Data Efficient Deep Reinforcement Learning with Bayesian Optimization

Submitted By: Riashat Islam; Supervisor: Marc Deisenroth

15th Jan 2016

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 episodes 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 using pixel information only with approaches based on Bayesian Optimization. The fundamental goal of our approach is to learn quickly using least number of system interactions and incorporate uncertainty into learning optimal policies.

In our work, we will consider probabilistic approaches to learning controllers, using approximate inference and Bayesian optimization methods for trajectory optimization. Following a policy based on the optimized trajectory, we can learn optimal complex non-linear neural network policies much faster than current approaches based on action-value or 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 our work, we will therefore consider probabilistic models of controller learning, evaluate approximate inference methods for trajectory optimization, and use Bayesian optimization methods, combined with guided policy search approaches for optimal policy search. Our approximate inference methods can create locally linear approximations of the dynamics incorporating exploration and data efficiency. Bayesian optimization based approaches can further enhance trajectory optimization in guided policy search approaches, further taking account of the exploration-exploitation tradeoff. 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 robotic tasks based experiments, preferrably using the MuJoCo simulator, comparing with other state of the art benchmark tasks.

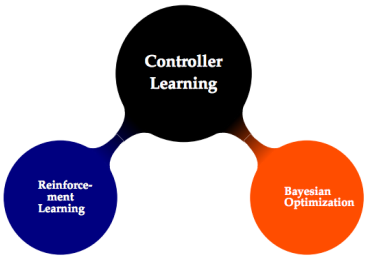

Figure 1: Controller Learning with Reinforcement Learning and Bayesian Optimization

Figure 2: Humanoid Robot iCub

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 ((2013)) 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 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.

2.3 Bayesian Optimization

Bayesian optimization has been successfully applied to policy search methods in robotics and reinforcement learning. In robotics applications, policy parameterization and policy search techniques are used to navigate a robot, while minimizing uncertainty about its own location and map estimates. Furthermore, BO has been used in hierarchical reinforcement learning, to tune the parameters of a neural network policy Brochu et al. ((2010)).

There has been a recent surge of interests in practical Bayesian optimization for machine learning algorithms Snoek et al. ((2012)). Furthermore, Wilson et al. ((2014)) considered using Bayesian optimization for optimization of parametric policies, considering Bayesian prior information about the expected return. Wilson et al. ((2014)) discusses how the process of selecting new policies accounts for the agent's uncertainty in performance estimates, and directs the agent to epxlore new parts of the policy space where uncertainty is high, and BO can tackle such problems. BO provides a method of planning a sequence of queries from an objective function for the purpose of seeking the maximum, and how the uncertainty of the objective function can be encoded in a Bayesian prior distribution that can estimate the performance of policies.

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 Bayesian Optimization in Reinforcement Learning

In Bayesian optimization, we consider finding the minimum of a function $f(x)$ using relatively few evaluations, by constructing a probabilistic model over $f(x)$. For continuous functions, BO works by assuming the unknown function was sampled from a Gaussian process, and maintains a posterior distribution for this function as observations are made. The main philosophy of BO is to find the minimum of an unknown non-convex function, by using all the information from previous evaluations of $f(x)$ and not simply relying on the local gradient.

In our work, we will consider Bayesian Optimization with effective acquistion functions to learn policies with as few samples as possible. In direct policy search based approaches, evaluation of expected returns using MC simulations can be very expensive, especially with recent approaches based on neural network policies. This requires learning policies, and find a peak of the cumulative reward objective function with as few policy iterations as possible. Similarly problems arise when neural network based value function approximators are to be learnt only over the relevant regions of the state space. In such cases, BO can be used to provide an extensive exploration-exploitation mechanism for finding the relevant regions and fitting value functions.
Bayesian optimization is particularly useful in policy gradient based direct policy search approaches. This is because it is derivative free, and hence is less prone to being stuck in a local optima, and can be explicitly designed to minimize the number of expensive value function evaluations. Furthermore, BO and Bayesian active exploration can be used to significantly speed up the learn process in hierarchical reinforcement learning tasks (HRL) here the idea is to structure the policy into tasks composed of subtasks, and are specific to a subset of the total world state space.

Bayesian optimization methods can further have applications in attention based models that uses policy gradient based reinforcement learning. Recently, Mnih et al. ((2014)) applied the REINFORCE algorithm for hard attention using a RNN model capabale of extracting information from an image or video by adaptively selecting a sequence of regions or locations and only processing the selected regions at high resolution. Mnih et al. ((2014)) considered training the model using reinforcement leanring methods to learn task-specific policies, where the agent is built on a RNN where at each time it processes the sensor data, and chooses how to act and deploy its sensor at next time step.

Our hypothesis is that BO based attention models can further have applicatons in data-efficient robot learning tasks based on pixel-information only, where attentional models can be used for simulatenous object tracking and recognition drived by gaze data. Such an approach was considered in Denil et al. ((2012)) where using attention models, the control pathway can model the location, orientation, scale and speed of the atteneded object, and learn to select gaze so as to minimize tracking uncertainty.
Bayesian optimization can therefore be successfully used to optimize parametric policies in challenging RL applications. It can be particularly useful for data-efficient learning, since it exploits Bayesian prior information about the expected return and exploits this knowledge to select new policies to execute. It also effectively addresses the exploration-exploitation tradeoff. Furthermore, considering our Guided Policy Search (GPS) approach, we can further use BO to exploit the sequential trajectory information generated by agents. Inspired by work from Wilson et al. ((2014)), we can further BO with Gaussian Process kernels to measure the similarity between policies using trajectory data generated from policy executions, that can perhaps further improve posterior estimates of the expected return, in turn improving the quality of exploration.

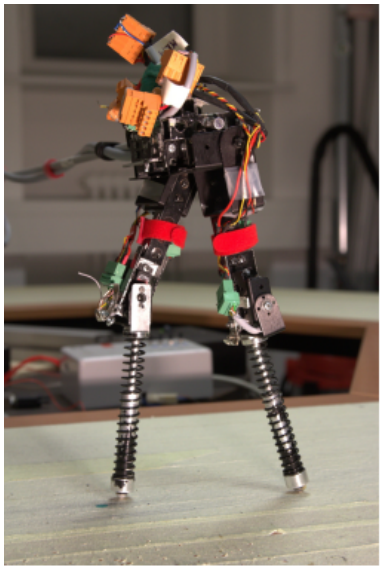

Figure 3: Bio-inspired dynamical bipedal walker using Bayesian Optimization

5 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. We will assess the proposed methodology of data-efficient learning using Bayesian optimization for autonomous learning from high-dimensional synthetic image data, and also consider experiments on single and planar double pendulum. We will evaluate our model on state of the art RL methods for continuous state and actions, and how the approach can be used over high-dimensional state spaces. Our results will be an important step towards data-effiicient fully autonomous end to end learning based on pixel information or observations.

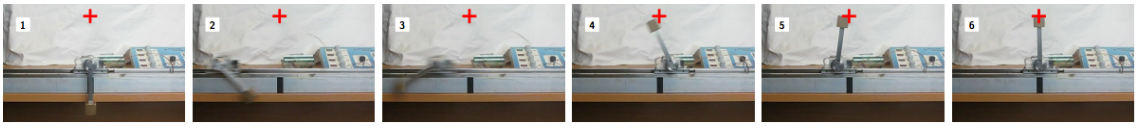

Figure 4: Real cart-pole system: Swing-up and balanc

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.

6 Initial Six Month Project Outline

In the first month of the project, we will consider building up work from Levine and Abbeel ((2014)), and build our experimental framework. In the subsequent three months, we will consider implementing approximate inference and Bayesian optimization based trajectory optimization methods for data-efficient learning. Once the model has been built, the last two months will consider evaluating the performance of our approach, compare it with other existing methods, and finally show how, based on our hypothesis, our model can achieve better data-efficient learning on similar tasks that were also done for robot learning and by using the MuJoCo simulator.

7 Conclusion

Our work will therefore investiagate Bayesian optimization and approximate inference based probabilistic approaches for data efficient 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. We will also present how Bayesian optimization can be considered as a useful tool in deep reinforcement learning. Based on our approach, we will investigate learning both stochastic and deterministic policies for continuous action.

References

E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010. URL `http://arxiv.org/abs/1012.2599`.

M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 24(8):2151–2184, 2012. doi: 10.1162/NECO_a_00312. URL `http://dx.doi.org/10.1162/NECO_a_00312`.

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-learningneuralnetworkpolicieswithguided-policy-searchunderunknowndynamics`.

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 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`.

V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2204–2212, 2014. URL `http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention`.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 2960–2968, 2012. URL `http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms`.

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`.

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`.