Bayesian-Adaptive Deep Reinforcement Learning using Model Ensembles

Gilwoo Lee gilwoo@cs.uw.edu
Jeongseok Lee jslee02@cs.uw.edu
Brian Hou bhou@cs.uw.edu
Aditya Vamsikrishna adityavk@cs.uw.edu

Abstract

lalala abstract!

1 Introduction

While reinforcement learning is capable of controlling complex autonomous systems, RL algorithms typically require huge amounts of data, can overfit to a particular task, or may learn brittle policies that are prone to disturbances. One of the main challenges that needs to be addressed is to train a policy that is robust to various model uncertainties and disturbances. In this project, we aim to address this challenge via an ensemble policy for Bayes-Adaptive Reinforcement Learning [1].

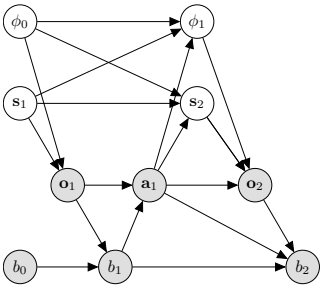
We model b_ϕ as a network capable of modeling evolving state change, e.g., Recurrent Neural Networks, or as a Bayes filter. At the low level, we plan to discretize Φ and have one actor-critic network per discretized value of ϕ: each critic estimates $V^*(\cdot;\phi)$ and each actor has an optimal policy for a particular discretized value of $\pi^*(\cdot;\phi)$. Given b_ϕ and the set of value-function approximators, it is straightforward to compute (2).

2 Related Works

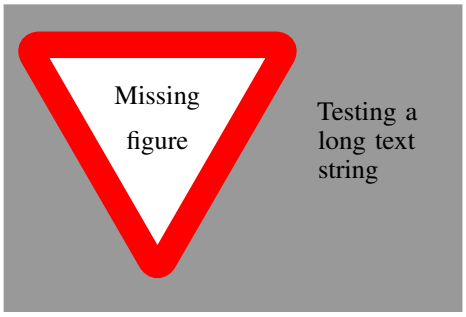
Our work is closely related to QMDP [2, 3] which is an approximation for POMDP. QMDP approximates POMDP by assuming fully-observable MDP after 1-step, and approximating the Q-value at the current belief state $b(s)$ as $Q_a(b) = \sum_s b(s) Q_{\text{MDP}}(s,a)$. In our problem setup, we have a belief over the physics parameters ϕ of the MDP, $b(\phi)$, and we compute the policy $Q_a(s;b) = \sum_\phi b(\phi) Q_{\text{MDP}}(s,a;\phi)$.

The BAMDP formulation is also similar to POMDP formulation used in POMDP-lite [4] which assumes that the hidden state variables are constant or only change deterministically. In our case, the hidden state variables correspond to the physics parameters ϕ. The authors of POMDP-lite have shown that such formulation is "equivalent to a set of fully observable Markov decision processes indexed by a hidden parameter" [4], which, in our case, is a discretization of ϕ.

3 Bayes-Adaptive Reinforcement Learning



(a) Graphical Model



(b) System structure

We assume that there exists a latent physics variable ϕ which determines the transition function of the underlying MDP, i.e., the transition function $P(s', \phi' | s, \phi, a)$ is now a function of state, action, and ϕ. We would like to learn a policy which maximizes the long term reward given ϕ. Formally, this is called a Bayes-Adaptive MDP [1, 5, 6], defined by a tuple $\langle \mathcal{S}', \mathcal{A}, P, P_0, R \rangle$ where

- $\mathcal{S}' = \mathcal{S} \times \Phi$ is the set of hyper-states (states, physics variable),
- $\mathcal{A}$ is the set of actions,
- $P(s', \phi' | s, \phi, a)$ is the transition function between hyper-states, conditioned on action a being taken in hyper-state (s, ϕ),
- $P_0 \in \mathcal{P}(\mathcal{S} \times \Phi)$ combines the initial distribution over hyper-states,
- $R(s, \phi, a)$ represents the reward obtained when action a is taken in hyper-state (s, ϕ).

We would like to find the Bayes-optimal policy for the following Bellman equaton:

$$V^*(b, s) = \max_a \left\{ R(b, s, a) + \gamma \sum_{s'} P(s' | b, s, a) V^*(b', s') \right\}. \tag{1}$$

where b is the belief over the set of latent physics parameters $\phi \in \Phi$.

We make a simplification to the BARL formulation. We assume that the latent variable ϕ is either constant or the rate of change is slow enough that approximating the long-term value with a determinized ϕ is a reasonable short-term approximation for choosing one-step action, i.e. we can treat $V^*(s_t, \phi_t) \approx V^*(s_t, \phi_{t:\infty})$ for the purpose of one-step Bellman update.

This assumption allows us to simplify BARL with an ensemble policy learning method. At a high level, we have a network that updates the *belief* of the physics parameters at time t,

$$b(\phi_t) = P(\phi_t | s_{t-1}, \phi_{t-1}, a_{t-1})$$

which is then used to compute the best policy from an ensemble of ϕ-dependent optimal policies, i.e., $\pi^*(\cdot; \phi)$ and $V^*(\cdot; \phi)$ are computed with typical RL algorithms for MDPs. Then the remaining task is to compute the one-step best action a:

$$a^* = \arg\max_a \mathbb{E}_{\phi \sim b(\phi)} \left[R(s, a, \phi) + \gamma \sum_{s', \phi'} P(s', \phi' | s, \phi, a) V^{*'}(s', \phi') \right]. \tag{2}$$

4 Bayes-Adaptive Policy Network

TODO: Explain how the system works (input to the network, etc.)

5 Experiments

We have setup a set of simulated examples and a set of RL algorithms to be utilized in our ensemble approach. For simulated examples, we have the following agents: **ant, reacher, swimmer, half-cheetah**, each with a predefined reward function defined similar to those given by OpenAI Gym [7].

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
    read current;
    if *understand* **then**
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end

Algorithm 1: TODO: write the algoruthm. (Change this to a prettier algorithm package)

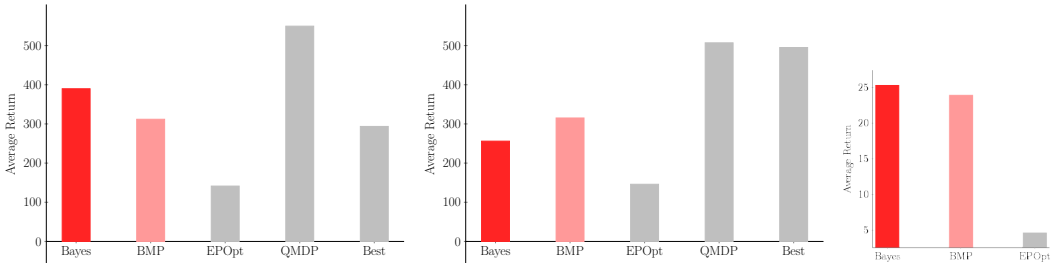
We are utilizing TRPO [8], VPG, DDPG [9] provide by `rllab` [10] as a set of algorithms to be utilized in our ensemble approach. In addition, we plan to implement PPO and a PID controller for RACECAR.

Our algorithm will be compared against two classes of algorithms: (1) sample-based algorithms which chooses an MDP and commit to this policy for a fixed horizon, and (2) ensemble algorithms which train a policy over an ensemble of MDP models. A greedy algorithm which chooses the maximum-likely MDP, or one that samples from a posterior distribution of MDPs given previous observations (e.g. Posterior Sampling Reinforcement Learning [11]) would fall into the former, and EPOpt[12] and Ensemble-CIO [13] would fall into the latter.

We are currently in the process of setting up baseline algorithms which may be used for direct comparison or as internal policy update algorithms for each of MDP in our algorithm.

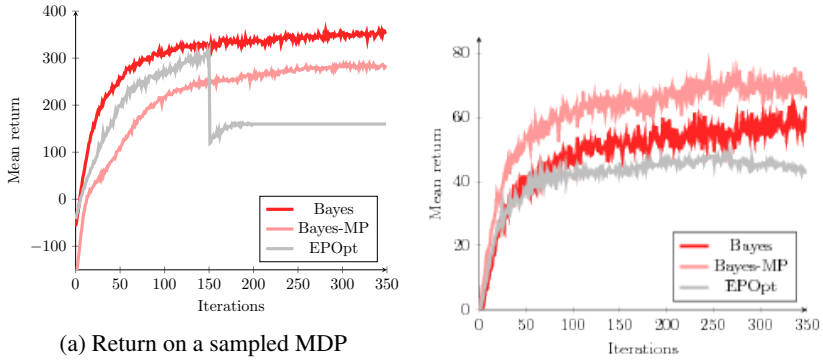
(a) Nominal (b) Key frames (c) Key frames

Figure 2: Keyframes from rollouts on various MDPs. The universal policy network learns to use different policies on different MDPs.



(a) Return on a sampled MDP (b) Return on one of the chosen MDPs (c) Average return

Figure 3: Return on sampled and prechosen MDPs. BARL is better than EPOpt. On average across K pre-chosen MDPs, it outperforms EPOpt by a large margin. TODO: remove BMP.



(a) Return on a sampled MDP

Figure 4: Training curves

References

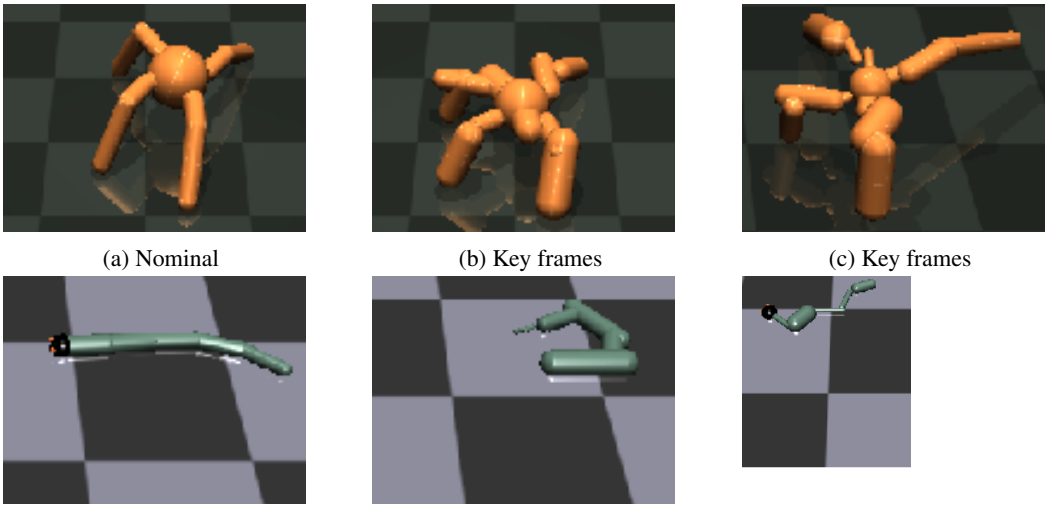
[1] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, *et al.*, "Bayesian Reinforcement Learning: A Survey," *Foundations and Trends® in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.

[2] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Machine Learning Proceedings 1995*, pp. 362–370, Elsevier, 1995.

[3] P. Karkus, D. Hsu, and W. S. Lee, "QMDP-Net: Deep Learning for Planning under Partial Observability," in *Advances in Neural Information Processing Systems*, pp. 4697–4707, 2017.

[4] M. Chen, E. Frazzoli, D. Hsu, and W. S. Lee, " POMDP-lite for Robust Robot Planning under Uncertainty," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pp. 5427–5433, IEEE, 2016.

[5] S. Ross, B. Chaib-draa, and J. Pineau, "Bayes-Adaptive POMDPs," in *Advances in Neural Information Processing Systems*, pp. 1225–1232, 2008.



(a) Nominal (b) Key frames (c) Key frames

(d) Nominal (e) Key frames (f) Key frames

Figure 5: Keyframes from rollouts on various MDPs. Some of the MDPs are gemeotrically significatnly differnt that they require drastically different policies to achieve optimal returns.

[6] A. Guez, D. Silver, and P. Dayan, "Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search," in *Advances in Neural Information Processing Systems*, pp. 1025–1033, 2012.

[7] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," 2016.

[8] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, pp. 1889–1897, 2015.

[9] 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," *arXiv preprint arXiv:1509.02971*, 2015.

[10] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking Deep Reinforcement Learning for Continuous Control," in *International Conference on Machine Learning*, pp. 1329–1338, 2016.

[11] I. Osband, D. Russo, and B. Van Roy, "(More) Efficient Reinforcement Learning via Posterior Sampling," in *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

[12] A. Rajeswaran, S. Ghotra, B. Ravindran, and S. Levine, "EPOpt: Learning Robust Neural Network Policies Using Model Ensembles," *arXiv preprint arXiv:1610.01283*, 2016.

[13] I. Mordatch, K. Lowrey, and E. Todorov, "Ensemble-CIO: Full-body dynamic motion planning that transfers to physical humanoids," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 5307–5314, IEEE, 2015.