Motivation

Subtle differences in the underlying dynamics of similar, but not identical, processes provide an intriguing application of transfer learning. Processes that probe similar, but not identical, dynamics (e.g. medical treatment of patients) often allow the agent to learn in one instance, but the agent for the other instance has few observations. The agent for one instance explored, but the agent for the other instance has few observations. Depending on interactions between the state space and hidden parameters or interactions between state dimensions.

Hidden Parameter Markov Decision Processes (HiP-MDP)

Doshi-Velez and Konidaris\(^1\) introduced the HiP-MDP to address the transfer between closely related tasks. The approximate transition model is defined by:

\[
K(s, \pi, s') = \sum_{a} P_{\text{BNN}}(s'|s, a) \pi(a|s) + \epsilon
\]

where \(P_{\text{BNN}}(s'|s, a)\) is the BNN predictive distribution for state \(s'\) given state \(s\) and action \(a\), \(\epsilon\) is drawn from a normal distribution.

HiP-MDP Control Policy Learning

We demonstrate the capability of the updated HiP-MDP to flexibly learn separate and optimal policies for different instances of the canonical acrobot domain and a healthcare domain. HiP-MDP learns optimal control policy more efficiently than other model baselines.

HiP-MDP Performance Improvement

We demonstrate the improved scalability of our approach by comparing the time taken to train the model as more observations are collected from new instances with different dynamics.

Inference and Policy Learning

BNN weights and latent parameters: The structure of the BNN allows for iterative and independent updates of both the network parameters as well as the latent weights \(w_b\) using a procedure introduced by Depeweg et al.\(^2\)

Deep Q-Network (DDQN): The control policy is \\(\epsilon\)-greedy policy that is learned by approximating the action-value function with a Double Deep Q Network \(^3\) with prioritized experience replay \(^4\).

Toy Example: The agent starts in the lower left corner and tries to reach the goal region. Each instance is assigned a hidden latent class that determines whether the wall is on the bottom of or the left side of the goal region.

Single Instance Example: Agent trajectories converge to optimal policy from single instance of a simple toy example.

Quantifying Uncertainty

We demonstrate the capability of the HiPMMDP to model the joint uncertainty between the latent parameters and the state space by comparing the variance of the BNN’s predictions using the latent weights of two instances (red/blue) in regions where the model has captured the correct behavior. The BNN’s predictive variance is 3x greater using the latent parameters from the instance with few observations from the region.

References and Acknowledgement


We gratefully acknowledge the fruitful discussions and advice gained from members of Harvard DTAK. Taylor Killian is grateful to MIT LL for their sponsorship.

HiV Treatment

Long run policy learning comparison

Cumulative Reward Per Episode

HiP-MDP with linear \(w_b\) vs. HiP-MDP with learned \(w_b\)

References:


We gratefully acknowledge the fruitful discussions and advice gained from members of Harvard DTAK. Taylor Killian is grateful to MIT LL for their sponsorship.