Bayesian-Adaptive Deep Reinforcement Learning using Model Ensembles

Gilwoo Lee, Brian Hou, Jeongseok Lee, Aditya Mandalika

Introduction

Problem: Decision-making under uncertainty and partial observability requires policies that:

- generalize over a broad range of target domains.
- are robust to model uncertainties and disturbances.

Goal: A Bayes-optimal policy satisfying the following Bellman equation (with b representing belief over MDPs):

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

Bayes-Adaptive Reinforcement Learning

Proposal: A policy-gradient algorithm for Bayes-Adaptive RL which takes belief as part of observation to learn a robust policy that is Bayes-optimal to each belief.

- **Idea**: Maintain belief over k MDPs and combine with observation as input to a policy network.
- **Motivation:**
  - k can be much smaller than full discretization of the parameter space (as required by QMDP) and still suffice to cover the range of dynamics
  - The learner optimizes for the set of (belief, observation) pairs that are expected to be experienced in test time, so it can be Bayes-optimal in this space.

Related Work

QMDP (Littman et al., 1995)

- **Idea**: Approximates POMDP Q-function by assuming a fully-observable MDP after the first action.
- **Limitation**: Requires (approximate) MDP Q-function. If belief space is discretized, the belief-space dimension gets large. Assumes determinized belief after 1-step, so it's not Bayes-optimal.

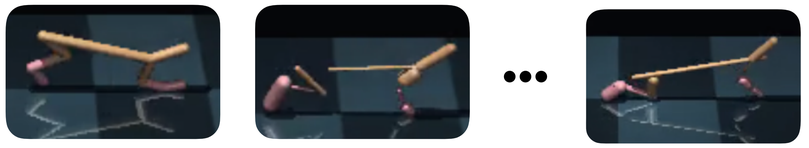
EPOpt (Rajeswaran et al., 2017)

- **Idea**: Learns policy that maximizes worst-case performance across multiple MDPs.
- **Limitation**: Minimax algorithm, designed for worst-case scenarios. If the MDP space is too large, this may result in a very pessimistic behavior.

UP-OSI (Yu et al., 2017)

- **Idea**: Learns policy from observation augmented with estimated parameters (via system identification).
- **Limitation**: No notion of "belief" or "uncertainty", so it can aggressively push for one policy when it should be more careful

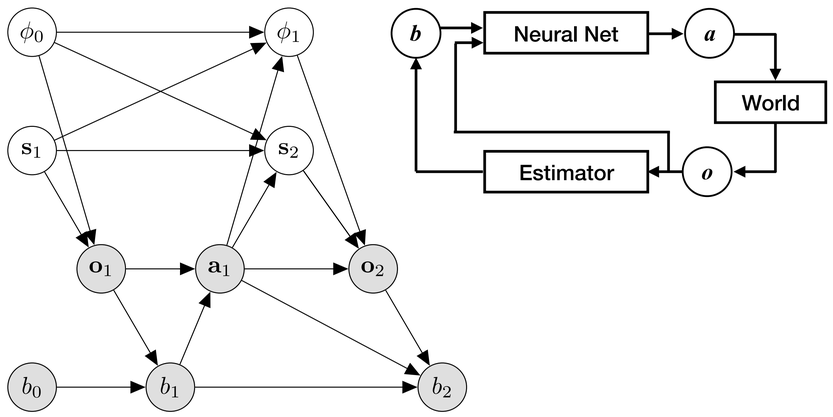
Offline: Learns different policies for different MDPs



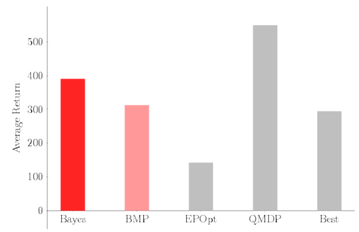
Front leg | Front foot | Back knee

Learned to use different moving rhythms per MDP to maximize forward velocity.

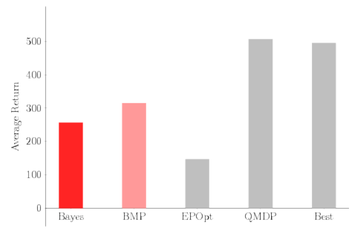
Overview of Bayes-Adaptive DRL



Robust yet higher rewards than EPOpt

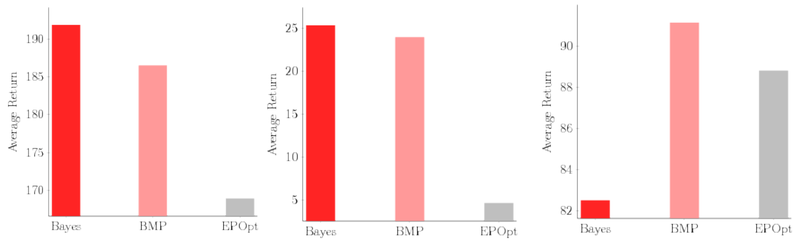


Cheetah: A sampled MDP

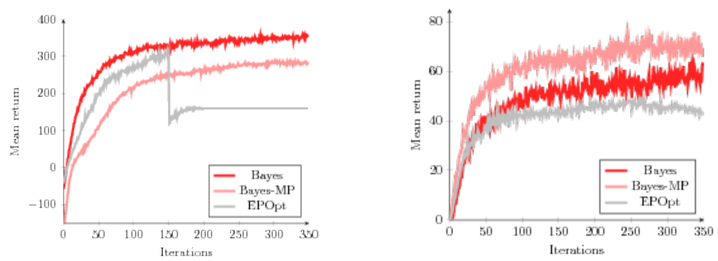


Cheetah: A pre-chosen MDP

Average performance is higher



Cheetah | Ant | Swimmer



Cheetah | Swimmer

Experiments

- MuJoCo environments: Ant, Swimmer, HalfCheetah
  - Uniformly sampled 20 MDPs across a range of parameters (e.g., body link length, mass, geometry size, joint damping, friction)
- Algorithm Implementations:
  - All policy networks were Gaussian MLPs with two layers, (64, 64), except for Bayes ones which had (128, 64) to account for the larger input space
  - TRPO implementation from rllab
  - EPOpt implemented based on the paper
  - Bayes-Mixture Policy uses TRPO policies trained on the 20 MDPs to mix actions based on the belief
  - QMDP uses TRPO policies trained on the 20 MDPs to rollout trajectories and approximate Q-functions

Analysis

- By augmenting the observation with a belief over MDPs, policy networks can learn to be robust against model uncertainty while maintaining some of the "optimal" actions w.r.t. each MDP.
- When the optimal policies across MDPs have a lot in common (e.g., Swimmer), simple "interpolation" of the deterministic policies provide good action proposals, suggesting that a mixture of policies (with a large number of policies that cover the space), may reduce sample complexity and offer even better performance.

Future Work

- Extend to a continuous version which has Gaussian belief distribution as input
- Bootstrap a set of stochastic/deterministic policies, each trained for one MDP or a small range of parameters
- Train with additional reward bonus for information gain which helps distinguishing policies (not just beliefs)

Chen et al. POMDP-lite for robust robot planning under uncertainty. *ICRA*, 2016.
Duan et al. Benchmarking Deep Reinforcement Learning for Continuous Control. *ICML*, 2016.
Ghavamzadeh et al. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 2015.
Guez et al. Efficient Bayes-adaptive reinforcement learning using sample-based search. *NIPS*, 2012.
Karkus et al. QMDP-Net: Deep learning for planning under partial observability. *NIPS*, 2017.
Littman et al. Learning policies for partially observable environments: Scaling up. *Machine Learning Proceedings*, 1995.
Rajeswaran et al. EPOpt: Learning robust neural network policies using model ensembles. *ICLR*, 2017.
Ross et al. Bayes-Adaptive POMDPs. *NIPS*, 2008.
Yu et al. Preparing for the Unknown: Learning a Universal Policy with Online System Identification. *RSS*, 2017.