Robust and Efficient Transfer Learning using Hidden Parameter Markov Decision Processes

Taylor W. Killian[1*] Samuel Daulton[1*] George D. Konidaris[2] Finale Doshi-Velez[1]

[1]Harvard University, John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA

[2]Brown University, Department of Computer Science, Providence, RI

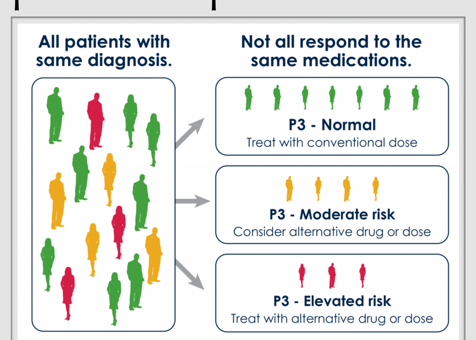
*Contributed equally as primary authors

HARVARD John A. Paulson School of Engineering and Applied Sciences

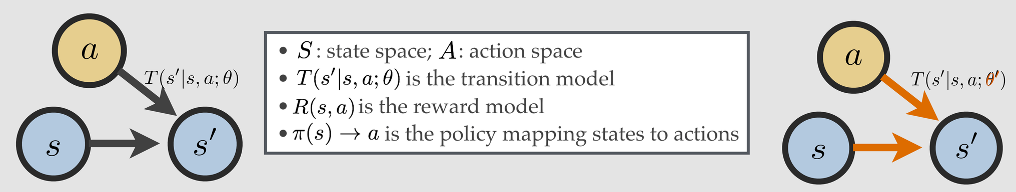
BROWN

Motivation

Subtle differences in the underlying dynamics of similar, but not identical, processes provide an intriguing application of transfer learning.



Markov Decision Processes $(S, A, T, R, \gamma) \rightarrow \pi$



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:

Equations

$$s'_d \approx \sum_{k=1}^{K} w_{kb} \hat{T}^{(GP)}_{kad}(s) + \epsilon$$
$$w_{kb} \sim \mathcal{N}(\mu_{w_k}, \sigma^2_w)$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2_{ad})$$

Variables

s : current state

s'_d : d^{th} dimension of next state

$\hat{T}^{(GP)}_{kad}$: GP basis function indexed by the k^{th} latent parameter, state dimension d, and action a

w_{kb} : k^{th} latent parameter for instance b, initialized for instance b using mean μ_{w_k} and variance σ^2_w of learned latent parameters

ϵ : random normal noise for state dimension d and action a

K : number of latent parameters

However, the original HiP-MDP had the following shortcomings:

1. It was not designed to model dynamical systems that depend on interactions between the state space and hidden parameters or interactions between state dimensions.
2. The inference procedure scaled poorly to higher dimensions.
3. It required overlapping observations of the state-action space from separate task instances in order to infer the latent parameters, which is infeasible in many settings (e.g. medical treatment of patients)

HiP-MDP with Joint Uncertainty

We augment the form the original HiP-MDP, improving the robustness and efficiency of the approximation of the transition model by:

- Embedding the latent representation w_b of the dynamics with the input
- Replacing the GP basis functions with a single Bayesian Neural Network (BNN)
- Jointly representing the full state and latent representation uncertainty via the BNN

Equations

$$s' \approx \hat{T}^{(BNN)}(s, a, w_b) + \epsilon$$
$$w_b \sim \mathcal{N}(\mu_w, \Sigma_b)$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2_n)$$

Variables

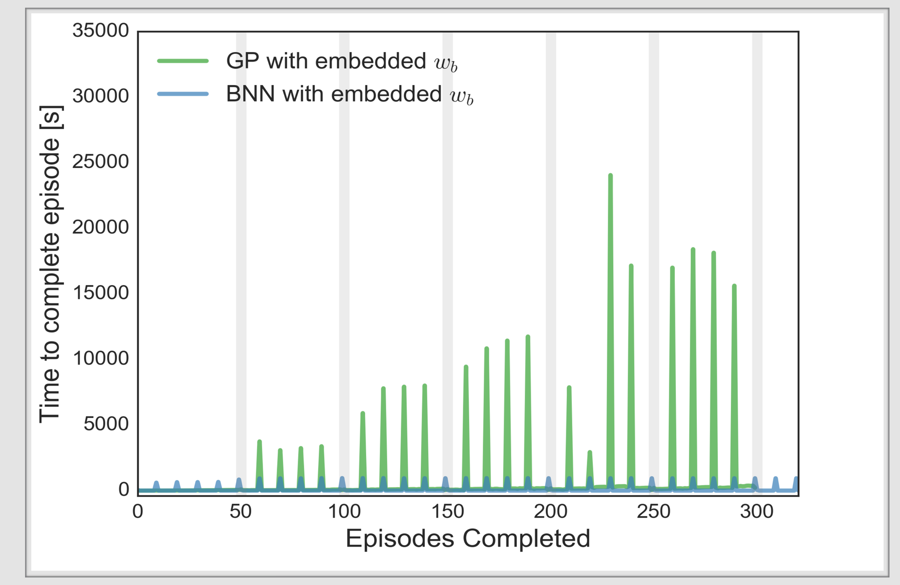
s' : next state

$\hat{T}^{(BNN)}$: BNN approximate transition function

w_b : latent embedding for instance b, initialized for instance b using mean μ_w and covariance matrix Σ_b of learned latent parameters

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.



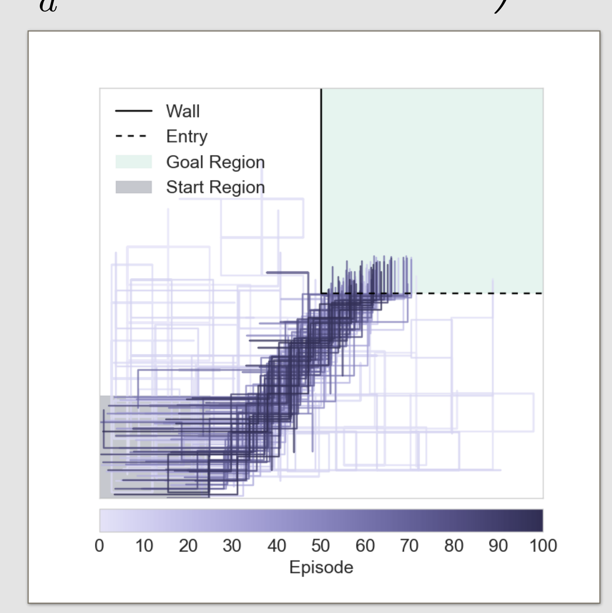
Inference and Policy Learning

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 e-greedy policy that is learned by approximating the action-value function with a Double Deep Q Network[3] with prioritized experience replay[4].

$$Q^{(DoubleQ)} \equiv R_{t+1} + \gamma\, Q\left(S_{t+1},\ \arg\max_a Q\left(S_{t+1}, a, \Phi_t\right), \Phi_t^-\right)$$

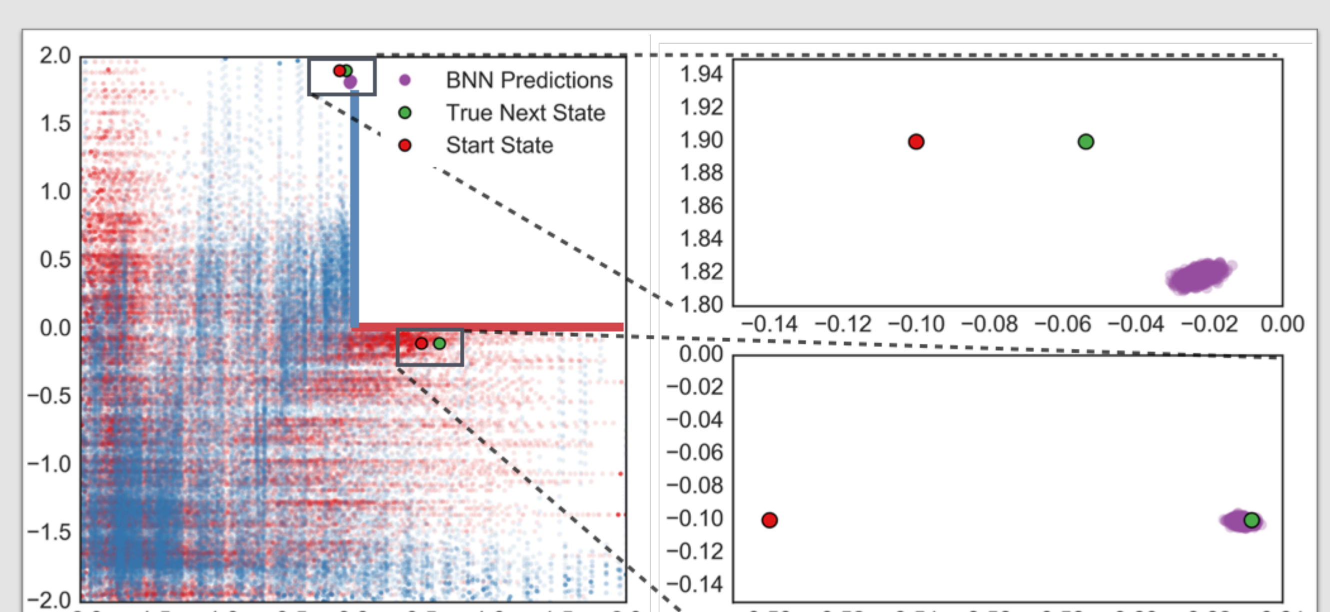
Toy Example: The agent starts in the lower left corner and tries to reach the goal region. Each instance is assigned a hidden latent class that determines whether the wall is on the bottom of or the left side of the goal region.



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

Quantifying Uncertainty

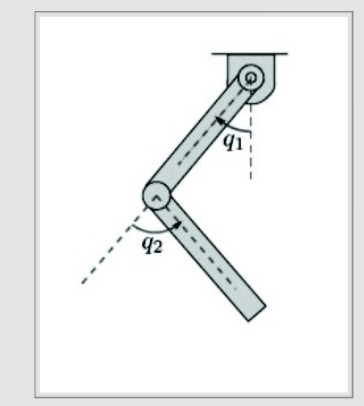
We demonstrate the capability of the HiPMDP to model the joint uncertainty between the latent parameters and the state space by comparing the variance of the BNN's predictions using the latent weights of two instances (red/blue) in regions where the agent for one instance explored, but the agent for the other instance has few observations. The BNN's predictive variance is 3x greater using the latent parameters from the instance with few observations from the region.



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.**

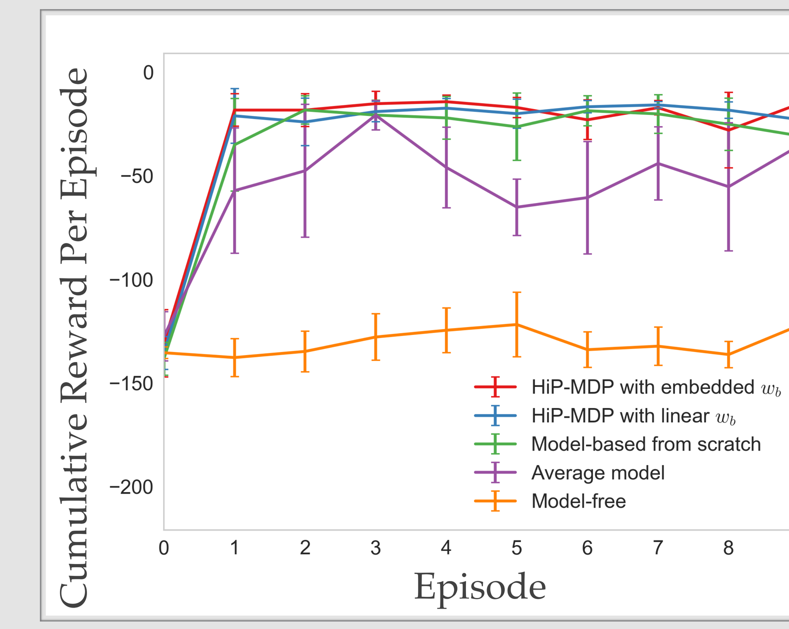
Acrobot



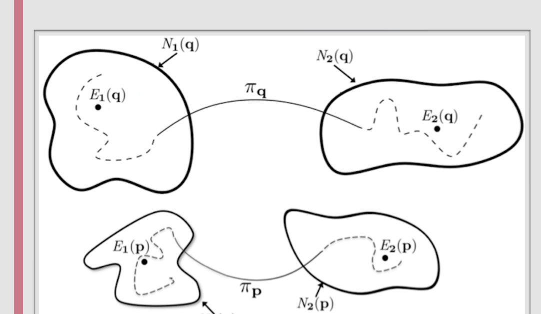
$s \in \mathbb{R}^4$

$|A| = 3$

$w_b \in \mathbb{R}^5$



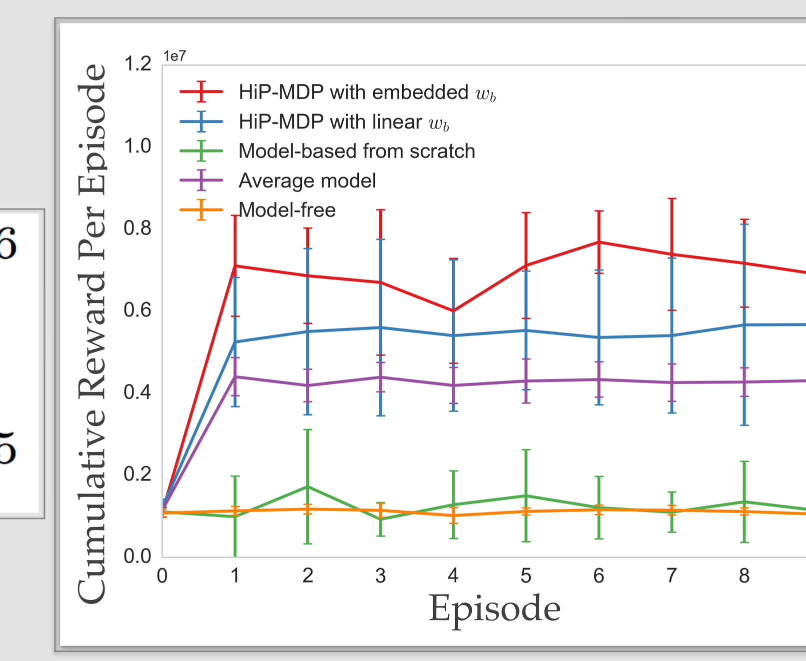
HIV Treatment



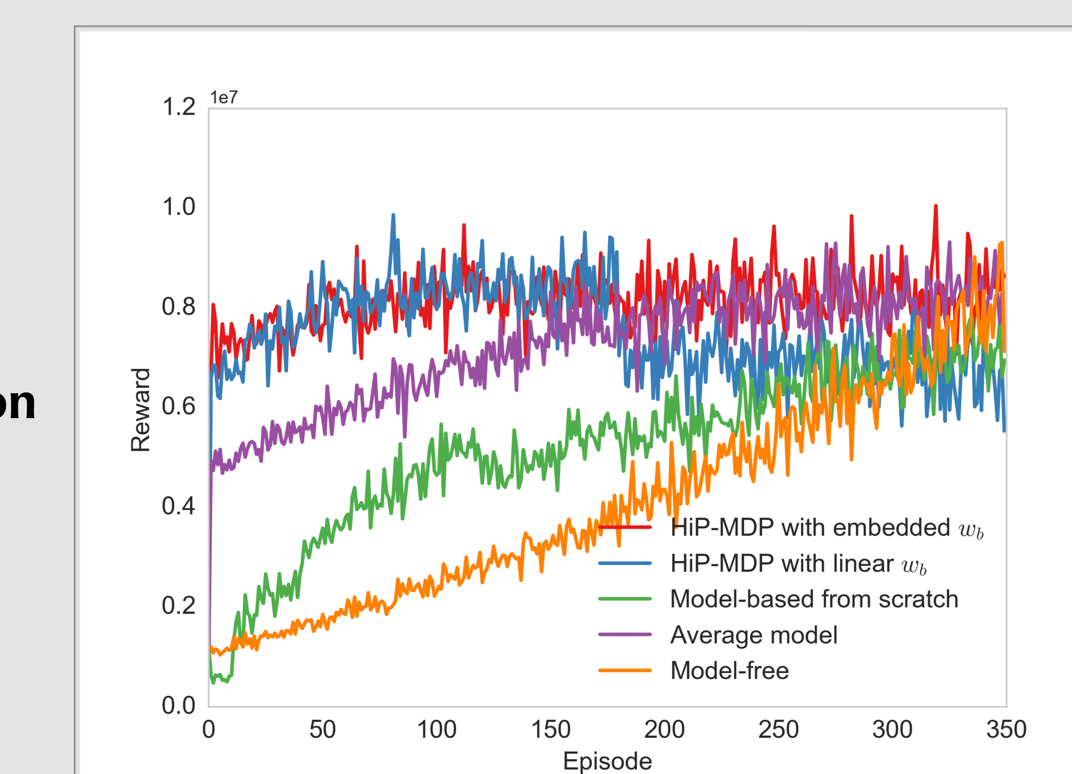
$s \in \mathbb{R}^6$

$|A| = 4$

$w_b \in \mathbb{R}^5$



Long run policy learning comparison



References and Acknowledgement

[1] Doshi-Velez, F., Konidaris, G.D; (2013). Hidden Parameter Markov Decision Processes: A Semiparametric Regression Approach for Discovering Latent Task Parametrizations. CoRR [Internet] abs/1308.3513

[2] Depeweg, S., Hernández-Lobato J.M, Doshi-Velez, F., and Udluft, S. Learning and policy search in stochastic dynamical systems with bayesian neural networks. *arXiv preprint arXiv:1605.07127*, 2016.

[3] van Hasselt, H.; Guez, A.; and Silver, D. (2016). Deep reinforcement learning with double q-learning. In Thirtieth AAAI Conference on Artificial Intelligence.

4] Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. (2015). Prioritized experience replay. arXiv preprint arXiv:1511.05952.

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