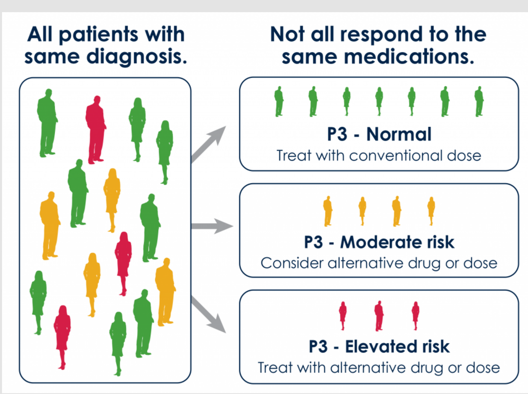
# Robust and Efficient Transfer Learning using Hidden Parameter Markov Decision Processes

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

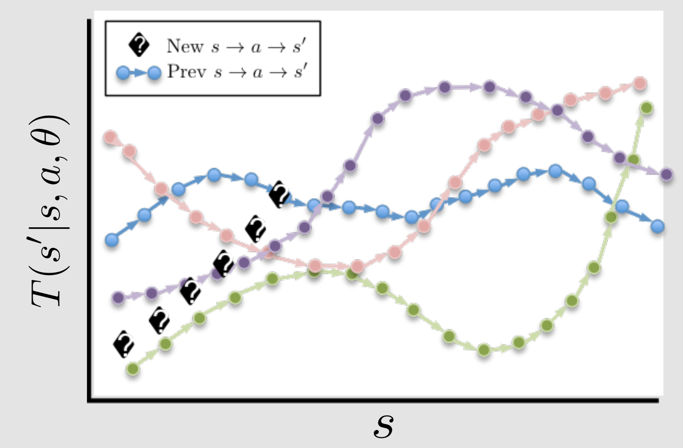
**Harvard University , Paulson School of Engineering and Applied Sciences, Cambridge, MA[1]**
**Brown University, Department of Computer Science, Providence, RI[2]**

## Motivation

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




By exploiting statistical similarities in the distributions of latent processes, one can approximate a transition model $T(s'|\,s,a,\theta_b)$ given previous observations




## Hidden Parameter Markov Decision Processes (HiP-MDP)

Doshi-Velez and Konidaris[1] introduced the HiP-MDP to address the transfer between closely related tasks. While expressive, the model is neither scalable nor efficient.

$$(s'_d - s_d) \approx \sum_k^K z_{kad} w_{kb} f_{kad}^{(GP)}(s) + \epsilon$$

$$w_{kb} \sim \mathcal{N}(\mu_{w_k}, \sigma_w^2)$$

$$\epsilon \sim \mathcal{N}(0, \sigma_{nad}^2)$$

## ~~HiP-MDP~~ with Joint Uncertainty

We augment the form the original HiP-MDP, improving the robustness and efficiency of the approximation of $T(s'|\,s,a,\theta_b)$ by:

- Embedding the latent representation $w_b$ of the dynamics $\theta_b$ with the input
- Replacing the Gaussian Process basis functions with a BNN
- Jointly representing the full state and latent representation uncertainty via the BNN

$$(s' - s) \approx f^{(BNN)}(s, a, w_b) + \epsilon$$

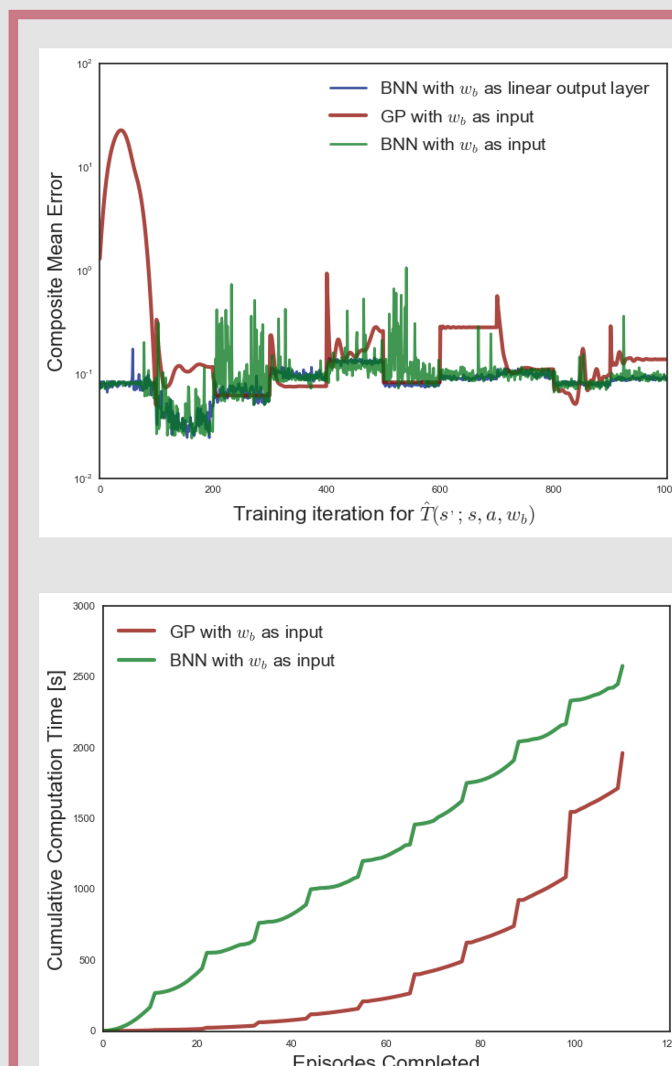
$$w_b \sim \mathcal{N}(\mu_w, \sigma_w^2)$$

$$\epsilon \sim \mathcal{N}(0, \Gamma_b)$$

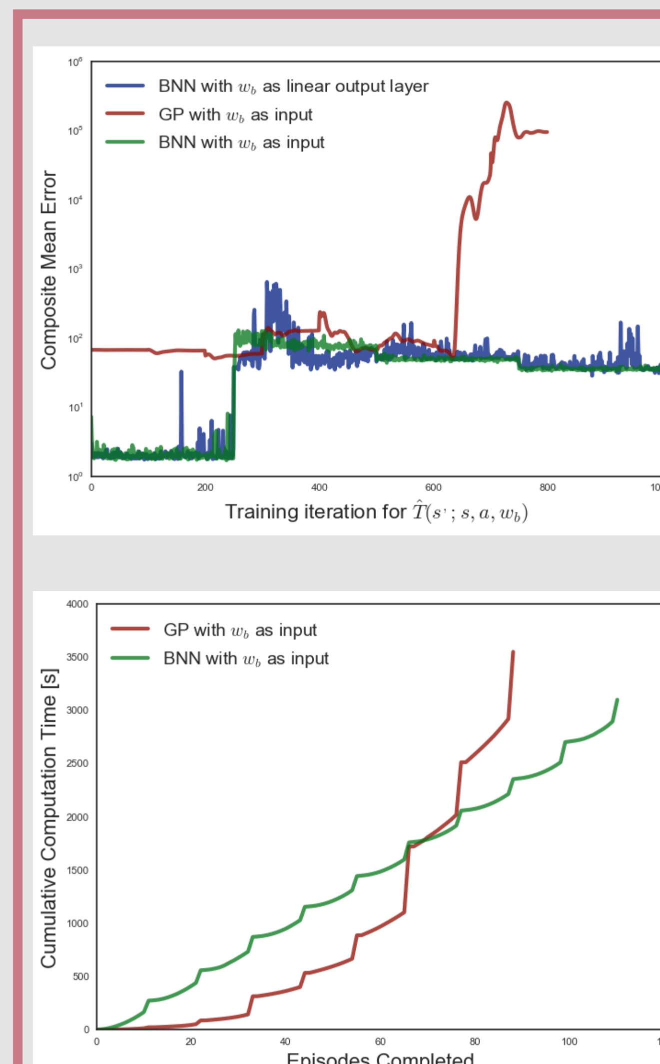
We can demonstrate the contributions of this shift in the modeling by visualizing the BNN approximation's robustness as well as it's scalability in comparison with a GP-based model

**Robust to unobserved instances of the task**

**Scalable to large state domains at higher data rates**




Toy problem
$s \in \mathbb{R}^6,\ a \in \mathbb{R}^4,\ w_b \in \mathbb{R}^3$




HIV Treatment Simulator
$s \in \mathbb{R}^6,\ a \in \mathbb{R}^4,\ w_b \in \mathbb{R}^5$

## Parameter Learning and Agent Training

The structure of the BNN allows for iterative and independent updates of both the network parameters as well as the latent weights $w_b$ following the procedure introduced by Deisenroth and Rasmussen[2].

The control policy is trained via a Double Deep Q Network[3] using prioritized experience replay[4].

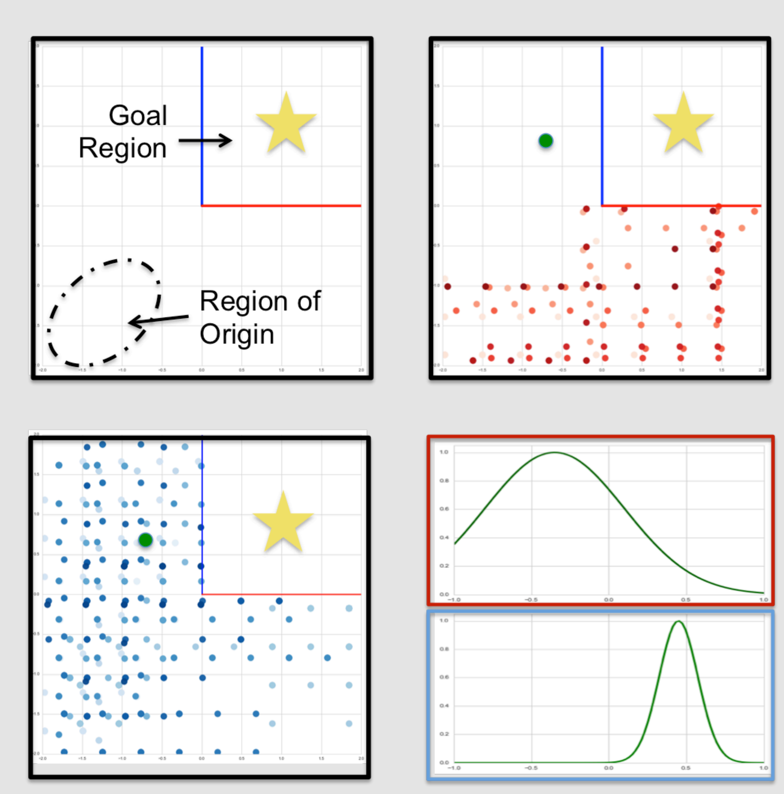
$$Q^{(DoubleQ)} \equiv R_{t+1} + \gamma\, Q\left(S_{t+1},\ \arg\max_a Q\left(S_{t+1}, a, \Phi_t\right), \Phi_t^-\right)$$

## Demonstration

We demonstrate the capability of the updated HiP-MDP with a simple toy domain. Here an agent is assigned a hidden latent class that determines how it can transition into a goal region.

Our updated model is able to flexibly learn separate policies for the different latent classes. The model is also able to infer transition uncertainty under separate latent class assumptions.

The performance of the HiP-MDP on this toy problem is encouraging for eventual application to more complex and critical domains.




## References

[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] Deisenroth, M., and Rasmussen, C. (2011). *Pilco: A model-based and data-efficient approach to policy search.* In *Advances in Neural Information Processing Systems,* volume 16, pp. 329-336

[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.*

## Acknowledgments

We gratefully acknowledge the support of Sam Daulton in finalizing the computational paradigm needed to complete this work. We also acknowledge the fruitful discussions and advice gained from members of Harvard DTAK. TWK expresses deep gratitude to MIT LL for their sponsorship.