Data Efficient Deep Reinforcement Learning using Approximate Inference

Submitted By: Riashat Islam; Supervisor(s): Zoubin Ghahramani, Shane Gu

18th November 2015

1 Introduction

Deep Reinforcement Learning, with non-linear policies parameterized by deep neural networks are still limited by the fact that learning and policy search methods requires larger number of interactions and training espisodes with the environment to find solutions. In our work, we consider data-efficient reinforcement learning in high dimensional state-action spaces for robot locomotion tasks and playing games (using pixel information) or on intelligent tutoring systems (language understanding) frameworks. The fundamental goal of our approach is to learn quickly using least number of system interactions and incorporate uncertainty into learning optimal policies. Below we will consider in detail some of the approaches towards a data-efficient reinforcement learning.

Our hypothesis is that using approximate inference and Bayesian methods for trajectory optimization, and then following a policy based on the optimized trajectory, we can learn optimal complex non-linear neural network policies much faster than current approaches using action-value (Q) functions. Bayesian approximate inference and sampling methods can provide data efficient policy search methods, while also taking account of exploration in an unknown environment, solving problems towards exploration and exploitation dilemma in unknown environments.

In this work, we will investigate using approximate inference for trajectory optimization and then use guided policy search based on optimized trajectory for optimal policy search. Our approximate inference methods can create locally linear approximations of the dynamics incorporating exploration and data efficiency. Finally, with the locally linear optimal policies along the trajectories, we can use supervised learning or Bayesian methods (to avoid overfitting) to train complex deep neural network policies to reproduce the state-to-action mapping found in the trajectory optimization phase. Our experimental framework will include playing Atari games and using the MuJoCo simulator for current state of the art benchmark tasks.

2 Prior Work

Our approach will be based on several prior methods. We consider some of the prior work based on which we can define our approximate inference based guided policy search for learning deep neural network policies.

2.1 Trajectory Optimization and Guided Policy Search

Levine and Koltun ((2013a)) considers a guided policy search algorithm that can direct policy learning. Policy gradient methods estimate the gradient of the expected return $\nabla_\theta J(\theta)$ using samples drawn from the current policy and then improve the policy by taking a step along the gradient. The guided policy search approach learns complex neural network policies with hundreds of parameters by incorporating guided samples, generated using importance sampling, into the policy search. Furthermore, Levine and Koltun ((2014)) formulates the problem as an optimization over trajectory distributions, alternating between optimizing the policy to match the trajectories and optimize the trajectory to match the policy and minimize the expected cost. Considering model-based reinforcement learning, when the dynamics model is available, trajectories

can be optimized directly with respect to the actions without a parametric policy.

Levine and Koltun ((2014)) uses a constrained guided policy search that can gradually bring the trajectories in agreement with the policy. Such a process that employs trajectory optimization avoids the need for random exploration in the environment, solving the problems towards the exploration exploitation dilemma in reinforcement learning. Optimizing the policy based on the optimized trajectory is then simply a step to follow the actions in each trajectory, Training the policy is then simply a supervised learning step, where it can be specifically done on distributions that were generated by trajectory optimization. Alternating policy and trajectory optimization is therefore gradually brought into agreement so that the final policy is trained on its own distribution.

2.2 Variational Policy Search

Levine and Koltun ((2013b)) further uses trajectory optimization as a powerful exploration strategy that can guide the policy search. They show how a variational decomposition of a maximum likelihood policy objective allows the use of standard trajectory optimization algorithm to be interleaved with standard supervised learning for the policy itself. In this method, exploration is performed by a trajectory optimization algorithm. Their work converts the task of finding an optimal policy into an inference problem, where using the dynamics distribution and the policy, a Bayesian network can be defined that relates states and actions. Policy optimization can then be performed by learning the maximum likelihood values for the policy parameters. The policy optimization objective function can further be decomposed by using a variational distribution $q(\alpha)$ which introduces a KL divergence term. The variational policy search algorithm therefore alternates between minimizing the KL divergence with respect to $q(\alpha)$ and maximizing a bound with respect to the policy parameters.

2.3 Approximate Inference in Trajectory Optimization

In our work, we will mainly build on top of Toussaint ((2009)). A classical solution in stochastic optimal control is to compute an optimal deterministic trajectory and then solve a linear-quadratic Gaussian (LQG) model to handle system stochasticity. The algorithm Approximate Inference Control (AICO) presents a probabilistic model for which the maximum likelihood trajectory coincides with the optimal trajectory. The algorithm then uses approximate inference to generalize to non-LQG systems.

In AICO, the trajectory optimization problem is solved using sequential quadratic programming (SQP) schemes, and this algorithm rather than computing the global (over the full time interval) trajectory, iteratively updates the local messages. The way the approach differs is that it considers a maximum likelihood solution that can coincide with the optimal trajectory, and then use approximate inference methods to efficiently find the ML trajectory and the local policy around this trajectory.

We now describe the probabilistic inference approach that AICO takes. While the classical approach to design a good trajectory is to define a cost function and minimize the expected cost given a stochastic control model, AICO takes the approach to design a good trajectory model as conditioning the probabilistic trajectory model on desired criteria and then consider the problem of inferring the posterior distribution of trajectories conditioned on these criteria. Similar to the classical RL approach where cost is defined in terms of states and actions, in the probabilistic approach, we will consider the negative log approach by introducing an extra binary random variable. AICO then proves that the maximized ML trajectory is equivalent to the expected cost minimization.

2.4 Learning Transition Dynamics

To work in a fully Bayesian manner in deep reinforcement learning, we will further consider learning the transition model dynamics using a probabilistic non-parametric Gaussian processes transition model based on

work from Deisenroth et al. ((2015)). By learning a probabilistic model, we can place a posterior distribution over plausible transition functions and express the level of uncertainty about the model itself. Such model uncertainty can further be incorporated into planning and policy evaluation. Work in this paper therefore involves using a nonparametric Gaussian process probabilistic model, and then using computationally efficient deterministic approximate inference for long-term predictions and policy evaluation. It further shows that combining a probabilistic dynamics model and Bayesian inference leads to automatic exploration as long as the predictions are far from the target-even for a policy, which greedily minimizes the cost objective function we want to minimize to maximize reward. Once close to the target, the policy does not substantially deviate from a confident trajectory that leads the system close to the target.

3 Data-Efficient Approximate Inference for DNN policies

Based on the prior work, we will therefore devise our algorithm that uses approximate inference or sampling methods for obtaining a posterior distribution over trajectories. Once we formulate the problem as an optimization over trajectory distribution using approximate inference, then similar to work done in Levine and Koltun ((2014)) we can then alternate between optimizing the policy to match the trajectories and optimize the trajectory to match the policy and minimize cost. Our work will also be based on guiding samples of trajectories using scalable inference methods; where the major benefit of using such an approach is data efficiency. By using message passing algorithms and scalable expectation propagation or variational methods based approaches, we can optimize this step using less roll outs of trajectories, working in a fully Bayesian manner, while also automatically incorporating uncertainty and exploration in the environment. By obtaining a posterior distribution over trajectories, we will incorporate more automatic exploration of the state space, while the policy learning step would be simply be based on training using the optimal trajectories.

Although sampling based methods such as sequential Monte Carlo can provide more accurate estimates of trajectory distributions, but it will be less data inefficient since we would need to draw large number of samples from trajectories, leading to lots of roll outs of trajectories; therefore using sampling based methods leaves room for comparison. Compared to that, we can formulate a message passing based approach such as Expectation Propagation to compute the approximate posterior distribution over trajectories much more efficiently. Our work even though initally will be based on model-based RL, but future work will further depend on whether we can extend our approximate inference guided policy search method to unknown dynamics based on work from Levine and Abbeel ((2014)).

4 Experimental Framework and Infrastructure

Robotic Locomotion: We will conduct our experiments using the MuJoCo simulator Todorov et al. ((2012))and include models such as swimmer, hopper, and walker for evaluation, and compare our results with current RL approaches based on using the MuJoCo simulator.
Playing Games from Images: We will further investigate our method for playing Atari games based on recent work from Lillicrap et al. ((2015)). We will train our policies for playing Atari games using raw images as input and test our algorithm on the same seven games reported in Lillicrap et al. ((2015))

Infrastructure for project: This project would not require any large scale computing resources. We would run our experiments using the MuJoCo simulator discussed above, which is a physics simulation engine to run benchmark RL experiments. We might need, however, to run our experiments using the mlsalt servers, which would be available throughout the course of the project.

5 Conclusion

We will therefore investigate using approximate inference methods for trajectory optimization to achieve data efficiency in deep reinforcement learning. Similar to guided policy search, we will first decompose the problem into learning locally linear approximations of the dynamics using scalable approximate inference

methods, and then use optimal control to find locally-linear optimal policy along these trajectories. Finally, we can use supervised learning or Bayesian methods to train a complex non-linear policy, such as deep neural network policies, to reproduce the state to action mapping found in the first phase. Based on our approach, we will investigate learning both stochastic and deterministic policies for continuous action.

References

M. P. Deisenroth, D. Fox, and C. E. Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *CoRR*, abs/1502.02860, 2015. URL `http://arxiv.org/abs/1502.02860`.

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, 2013a. URL `http://jmlr.org/proceedings/papers/v28/levine13.html`.

S. Levine and V. Koltun. Variational policy search via trajectory optimization. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 207–215, 2013b. URL `http://papers.nips.cc/paper/5178-variational-policy-search-via-trajectory-optimization`.

S. Levine and V. Koltun. Learning complex neural network policies with trajectory optimization. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 829–837, 2014. URL `http://jmlr.org/proceedings/papers/v32/levine14.html`.

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`.

E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pages 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109. URL `http://dx.doi.org/10.1109/IROS.2012.6386109`.

M. Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1049–1056, 2009. doi: 10.1145/1553374.1553508. URL `http://doi.acm.org/10.1145/1553374.1553508`.