The Infinite Regionalized Policy Representation

Miao Liu, Xuejun Liao, Lawrence Carin MIAO.LIU,XJLIAO,LCARIN@DUKE.EDU
Duke University, Durham, NC 27519, USA

Abstract

We introduce the infinite regionalized policy presentation (iRPR), as a nonparametric policy for reinforcement learning in partially observable Markov decision processes (POMDPs). The iRPR assumes an unbounded set of decision states *a priori*, and infers the number of states to represent the policy given the experiences. We propose algorithms for learning the number of decision states while maintaining a proper balance between exploration and exploitation. Convergence analysis is provided, along with performance evaluations on benchmark problems.

1. Introduction

The policy controlling the behavior of an agent in a partially observable Markov decision process (POMDP) is represented as a mapping from the belief-state space to the action space (Kaelbling et al., 1998). A belief state is the probability distribution over the states of the world the agent interacts with; it is a sufficient statistic of the history of past actions and observations, and summarizes all information necessary to determine the next non-myopic action.

Computation of belief states requires knowledge of the true POMDP model. Therefore, belief states are unobservable to a reinforcement learning (RL) agent, who does not know the true model, but tries to learn the policy based on the experiences. The question then arises as to how to represent the policy and learn it, in the absence of the true model. One approach is to obtain an estimate of the model from the agent's experiences, and then compute an (approximate) optimal policy for the estimated model, with this used as an approximation to the optimal policy for the true model. An alternative is to learn the policy directly from experiences, without an intermediate step of estimating the model. We refer to the former as a model-based approach and the latter as a policy-based approach.

An indispensable ingredient, for both of the aforementioned approaches, is a mechanism for maintaining a proper balance between exploration and exploitation. Until the current policy is optimal, the agent should always explore the consequences of actions that are not encouraged by the current policy, to see whether the new actions will lead to higher expected long-term rewards. Exploration is the only way to ensure continual improvement of the policy. However, excessive exploration makes the policy converge unnecessarily slowly. To keep a balance, the agent needs to switch appropriately between exploration and exploitation.

Model-based approaches usually employ Bayesian RL for an implicit exploration and exploitation trade-off, treating the uncertain POMDP parameters as additional states, and attempting to solve an augmented POMDP with the model uncertainty incorporated into the augmented belief states (Poupart & Vlassis, 2008). However, the augmented POMDP is intractable and approximations are used. The approximations are usually based on policy-solving of model samples, ignoring the model uncertainty in future steps (Doshi-Velez et al., 2009; Doshi-Velez, 2010); as a result, their ability to balance exploration and exploitation is limited.

There has been little work on exploration and exploitation in POMDPs using policy-based approaches. One recent study addressing this problem is reported in (Cai et al.), in which an explicit exploration and exploitation algorithm is given for POMDPs, employing an idea motivated by E^3 (Kearns & Singh, 1998) and R-MAX (Brafman & Tennenholtz, 2002), two RL algorithms in Markov decision processes (MDPs). The method employs a primary policy for choosing the regular actions, and an auxiliary policy for switching between exploration and exploitation. The primary policy is a regionalized policy representation (RPR) (Li et al., 2009). The auxiliary policy is affiliated with the primary one, using the same RPR but with a different set of local policies (see Section 2).

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

The RL algorithm in (Cai et al.) is guaranteed to converge to the optimal policy. However, the optimality is based on the assumption that the RPR has an appropriate number of decision states. In practice, this number is never known and has to be set manually. When this number is too small, the RPR cannot express the optimal policy; when it is too large, the RPR requires an unnecessarily large amount of exploration to converge. Therefore, an appropriate choice of this number is crucial to the success of the algorithm.

In this paper, we introduce the iRPR as a solution to this problem. An iRPR is an extension of the RPR where the parameters are *a priori* drawn from the hierarchical Dirichlet process (HDP) (Teh et al., 2006), which allows an infinite number of decision states and yet is biased towards a small number of states. Given the agent experiences, we infer the number of decision states while maintaining a proper balance between exploration and exploitation. We provide theoretical analysis which guarantees the iPRP converges to the optimal policy as the rate of exploration decreases to zero. Experiments on benchmark problems demonstrate the performance of the iRPR.

2. Regionalized Policy Representation

Definition 1. (Li et al., 2009) A regionalized policy representation is a tuple $(\mathcal{A}, \mathcal{O}, \mathcal{Z}, W, \mu, \pi)$, where $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{Z}$ are respectively a finite set of actions, observations, and decision states; W is a set of Markov transition matrices, with $W^{zao}_{z'}$ denoting the probability of transiting from z to z' when taking action a in z results in observation o; μ is the initial distribution of decision states, with μ_z the probability of initially being in z; π is a set of stochastic policies, with π^z_a the probability of taking action a in z.

For simplicity, $\mathcal{Z}$ is denoted as $\{1, 2, \cdots, |\mathcal{Z}|\}$, where $|\mathcal{Z}|$ is the cardinality, and $\mathcal{A}$ and $\mathcal{O}$ are denoted in similar ways. The set of RPR parameters are denoted as $\Theta = \{\pi, \mu, W\}$. A consecutively indexed variable is abbreviated as the variable with its index range; for example, $a_{0:T} = (a_0, a_1, \ldots, a_T)$, $W^{zao}_{1:|\mathcal{Z}|} = (W^{zao}_1, W^{zao}_2, \cdots, W^{zao}_{|\mathcal{Z}|})$, $\beta_{1:\infty} = (\beta_1, \beta_2, \cdots, \beta_\infty)$, etc.

Given $h_t = \{a_{0:t-1}, o_{1:t}\}$, the history of actions and observations up to t, the RPR chooses action a_t according to

$$p(a_t|h_t, \Theta) = \frac{p(a_{0:t}|o_{1:t}, \Theta)}{p(a_{0:t-1}|o_{1:t}, \Theta)} = \frac{p(a_{0:t}|o_{1:t}, \Theta)}{p(a_{0:t-1}|o_{1:t-1}, \Theta)}, \quad (1)$$

where the second equality arises because o_t has no influence on the actions before t, and $p(a_{0:t}|o_{1:t}, \Theta)$ results from

$$p(a_{0:t}, z_{0:t}|o_{1:t}, \Theta) = \mu_{z_0} \pi^{z_0}_{a_0} \prod_{\tau=1}^{t} W^{z_{\tau-1} a_{\tau-1} o_\tau}_{z_\tau} \pi^{z_\tau}_{a_\tau}, \quad (2)$$

by marginalizing out latent decision states $z_{0:t}$. From equation (1) follows $p(a_{0:t}|o_{1:t}, \Theta) = \prod_{\tau=0}^{t} p(a_\tau|h_\tau, \Theta)$.

The RPR parameters are learned from the agent experiences by using an empirical value function defined below. Assuming the interaction between the POMDP and the agent is episodic (Sutton & Barto, 1998), the experiences are represented as a set of episodes. An episode of length T_k is denoted by $(a^k_0 r^k_0 o^k_1 a^k_1 r^k_1 \cdots o^k_{T_k} a^k_{T_k} r^k_{T_k})$, where r is a nonnegative immediate reward, k indexes the episodes, and the subscripts index discrete time steps.

Definition 2. (Li et al., 2009) Let $\mathcal{D}^{(K)} = \{(a^k_0 r^k_0 o^k_1 a^k_1 r^k_1 \cdots o^k_{T_k} a^k_{T_k} r^k_{T_k})\}_{k=1}^{K}$ be a set of episodes resulting from the interaction between the POMDP and an agent who chooses actions according to Π, an arbitrary stochastic policy with action-selecting distributions $p^\Pi(a|h) > 0$, $\forall$ action a, $\forall$ history h. The *empirical value function* is defined as

$$\widehat{V}(\mathcal{D}^{(K)}; \Theta) \stackrel{def.}{=} \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T_k} \gamma^t r^k_t \frac{\prod_{\tau=0}^{t} p(a^k_\tau|h^k_\tau, \Theta)}{\prod_{\tau=0}^{t} p^\Pi(a^k_\tau|h^k_\tau)} \quad (3)$$

where $h^k_t = (a^k_{0:t-1}, o^k_{1:t})$, $0 < \gamma < 1$ is the discount as defined in the POMDP.

It can be shown that $\lim_{K\to\infty} \widehat{V}(\mathcal{D}^{(K)}; \Theta)$ is the expected sum of discounted rewards by following the RPR parameterized by Θ for an infinite number of steps (Li et al., 2009). Therefore, the RPR resulting from maximization of the empirical value function is an approximation of the optimal policy, assuming the number of decision states, i.e., $|\mathcal{Z}|$, is large enough to accommodate the optimal policy. Sondik (1978) has shown that the optimal policy of any POMDP can be approximated, to arbitrary precision, by a finite state controller (FSC) with a sufficiently large number of internal nodes (the internal nodes correspond to the decision states in an RPR). Meanwhile, Li et al. (2009) have shown that the RPR subsumes the FSC as a special case, when the parameters W and π take particular forms. Therefore, the optimal policy can be approximated by an RPR, and the approximation can be made arbitrarily accurate by using a sufficiently large number of decision states. The optimal number of decision states can be inferred by the method that will be presented in Section 3.

2.1. Bayesian Policy Learning

In addition to value maximization, Li et al. (2009) have also given a Bayesian approach to learning the

RPR, which employs $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ as the likelihood function of Θ given the episodes $\mathcal{D}^{(K)}$. This Bayesian approach is nonstandard, as $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ is not equal to the probability of $\mathcal{D}^{(K)}$ given Θ. But this does not prevent one from obtaining a legitimate posterior of Θ, because, during the posterior inference, one is only interested in $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ as a function of Θ. Since the Bayesian approach forms the basis for the technical developments in Section 3, we provide a review of it below.

With a prior distribution $G_0(\Theta)$, the posterior is defined as

$$p(\Theta|\mathcal{D}^{(K)}) \stackrel{def.}{=} \widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)\,[\widehat{V}(\mathcal{D}^{(K)})]^{-1} \quad (4)$$

where $\widehat{V}(\mathcal{D}^{(K)}) = \int \widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)d\Theta$ is the marginal empirical value. The posterior generally does not have an analytic form. However, by employing the variational Bayesian technique (Beal, 2003), one may obtain an approximation to the posterior, along with the following byproducts: approximations to $p(z^k_{0:t}|a^k_{0:t},o^k_{1:t}),\forall t,k$, and an approximation to the nonnegative sequence $\nu = \{\nu^k_{0:T_k}\}_{k=1}^K$, where $\nu^k_t = \frac{\gamma^t r^k_t p(a^k_{0:t}|o^k_{1:t})}{\prod_{\tau=0}^t p^\Pi(a^k_\tau|h^k_\tau)\widehat{V}(\mathcal{D}^{(K)})}$ is a rescaled discounted reward[1] averaged over the RPRs drawn from G_0 (Li et al., 2009). The rescaling leads to $\frac{1}{K}\sum_{k=1}^K\sum_{t=0}^{T_k}\nu^k_t = 1$, which normalizes $\widehat{V}(\mathcal{D}^{(K)})$ to a unit. Denote by $g(\Theta)$ the approximation to $p(\Theta|\mathcal{D}^{(K)})$, by $\hat{\nu} = \{\hat{\nu}^k_{0:T_k}\}_{k=1}^K$ the approximation to ν, and by $q^k_t(z^k_{0:t})$ the approximation to $p(z^k_{0:t}|a^k_{0:t},o^k_{1:t}),\forall t,k$. Letting $\mathrm{KL}(q\|p)$ denote the Kullback-Leibler (KL) distance between probability measures q and p, the approximations are found by point-wise maximization of

$$\mathrm{LB}(g(\Theta),\hat{\nu},\{q^k_t\}) = \widehat{V}(\mathcal{D}^{(K)}) - \mathrm{KL}(\tfrac{\hat{\nu}}{K}||\tfrac{\nu}{K}) - \frac{1}{K}\sum_{k=1}^K\sum_{t=0}^{T_k}\hat{\nu}^k_t\mathrm{KL}\big(q^k_t(z^k_{0:t})g(\Theta)\big\|p(z^k_{0:t},\Theta|a^k_{0:t},o^k_{1:t})\big), \quad (5)$$

subject to the non-negativity and normalization constraints on $g(\Theta)$, $\frac{\hat{\nu}}{K}$, $\{q^k_t\}$. The first term on the right side, the marginal empirical value, is a constant, thus the maximization is equivalent to minimization of the KL distance between each approximation and the associated true. The joint distribution $p(z^k_{0:t},\Theta|a^k_{0:t},o^k_{1:t})$ is factorized into a product of two marginals, i.e., $q^k_t(z^k_{0:t})g(\Theta)$, in the approximation, and their KL distance is minimized in proportion to the weight $\hat{\nu}^k_t$. Since the weight sequence $\hat{\nu}$ is an approximation to the reward sequence ν, this ensures the approximations associated with higher rewards are more accurate.

[1]The ν^k_t results from the fact one is evaluating the RPR using episodes collected by a different policy Π.

Maximization of (5) leads to analytic solutions when G_0 is a product of Dirichlet distributions,

$$G_0(\Theta) = \Big[\mathrm{Dir}\big(\mu_{1:|\mathcal{Z}|}\big|\upsilon_{1:|\mathcal{Z}|}\big)\Big]\Big[\prod_{i=1}^{|\mathcal{Z}|}\mathrm{Dir}\big(\pi^i_{1:|\mathcal{A}|}\big|\rho^i_{1:|\mathcal{A}|}\big)\Big] \times\Big[\prod_{a=1}^{|\mathcal{A}|}\prod_{o=1}^{|\mathcal{O}|}\prod_{i=1}^{|\mathcal{Z}|}\mathrm{Dir}\big(W^{iao}_{1:|\mathcal{Z}|}\big|\omega^{iao}_{1:|\mathcal{Z}|}\big)\Big], \quad (6)$$

with hyper-parameters (υ,ρ,ω), where $\upsilon = \upsilon_{1:|\mathcal{Z}|}$, $\rho = \{\rho^i_{1:|\mathcal{A}|}\}_{i=1}^{|\mathcal{Z}|}$, and $\omega = \{\omega^{iao}_{1:|\mathcal{Z}|}\}_{i=1:|\mathcal{Z}|,a=1:|\mathcal{A}|,o=1:|\mathcal{O}|}$. The solutions, which are given in (Li et al., 2009), are re-stated in Theorem 3.

Theorem 3. *Let $g(\Theta)$ initially be the form of (6) with hyper-parameters $(\hat{\upsilon},\hat{\rho},\hat{\omega})$, then iterative application of the following updates leads to monotonic increase of (5), until convergence to a maxima. The updates of $\hat{\nu}$ and $\{q^k_t\}$ are given by*

$$\hat{\nu}^k_t = \frac{\gamma^t r^k_t p(a^k_{0:t}|o^k_{1:t},\widetilde{\Theta})}{\prod_{\tau=0}^t p^\Pi(a^k_\tau|h^k_\tau)\widehat{V}(\mathcal{D}^{(K)}|\widetilde{\Theta})},\forall t,k, \quad (7)$$

$$q^k_t(z^k_{0:t}) = p(z^k_{0:t}|a^k_{0:t},o^k_{1:t},\widetilde{\Theta}),\forall t,k, \quad (8)$$

where $\widetilde{\Theta} = \{\widetilde{\pi},\widetilde{\mu},\widetilde{W}\}$ is a set of under-normalized probability mass functions[2], with $\widetilde{\pi}^i_m = e^{\psi(\hat{\rho}^i_m)-\psi(\sum_{a=1}^{|\mathcal{A}|}\hat{\rho}^i_a)}$, $\widetilde{\mu}_i = e^{\psi(\hat{\upsilon}_i)-\psi(\sum_{j=1}^{|\mathcal{Z}|}\hat{\upsilon}_j)}$, and $\widetilde{W}^{iao}_j = e^{\psi(\hat{\omega}^{iao}_j)-\psi(\sum_{z=1}^{|\mathcal{Z}|}\hat{\omega}^{iao}_z)}$, and ψ is the digamma function. The hyper-parameters of $g(\Theta)$ are updated as

$$\begin{aligned}\hat{\upsilon}_i &= \upsilon_i + \tfrac{1}{K}\textstyle\sum_{k=1}^K\sum_{t=0}^{T_k}\hat{\nu}^k_t\phi^k_{t,0}(i)\\ \hat{\rho}^i_a &= \rho^i_a + \tfrac{1}{K}\textstyle\sum_{k=1}^K\sum_{t=0}^{T_k}\hat{\nu}^k_t\sum_{\tau=0}^t\phi^k_{t,\tau}(i)\delta(a^k_\tau,a) \quad (9)\\ \hat{\omega}^{iao}_j &= \omega^{iao}_j + \tfrac{1}{K}\textstyle\sum_{k=1}^K\sum_{t=0}^{T_k}\hat{\nu}^k_t\sum_{\tau=1}^t\xi^k_{t,\tau-1}(i,j)\\ &\quad\times\delta(a^k_{\tau-1},a)\delta(o^k_\tau,o),\end{aligned}$$

where

$$\xi^k_{t,\tau}(i,j) = p(z^k_\tau = i, z^k_{\tau+1} = j|a^k_{0:t},o^k_{1:t},\widetilde{\Theta}) \quad (10)$$

$$\phi^k_{t,\tau}(i) = p(z^k_\tau = i|a^k_{0:t},o^k_{1:t},\widetilde{\Theta}) \quad (11)$$

are marginals of $q^k_t(z^k_{0:t})$.

2.2. Exploration-Exploitation Trade-off

The method in (Cai et al.) employs an auxiliary policy to decide between exploration and exploitation at any time during an episode, and the decision is conditional on the history of past actions and observations. The auxiliary policy, which also is an RPR, is affiliated with the primary RPR that controls the regular actions. The auxiliary RPR has parameters (σ,μ,W), where

[2]Note that $q^k_t(z^k_{0:t}) = p(z^k_{0:t}|a^k_{0:t},o^k_{1:t},\widetilde{\Theta})$ is always properly normalized by $p(a^k_{0:t}|o^k_{1:t},\widetilde{\Theta}) = \sum_{z_0,\cdots,z_t=1}^{|\mathcal{Z}|}p(z^k_{0:t},a^k_{0:t}|o^k_{1:t},\widetilde{\Theta})$.

(μ, W) are shared with the primary RPR, and σ is distinct from π, with σ^z_y denoting the probability of choosing exploration ($y = 1$) or exploitation ($y = 0$) in decision state z. The primary RPR is learned as discussed above. The auxiliary RPR only updates σ, using (μ, W) as they are learned for the primary RPR. The σ is governed by a set of beta distributions,

$$\sigma^z_0 \sim \text{Beta}(u^z_0, u_1), \text{ with } \sigma^z_1 = 1 - \sigma^z_0, \ \forall\, z \in \mathcal{Z}, \quad (12)$$

where $u_1 > 1$ is a given constant and $\{u^z_0\}_{z=1}^{|\mathcal{Z}|}$ are updated using the rule,

$$u^i_0 = \sum_{k=1}^{K} \sum_{t=0}^{T_k} \hat{\nu}^k_t \sum_{\tau=0}^{t} \phi^k_{t,\tau}(i), \ \forall\, i \in \mathcal{Z}, \quad (13)$$

where $\hat{\nu}^k_t$ and $\phi^k_{t,\tau}(i)$ are as given in (7) and (11), respectively. At time t during an episode, the probability $p(y_t|h_t)$ is computed using (1) and (2), replacing a_t with y_t and π with σ. The update in (13) is performed concurrently with (9), upon completion of an episode.

Intuitively, u^z_0 represents the total amount of immediate and future rewards[3] (over all time steps in previous episodes) that the agent receives when executing $\pi^z_{1:|\mathcal{A}|}$, the local policy in decision state z. Since $\pi^z_{1:|\mathcal{A}|}$ is executed only when z is occupied, the reward at t is allocated to u^z_0 in proportion to $\sum_{\tau=0}^{t} \phi^k_{t,\tau}(z)$, where $\phi^k_{t,\tau}(z)$ is the probability that z is occupied at τ in episode k, given the actions and observations that have led to the reward at t, as one recalls from (11).

When $u^z_0 \gg u_1$, one has $\sigma^z_0 \gg \sigma^z_1$, which implies the agent almost never performs exploration. Therefore, u_1 defines, up to a constant multiplier, the total reward required in z for the agent to stop exploration in z.

As rewards accumulate in each decision state, the probability of exploration gradually decreases. It is shown in (Cai et al.) that, with a sufficiently large u_1, the RPR is guaranteed to converge to the optimal policy (assuming $|\mathcal{Z}|$ is appropriate).

3. The Infinite RPR

Definition 4. The infinite regionalized policy representation (iRPR) is a tuple $(\mathcal{A}, \mathcal{O}, \mathcal{Z}, \lambda, \alpha, \rho)$, where $\mathcal{A}$ and $\mathcal{O}$ are as in Definition 1; $\mathcal{Z}$ is an *unbounded* set of decision states indexed by positive integers; (λ, α, ρ) determine (W, μ, π), the RPR parameters in Definition 1, as follows (notations described under Definition 1),

$$\begin{aligned} \eta_i &\sim \text{Beta}(1, \lambda), \ i = 1, 2, \cdots, \infty, \\ \beta_i &= \eta_i \textstyle\prod_{j=1}^{i-1}(1 - \eta_j), \ i = 1, 2, \cdots, \infty, \\ W^{iao}_{1:\infty} &\sim \text{DP}(\alpha, \beta_{1:\infty}), \ i = 1, 2, \cdots, \infty, \\ \mu_{1:\infty} &\sim \text{DP}(\alpha, \beta_{1:\infty}), \\ \pi^i_{1:|\mathcal{A}|} &\sim \text{Dir}(\cdot\,|\,\rho_{1:|\mathcal{A}|}), \ i = 1, 2, \cdots, \infty, \end{aligned}$$

$\forall\, a \in \mathcal{A}$ and $\forall\, o \in \mathcal{O}$, where $\text{DP}(\alpha, \beta_{1:\infty})$ denotes a Dirichlet process (Ferguson, 1973) with concentration α and base probability measure $\beta_{1:\infty}$.

The iRPR is defined based on the hierarchical Dirichlet process (HDP) (Teh et al., 2006), which places a nonparametric prior on $\Theta = (W, \mu, \pi)$. The stick-breaking weights $\beta_{1:\infty}$ specify a probability measure on $\mathcal{Z} = \{1, 2, \cdots, \infty\}$. The initial state distribution (μ) and the next-state distributions (W) are independently drawn from $\text{DP}(\alpha, \beta_{1:\infty})$. The local policies (π) are drawn independently from a Dirichlet distribution with parameters $\rho_{1:|\mathcal{A}|}$.

Assuming $\rho_a = 1/|\mathcal{A}|$, $\forall\, a \in \mathcal{A}$ (i.e., the agent takes random actions *a priori*), we are interested in learning the RPR parameters Θ, along with the hyper-parameters (λ, α) if inference of the latter is desirable. The learning is based on Gibbs sampling, which iteratively samples the decision-state occupancies $\{z^k_t\}$, the RPR parameters Θ, the latent parameters $\beta_{1:\infty}$, and the hyper-parameters (λ, α) (if desirable), with the samples of one group of parameters conditional on all other parameters.

During the inference, one may employ Theorem 3 as a basic ingredient. To see why this is possible, we first note that the nonparametric prior can be expressed using the parametric prior in (6) by introducing a special state z^* summarizing all decision states currently unoccupied by the episodes. Given that the number of (distinct) unoccupied decision states is n, one may write a parametric prior as

$$\begin{aligned} G^n_0(\Theta) = &\left[\text{Dir}\left(\mu_{1:n+1}\middle|\alpha\beta_{1:n+1}\right)\right]\left[\textstyle\prod_{i=1}^{n+1}\text{Dir}\left(\pi^i_{1:|\mathcal{A}|}\middle|\rho_{1:|\mathcal{A}|}\right)\right] \\ &\times\left[\textstyle\prod_{a=1}^{|\mathcal{A}|}\prod_{o=1}^{|\mathcal{O}|}\prod_{i=1}^{n+1}\text{Dir}\left(W^{iao}_{1:n+1}\middle|\alpha\beta_{1:n+1}\right)\right], \quad (14) \end{aligned}$$

where z^* is indicated by $n + 1$, $\{\beta_i\}_{i=1}^{n}$ are the same as in definition 4 and $\beta_{n+1} = 1 - \sum_{i=1}^{n} \beta_i$.

Given $G^n_0(\Theta)$, one may employ Theorem 3 to obtain $\hat{\nu}$, $\{q^k_t(z^k_{0:t})\}_{\forall\, t,k}$, and $g(\Theta)$, which are the approximations to the rescaled rewards ν, the decision-state posterior $\{p(z^k_{0:t}|a^k_{0:t}, o^k_{1:t})\}_{\forall\, t,k}$, and the RPR posterior $p(\Theta|\mathcal{D}^{(K)})$, respectively. Given $\{q^k_t(z^k_{0:t})\}_{\forall\, t,k}$, one can obtain decision-state occupancies $\{z^k_{t,0:t}\}_{t=0:T_k, k=1:K}$, where $z^k_{t,0:t} \sim q^k_t(z^k_{0:t})$ with $q^k_t(z^k_{0:t})$ given in (8). Let $\mathbb{I}_s$ be an indictor function that equals one if s is true and zero otherwise. Define

$$\varphi^{iao}_j = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\hat{\nu}^k_t \sum_{\tau=1}^{t} \mathbb{I}_{z^k_{t,\tau-1}=i, a^k_{\tau-1}=a, o^k_\tau=o, z^k_{t,\tau}=j}$$

[3] Recall that $\hat{\nu}^k_t$ approximates a rescaled reward received by following the RPR.

which is the reward-weighted sum of transitions from i to j given that action a results in observing o.

If $\sum_{i,a,o} \varphi_{n+1}^{iao} = 0$, there is currently no occupancy at z^*, then β is updated as

$$\beta_{1:n+1} \sim \text{Dir}(\beta_{1:n+1} | \textstyle\sum_{i,a,o} m_1^{iao}, \cdots, \sum_{i,a,o} m_n^{iao}, \lambda)$$

with $\{m_j^{iao}\}$ a set of auxiliary variables sampled from

$$p(m_j^{iao}{=}m | z, \beta, \alpha) \propto S(\lceil \varphi_j^{iao} \rceil, m)(\alpha\beta_j)^m,\ m{=}1, \cdots, n,$$

where $S(\cdot,\cdot)$ is a Stirling number of the first kind (Teh et al., 2006); $\lceil x \rceil$ is the smallest integer no less than x.

If $\sum_{i,a,o} \varphi_{n+1}^{iao} > 0$, it indicates there is at least one occupancy at z^*, then one generates a new decision state to hold the occupancy, releasing z^* as a special state. Assuming $n+1$ and $n+2$ respectively indicates the new decision state and z^*, one first samples $\{m_{1:n+1}^{iao}\}$ and then samples $\beta_{1:n+2}$, similarly as above.

Given the samples of $\{\varphi^{iao}\}$, $\{m^{iao}\}$, and β, one may sample the concentration parameters (α, λ), assuming they have gamma priors. The details are similar to those as described in the appendix of (Teh et al., 2006).

So far, one has completed a single iteration of Gibbs sampling. Given the update of (α, β), the prior in (14) is updated, and one begins a new iteration. The process is repeated until the Gibbs sampler converges.

4. Exploration vs Exploitation in iRPR

We extend the exploration-exploitation method in Section 2.2 to account for the unbounded set of decision states. Definition 2 requires $p^\Pi(a|h) > 0$, $\forall$ action a and history h, so that the empirical value function converges to the true value function as the episodes grow. We consider Π as a mixture of the iRPR and the uniformly random policy, i.e., $\forall\, h$, we let $p^\Pi(a|h) = p(y=0|h)p(a|h,\Theta) + p(y=1|h)/|\mathcal{A}|$. The mixing proportions, $p(y|h) = \sum_{z\in\mathcal{Z}} \sigma_y^z p(z|h)$, are computed using the auxiliary policy in Section 2.2, with the special state z^* included in $\mathcal{Z}$ to represent any potential unseen experiences.

Before the $(K+1)$-th episode starts, the agent has learned Θ based on the previous episodes $\mathcal{D}^{(K)}$ and the prior in (14), where the special state $z^* = n+1$ is currently unoccupied and reserved to hold possible occupancies in future episodes.

Moreover, the agent has used (13) to allocate each previous reward to u_0^z in proportion to the probabilities that z is occupied at and *before* the time of receiving the reward. A large u_0^z, therefore, implies either or both of the following events: (i) the local policy $\pi_{1:|\mathcal{A}|}^z$ has contributed to large immediate or *future* rewards; (ii) there has been a large amount of visits to z leading to small rewards. Note that $u_0^{z^*} \equiv 0$, since z^* is not occupied in $\mathcal{D}^{(K)}$.

Since $p(y=0|h) \gg p(y=1|h)$ implies that $u_0^z \gg u_1$ for any $z \in \{z : p(z|h) \gg 0\}$, which in turn implies that the decision states closely associated with h have received a large amount (relative to u_1) of rewards and/or visits so far. As a result, when u_1 in (12) is sufficiently large, one can conclude

$$\mathcal{H}_{\text{known}} \overset{Def.}{=} \{h : p(y=0|h) \gg p(y=1|h)\} \tag{15}$$

represents the part of $\mathcal{H}$ in which the agent has acquired so much information, in the form of either a few large rewards or a large number of small rewards, that the iRPR policy it has learned is nearly optimal there, where $\mathcal{H}$ is the set of all possible histories. Let $u_1^{\min}$ denote the minimum u_1 such that this is true.

Define $\mathcal{H}_{\text{unseen}} = \{h : p(z = z^*|h) \gg 0\}$, which is the part of $\mathcal{H}$ that can not be represented by current decision states. Note that $h \in \mathcal{H}_{\text{known}}$ implies $h \notin \mathcal{H}_{\text{unseen}}$, because $h \in \mathcal{H}_{\text{unseen}}$ contradicts $p(y=0|h) \gg p(y=1|h)$, using the fact that $u_0^{z^*} \equiv 0$. Thus $\mathcal{H}_{\text{unseen}} \subset \mathcal{H}_{\text{unknown}} = (\mathcal{H} \setminus \mathcal{H}_{\text{known}})$. Letting $\mathcal{H}_{\text{seen}} = \mathcal{H}_{\text{unknown}} \setminus \mathcal{H}_{\text{unseen}}$, one may write $\mathcal{H} = \mathcal{H}_{\text{known}} \cup \mathcal{H}_{\text{seen}} \cup \mathcal{H}_{\text{unseen}}$, with the subsets mutually exclusive.

During the $(K+1)$-th episode, the agent follows the iRPR policy in $\mathcal{H}_{\text{known}}$ to obtain high rewards (exploitation), takes random actions to increase knowledge in $\mathcal{H}_{\text{seen}}$ or start new knowledge in $\mathcal{H}_{\text{unseen}}$ (both are exploration). Upon completion of the episode, $\{u_{0,1}^z\}_{z=1}^{\mathcal{Z}}$ are updated to reflect the increased knowledge, with a new decision state introduced to hold the new knowledge if $\mathcal{H}_{\text{unseen}}$ has been visited.

The iRPR always maintains the minimum $\mathcal{Z}$ for any given episodes $\mathcal{D}^{(K)}$, which is a key difference from the RPR apart from inferring $\mathcal{Z}$. Identification of $\mathcal{H}_{\text{unseen}}$ leads to incremental augmentation of $\mathcal{Z}$, and all new decision states (including z^*) are initially activated for constant exploration, recalling from Section 3 that $\rho_z = 1/|\mathcal{Z}|, \forall\, z \in \mathcal{Z}$.

In contrast, the RPR assumes the optimal policy is representable on a fixed $\mathcal{Z}$. When $|\mathcal{Z}|$ is underspecified, an apparently large $p(y=0|h)$ does not necessarily imply h is known, because the RPR cannot identify $\mathcal{H}_{\text{unseen}}$. Thus, even if $u_1^z \gg u_0$, $\forall\, z \in \mathcal{Z}$, the RPR cannot claim $\mathcal{H}_{\text{known}}$ as known; instead, the RPR will converge to a suboptimal policy in this case. It is clear then, the above definition for $\mathcal{H}_{\text{known}}$ is correct for the RPR only when $|\mathcal{Z}|$ is appropriate, while it

is correct for the iRPR even if $|\mathcal{Z}|$ is initially underspecified.

It is important to note that the agent does not have to see every history in $\mathcal{H}$, nor try all actions, to obtain a good policy. Recall an RPR policy is a mapping from $\mathcal{H}$ to $\mathcal{A}$. Given that any $h \in \mathcal{H}$ corresponds to a belief state in the underlying POMDP, one may define the similarity between two histories based on the similarity between the corresponding belief states. As similar belief states are likely to have the same optimal action (Sondik, 1978), so are similar histories. Therefore, the agent needs to see typical histories only, instead of every single history in $\mathcal{H}$.

When no new decision state emerges, it means $\mathcal{H}_{\text{unseen}}$ is null; when all existing decision states have their u_0's significantly exceeding u_1, it means $\mathcal{H}_{\text{seen}}$ is null. When both occur, one has $\mathcal{H} = \mathcal{H}_{\text{unseen}}$ and the learning stops.

Thus, we have provided an approach to balancing exploration and exploitation while at the same time inferring the number of decision states. A formal convergence analysis of the approach is given below.

4.1. Optimality and Convergence Analysis

The following theorem guarantees that an agent following the Π specified above will continue exploration until the iRPR has converged to the optimal policy. Moreover, the theorem quantitatively relates the exploration rate to the difference between the optimal value and the value of the current iRPR. The theorem extends the analysis in (Cai et al.) to account for the unbounded $\mathcal{Z}$. The proof is in the Appendix.

Let $\mathcal{M}$ denote the true model of the POMDP. Then

$$V(\mathcal{M};\Theta) = \sum_{t=0}^{\infty} \sum_{a_{0:t},o_{1:t},r_t} \gamma^t r_t \, p(a_{0:t}, o_{1:t}, r_t|\Theta, \mathcal{M}), \quad (16)$$

is the true value function of Θ, where $r_t = 0, \forall\, t > T$, for an episode of length T.[4]. Let $R_{\max}$ denote the maximum r. Since $r \geq 0$, as one recalls from Section 2, one must have $R_{\max} > 0$ (otherwise $r \equiv 0$).

Theorem 5. *Let Θ^* be the optimal iRPR for the underlying POMDP. Let Θ be the iRPR learned from $\mathcal{D}^{(K)}$, and σ be governed by (12), with $u_1 \geq u_1^{\min}$ and $\{u_0^z\}_{z=1}^{|\mathcal{Z}|}$ updated as in (13). For any $\epsilon \geq 0$, if $V(\mathcal{M};\Theta) < V(\mathcal{M};\Theta^*) - \epsilon$, then*

$$P_e = 1 - p(y_{0:\infty} = 0|\sigma, \Theta) > (1-\gamma)\epsilon/R_{\max}. \quad (17)$$

where P_e denotes the probability of exploration, and $y_{0:t} = 0$ is a shorthand for "$y_\tau = 0, \forall\, \tau \in [0, t]$".

[4] After an episode terminates, the agent stays in an absorbing state with zero reward (Sutton & Barto, 1998).

Theorem 5 shows that, when the value of the current iRPR is ϵ away from the optimal value, the agent will perform exploration with probability $P_e > (1-\gamma)\epsilon/R_{\max}$. Conversely, when $P_e \leq (1-\gamma)\epsilon/R_{\max}$, the value of the current iRPR is guaranteed to be ϵ close to the optimal value.

Given a history h, the agent may explore in either of the two cases: (i) $z^* \in \mathcal{Z}(h) = \{z : p(z|h) \gg 0\}$, (ii) $\mathcal{Z}(h)$ contains an occupied decision state z for which $u_0^z \gg u_1$ is false. In case (i), $h \in \mathcal{H}_{\text{unseen}}$, the agent explores to start the learning in h, while in case (ii), $h \in \mathcal{H}_{\text{seen}}$, the agent resumes the learning in h. When nether (i) nor (ii) occurs, the iRPR is optimal with the minimum necessary number of decision states.

5. Results

We study the empirical performance of the iRPR based on three benchmark POMDP models, i.e, Littman's noisy 1D maze and Hallway, and Tag. The models are available at `http://www.cs.brown.edu/research/ai/pomdp/examples/index.html` and `http://www.science.uva.nl/~mtjspaan/pomdp`.

In all the experiments, we assume gamma priors for the HDP concentration parameters, i.e., $\alpha \sim \text{Ga}(10, 10)$ and $\lambda \sim \text{Ga}(3, 10)$, where $\text{Ga}(a_g, b_g)$ is a gamma distribution with scale a_g and shape b_g. Upon completion of each episode, the iRPR parameters are updated using the inference algorithm presented in Section 3, based on all available episodes. All results shown result from an average over ten independent Monte Carlo runs, with error bars showing the variances.

5.1. Effects of History Information and u_1

We first examine the effects of u_1 and history information on the performance of balancing exploitation and exploration. Figure 1 plots the following experimental outputs as a function of $\log(k)$: (i) the cumulative discounted reward averaged over episodes 1 through k, with the optimal reward subtracted, (ii) the average exploration rate in the k-th episode, i.e., $\frac{1}{T_k+1}\sum_{t=0}^{T_k} p(y_t^k|h_t^k)$, (iii) the number of decision states $|\mathcal{Z}|$ learned from episodes 1 through k. For a curve labeled with u_1 only, the exploration or exploitation is determined by $y_t^k \sim p(y_t^k|h_t^k)$, using the u_1 shown. For a curve labeled with u_1 and P_e, the agent draws y_t^k from a Bernoulli distribution with $p(y_t^k = 1) = P_e$, where $P_e = \frac{1}{\sum_{i=1}^k T_i+1}\sum_{i=1}^k \sum_{t=0}^{T_i} p(y_t^k|h_t^k)$ is the average exploration rate for the curve that is labeled with the corresponding u_1 only.

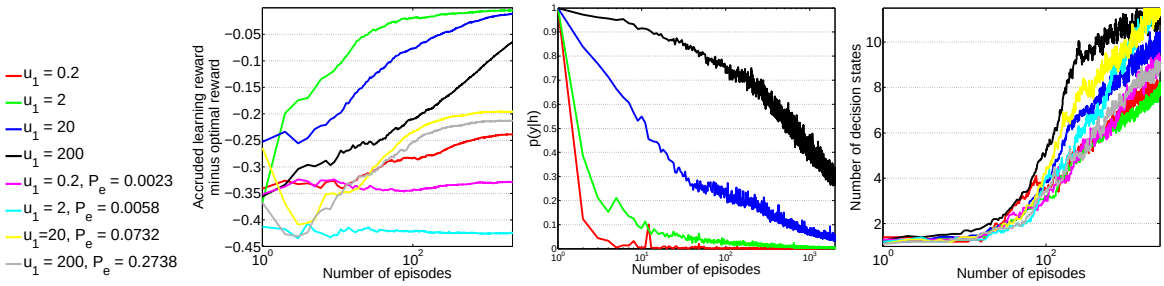

Figure 1. The iRPR's performance on the noisy 1D maze: (left) relative reward (middle) exploration rate (right) $|\mathcal{Z}|$.

It is seen from Figure 1 that, when $u_1 = 2$, the iRPR converges to optimality after 1000 learning episodes, and the exploration rate drops to zero accordingly. This indicates the amount of exploration allowed by u_1 is appropriate. When $u_1 = 20$, the iRPR converges to optimality, but at a lower convergence rate. This indicates that the amount of exploration as specified by $u_1 = 20$ is excessively large. With $u_1 = 200$, the convergence is too slow to be seen within the number of episodes shown here. When $u_1 = 0.2$, the exploration rate quickly drops to zero, without giving the agent enough time for exploration, and as a result, the iRPR only converges to a suboptimal policy. The results show relations between exploration rates and values (accumulative discounted rewards) that are in agreement with Theorem 5.

The performances significantly degrade when using fixed exploration rate (not considering history information), demonstrating that the use of history information is crucial to balancing exploitation and exploration. The number of decision states inferred by the iRPR generally increases with the number of episodes, and with the increase of exploration rates.

5.2. Performance Comparisons

We compare the performance of the iPRP to those of the RPR and the iPOMDP (Doshi-Velez et al., 2009), the latter is a nonparametric model-based approach that infers the number of world states of a POMDP.[5]

We report the results on Hallway and Tag, in the form of the un-discounted reward summed over interactions. The results of iPOMDP and EM (which is the finite counterpart of iPOMDP), which are cited from (Doshi-Velez, 2010), are available only within a small portion of the interactions shown here. It is seen from Figure 2 that the iPRR performs much better than the iPOMDP and EM.

The RPR has its performance dependent on the number of decision states. The iRPR always achieves superior performance by using appropriate numbers of decision states. The advantage of iRPR is more prominent as the agent has accumulated more experiences to make the inference of $|Z|$ more accurate.

[5]The iPOMDP was employed in (Doshi-Velez, 2010) to implement an infinite FSC (iFSC), which can be regarded as a special case of the iRPR (refer to the text under Definition 2 for the relations between an iRPR and a FSC). However, the iFSC was learned by fitting to expert trajectories (using standard likelihood functions), while learning of the iRPR uses the empirical value function in (3). As shown in (Li et al., 2009), learning with the empirical value function is closely related to policy iteration (Sondik, 1978). The iFSC and the iRPR are based on totally different learning frameworks, which should not be confused.

6. Conclusions

We have extended the RPR to represent the POMDP policy on an *a priori* unbounded set of decision states. The resulting iRPR infers the *a posteriori* number of decision states, to match policy complexity dynamically to the experiences. We have given an approach to balancing exploration and exploitation while inferring the decision states. Convergence analysis guarantees that the iRPR performs exploration with a rate commensurate with the difference from the optimal value. Experimental results agree with the theoretical analysis and demonstrate the iRPR's superior performance over those of the RPR and the iPOMDP.

Acknowledgments

The research reported in this paper was supported by AFOSR, ARO, DOE, ONR and NGA.

Appendix

Proof of Theorem 5: We note that $V(\mathcal{M};\Theta^*)$ is upper bounded by

$$V_f(\mathcal{M};\sigma,\Theta)=\sum_{t=0}^{\infty}\sum_{a_{0:t},o_{1:t},r_t}\gamma^t r_t p(a_{0:t},o_{1:t},r_t,y_{0:t}=0|\sigma,\Theta,\mathcal{M})$$
$$+\sum_{t=0}^{\infty}\gamma^t R_{\max}\sum_{a_{0:t},o_{1:t},r_t}\sum_{y_{0:t}\neq 0}p(a_{0:t},o_{1:t},r_t,y_{0:t}|\sigma,\Theta,\mathcal{M}),\quad(18)$$

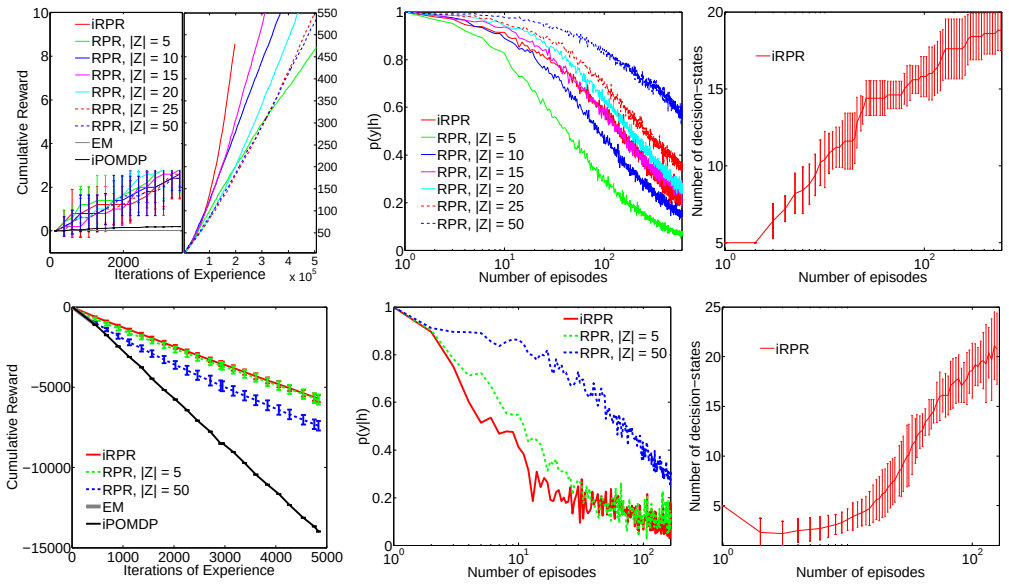

Figure 2. Performance comparison between iRPR, RPR, iPOMDP (top) Hallway (bottom) Tag

where $y_{0:t} = 0$ is an abbreviation for "$y_\tau = 0, \forall \tau \in [0, t]$" and $y_{0:t} \neq 0$ for "$y_\tau \neq 0, \exists \tau \in [0, t]$".

To verify the upper bound, one notes that V_f is constructed as an optimistic value function, in which the agent receives $R_{\max}$ at any time t unless $y_{0:t} = 0$. However, observing $y_{0:t} = 0$ implies $\{h_\tau : \tau \in [0, 1]\} \subset \mathcal{H}_{\text{known}}$, in which Θ is optimal (see (15) and the discussions thereabout). Note that the probability of observing $y_{0:t} = 0$, i.e., $p(y_{0:t} = 0)$ can be small, which means the first term in V_f may also be small.

The premise implies $\epsilon < V_f(\mathcal{E}; \Theta) - V(\mathcal{E}; \Theta)$. Substituting (16), (18), the equation $p(a_{0:t}, o_{1:t}, r_t|\Theta, \mathcal{M}) = \sum_{y_{0:t}} p(a_{0:t}, o_{1:t}, r_t, y_{0:t}|\sigma, \Theta, \mathcal{M})$, and integrating out a's and o's, one obtains

$$\begin{aligned}
\epsilon &< \textstyle\sum_{t=0}^{\infty} \sum_{r_t} \gamma^t (R_{\max} - r_t) \sum_{y_{0:t} \neq 0} p(r_t, y_{0:t}|\Theta, \sigma), \\
&< \textstyle\sum_{t=0}^{\infty} \sum_{r_t} \gamma^t R_{\max} \sum_{y_{0:t} \neq 0} p(r_t, y_{0:t}|\Theta, \sigma), \\
&= \textstyle\sum_{t=0}^{\infty} \gamma^t R_{\max} \sum_{y_{0:t} \neq 0} p(y_{0:t}|\Theta, \sigma), \\
&= \textstyle\sum_{t=0}^{\infty} \gamma^t R_{\max} (1 - p(y_{0:t} = 0|\Theta, \sigma)) \\
&< \textstyle\sum_{t=0}^{\infty} \gamma^t R_{\max} (1 - p(y_{0:\infty} = 0|\Theta, \sigma)) \\
&= \frac{R_{\max}}{1 - \gamma} (1 - p(y_{0:\infty} = 0|\Theta, \sigma)),
\end{aligned}$$

from which (17) follows. □

References

Beal, M. J. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, Univertisity College London, 2003.

Brafman, R. I. and Tennenholtz, M. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(OCT):213–231, 2002.

Cai, C., Liao, X., and Carin, L. Learning to explore and exploit in pomdps. In *Advances in Neural Information Processing Systems 22*.

Doshi-Velez, F. Nonparametric bayesian policy priors for reinforcement learning. In *Advances in Neural Information Processing Systems 23*. 2010.

Doshi-Velez, F., Wingate, D., Roy, N., and Tenenbaum, J. The infinite partially observable markov decision process. In *Advances in Neural Information Processing Systems 22*, pp. 477–485, 2009.

Ferguson, T. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

Kaelbling, L., Littman, M., and Cassandra, A. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

Kearns, M. and Singh, S. P. Near-optimal performance for reinforcement learning in polynomial time. In *Proc. ICML*, pp. 260–268, 1998.

Li, H., Liao, X., and Carin, L. Multi-task reinforcement learning in partially observable stochastic environments. *Journal of Machine Learning Research*, 10:1131–1186, 2009.

Poupart, P. and Vlassis, N. Model-based Bayesian reinforcement learning in partially observable domains. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

Sondik, E. J. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978.

Sutton, R. and Barto, A. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.