

# Thesis Summary: Nonparametric Bayesian Approaches for Reinforcement Learning in Partially Observable Domains

**Finale Doshi-Velez**

Massachusetts Institute of Technology

## Abstract

The objective of my doctoral research is bring together two fields: partially-observable reinforcement learning (PORL) and non-parametric Bayesian statistics (NPB) to address issues of statistical modeling and decision-making in complex, realworld domains.

## Problem Addressed

The field of reinforcement learning (Sutton and Barto 1998) addresses how an agent can learn to make good decisions in new environments through an online, interactive process. More formally, the agent chooses from a set of actions  $A$  which may affect the (unobserved) state  $s$  of its environment. It then receives an observation  $o$  and a reward  $r$  from the environment. The state transition, observation, and reward models are initially unknown, and the agent’s goal is to maximize long-term rewards. Especially when parts of the environment are hidden—that is, the observation  $o$  does not reliably provide complete information about the state  $s$ —traditional reinforcement learning approaches typically require the agent to gather a large amount of experience before it can start acting near-optimally. Despite the interest in partially-observable reinforcement learning (Hutter et al. 2009), research has had limited success in scaling to real-world domains.

The objective of this research is to apply nonparametric Bayesian techniques to address issues of sample-complexity and scalability in partially-observable reinforcement learning. We focus on partially-observable reinforcement learning problems with discrete sets of actions and discrete internal representations of state. In these settings, Bayesian approaches to reinforcement learning (Poupart and Vlassis 2008; Jaulmes, Pineau, and Precup 2005; Ross, Chaib-draa, and Pineau 2008; Doshi, Pineau, and Roy 2008) have already been shown to help an agent make better use of its experience: priors guide the agent’s decisions when data is scarce, allowing it to make reasonable decisions with limited information. Bayesian approaches also allow the agent to reason about its uncertainty about the world. However, for large problems, reasoning about all sources of uncertainty is computationally expensive. Nonparametric techniques

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

promise an elegant solution to these computational issues by reasoning only about parameters that the data suggest are relevant. However, current inference methods for nonparametric Bayesian models are not particularly well-suited for the online settings found in reinforcement learning. Thus, a key aspect of this research is developing new inference methods more appropriate for reinforcement learning.

## Proposed Plan of Research

We consider three nonparametric Bayesian approaches to reinforcement learning in partially observable domains:

- A model-based approach in which the agent’s internal state representation corresponds to the “true” states in the environment (assumed to be a partially observable Markov decision process),
- A model-free approach in which the agent’s internal state representation corresponds to nodes in a finite state controller with an unknown number of nodes, and
- A structure-discovery approach for more complex environments; specifically we consider a feature-discovery setting in which part of the agent’s internal state corresponds to what features may be relevant for a particular decision.

The third approach will be applied to a healthcare application for remotely monitoring chronically-ill patients.

**Model-based learning: infinite POMDP.** We extend the nonparametric version of the standard hidden Markov model, the infinite hidden Markov model (iHMM) (Beal, Ghahramani, and Rasmussen 2002; Teh et al. 2006), to an infinite POMDP model (Doshi-Velez 2009). Belief-monitoring in the iPOMDP is accomplished by adapting the beam sampler for the iHMM (van Gael et al. 2008) to draw models from the iPOMDP posterior given the agent’s experience. A key aspect of this research is to develop an online version of the beam-sampler. A second question is that of action-selection. We have considered a variety of approaches, including stochastic forward-search, for this particle-based representation of the belief and hope to characterize their properties.

Like the iHMM, our iPOMDP model is derived from the hierarchical Dirichlet process (HDP). As with other Bayesian techniques using Dirichlet-multinomial distribu-

tions, the agent infers its transition, observation, and reward models based on counts of how often it believes it has experienced those events. Thus, given a set of somewhat-similar possible models, deep look-aheads—many counts—are needed for the agent to differentiate these models and realize the value of reducing model uncertainty.

We hypothesize that biasing the agent toward proposing more radically different models and more confident learning can alleviate this problem. We are currently trying two approaches: adjusting the concentration parameters on the iHMM and developing a hierarchical Pitman-Yor (HPY-HMM) model to induce even stronger model differentiation and sparsity.

**Model-free learning: infinite state controllers.** We focus on learning policies that can be represented with stochastic state controllers (e.g. (Hansen 1998)). When the underlying state dynamics are complex, such policy-based approaches can provide more compact solutions; they are also more amenable to certain types of expert input, such as alarm-thresholds in the healthcare domain.

We are currently testing an approach that uses a Gaussian process to place a distribution over the value of a policy. Symmetries in the policy-encoding mean that many policies produce similar behavior, so we wish to compare policies based on the distributions of histories they produce (rather than directly comparing their parameters). We use importance weighting techniques from (Shelton 2001) to compare how likely a history produced by one policy is to be produced by another policy. By placing an optimistic mean function as the prior on the Gaussian process, we bias the agent to try new policies.

The above approach is principled but computationally expensive. In future work, we hope to leverage similarities between the iPOMDP and the iSC: both build discrete internal state representations with stochastic transitions between nodes that are conditioned on certain inputs (actions for the iPOMDP, observations for the iSC) and produce stochastic outputs (observations for the iPOMDP, actions for the iSC). The key difference is that the iPOMDP overlays a belief over its internal state, while the iSC “fully observes” its internal state. A second thrust of our iSC work to apply inference for the iPOMDP to learn policies instead of models.

**Online structure-discovery.** The first two approaches don’t explicitly consider structure that may present in the environment. We focus on a very specific structure-discovery problem relevant to our healthcare application, where we wish to predict upcoming incidents (hospitalizations, ER visits, etc.) based on a patient’s current history of vital signs, basic statistics, and previous incidents. Here, both the patient’s true condition as well as what features are relevant to predicting his or her true condition are initially unknown. Combinations of the basic inputs can produce very high-dimensional feature vectors.

Our goal is to simultaneously discover which features are relevant for different subpopulations of patients and segment patients into subpopulations based on their what features are relevant for them. As a first step, we propose to use nested combinations of Dirichlet and Indian Buffet processes to model the population segmentations in a batch setting. Our

next goal is to be able to quickly classify a new patient soon after enrollment, so that the approach can be applied to a real healthcare setting.

## Progress

My master’s theses separately studied Bayesian approaches to PORL (Doshi, Pineau, and Roy 2008) and scalable inference in the Indian Buffet Process, a discrete nonparametric Bayesian model. Last year, I began work on the infinite POMDP model (Doshi-Velez 2009), and we have already observed that the iPOMDP approach results in faster learning rates than using EM (tests using different priors are currently underway). I am also in the process of testing an initial implementation of the model-free infinite state controller algorithm on some toy problems. Finally, I obtained the healthcare-related data in February 2010 and am in the process of exploratory data analysis and testing initial feature discovery algorithms. My plan for the rest of the spring is to continue work on the model-free iSC approach and structure-discovery work with the healthcare data. I will also be writing my thesis proposal and forming my thesis committee.

## References

- Beal, M. J.; Ghahramani, Z.; and Rasmussen, C. E. 2002. The infinite hidden Markov model. In *Machine Learning*, 29–245. MIT Press.
- Doshi, F.; Pineau, J.; and Roy, N. 2008. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. In *ICML*, volume 25.
- Doshi-Velez, F. 2009. The infinite partially observable Markov decision process. In *NIPS*.
- Hansen, E. A. 1998. An improved policy iteration algorithm for partially observable MDPs. In *NIPS*, volume 10.
- Hutter, M.; Uther, W.; Poupart, P.; and Ng, K. S. 2009. Partially observable reinforcement learning nips minisymposium.
- Jaulmes, R.; Pineau, J.; and Precup, D. 2005. Learning in non-stationary partially observable Markov decision processes. *ECML Workshop*.
- Poupart, P., and Vlassis, N. 2008. Model-based Bayesian reinforcement learning in partially observable domains. In *ISAIM*.
- Ross, S.; Chaib-draa, B.; and Pineau, J. 2008. Bayes-adaptive POMDPs. In *NIPS*.
- Shelton, C. R. 2001. Policy improvement for POMDPs using normalized importance sampling. *AI Memo 2001-002*, MIT AI Lab.
- Sutton, R. S., and Barto, A. G. 1998. Reinforcement learning: An introduction.
- Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476):1566–1581.
- van Gael, J.; Saatci, Y.; Teh, Y. W.; and Ghahramani, Z. 2008. Beam sampling for the infinite hidden Markov model. In *ICML*, volume 25.