. Fern, and P. Tadepalli. Using trajectory data to improve bayesian optimization for reinforcement learning. *Journal of Machine Learning Research*, 15(1):253–282, 2014. URL http://dl.acm.org/citation.cfm?id=2627443.