Factorized Asymptotic Bayesian Policy Search for POMDPs

Masaaki Imaizumi
Institute of Statistical Mathematics
insou11@hotmail.com

Ryohei Fujimaki
NEC Corporation
rfujimaki@nec-labs.com

Abstract

This paper proposes a novel direct policy search (DPS) method with model selection for partially observed Markov decision processes (POMDPs). DPSs have been standard for learning POMDPs due to their computational efficiency and natural ability to maximize total rewards. An important open challenge for the best use of DPS methods is model selection, i.e., determination of the proper dimensionality of hidden states and complexity of policy functions, to mitigate overfitting in highly-flexible model representations of POMDPs. This paper bridges Bayesian inference and reward maximization and derives marginalized weighted log-likelihood (MWL) for POMDPs which takes both advantages of Bayesian model selection and DPS. Then we propose factorized asymptotic Bayesian policy search (FABPS) to explore the model and the policy which maximizes MWL by expanding recently-developed factorized asymptotic Bayesian inference. Experimental results show that FABPS outperforms state-of-the-art model selection methods for POMDPs, with respect both to model selection and to expected total rewards.

1 Introduction

Partially observed Markov decision processes (POMDPs) ([Åström, 1965]) are powerful and successful probabilistic models in reinforcement learning. By introducing hidden states into Markov decision processes (MDPs) ([Howard, 1960],[Puterman, 2014]), POMDPs achieve flexible but compact representations that lead to more accurate policy, better mitigation against the curse of dimensionality, and higher interpretability of the learned models than those of MDPs. With their powerful representation ability, POMDPs have been successfully applied in such various applications as controlling robots ([Capitan *et al.*, 2013],[Spaan *et al.*, 2010]), audio processing ([Young *et al.*, 2010]), business marketing ([Irissappane *et al.*, 2014]), and medical services ([Debbi *et al.*, 2013]).

For learning POMDP parameters, the use of directed policy searches (DPSs) is especially promising ([Konda and Tsitsiklis, 2000],[Peters and Schaal, 2007],[Cai *et al.*, 2009], [Deisenroth and Rasmussen, 2011] [Levine and Koltun, 2013],[Busa-Fekete *et al.*, 2014]). They directly search for the best policy in a policy space, while other frameworks, such as value-iteration or temporal difference learning, do their explorations in a value space. Also, DPSs have an advantage in computational efficiency over the others since they model policies directly, without the need to model how the hidden factor behaves, whose complexity would rapidly increase with the dimensionality of hidden states.

An important open problem for the best use of POMDPs with DPSs is model selection, i.e., determination of the proper dimensionality of hidden states and complexity of policy functions. Because of their highly flexible model representations, POMDPs are likely to be over-fitted to data if one overestimate the complexity of models. Due to the *singularity* of POMDPs (for the singularity of statistical models, see [Watanabe, 2009]), such classical approaches as Bayesian information criterion (BIC) ([Schwarz, 1978]) do not work, such methods as cross-validation are computationally expensive. As for a non-DPS framework, [Doshi-Velez *et al.*, 2015] have proposed a value function based method with model selection by taking advantage of recent progress in Bayesian nonparametrics. However, as far as we know, there exists no principled method which addresses the model selection issue in POMDPs with DPSs.

This paper proposes a Bayesian model selection algorithm for POMDPs with the DPSs. Our key contributions are summarized as follows. We first define marginalized weighted log-likelihood (MWL) and its asymptotic approximation, weighted factorized information criterion (wFIC), as our model selection criterion for POMDPs by extending that for MDPs ([Ueno *et al.*, 2012]), which can conduct Bayesian inference (model selection) and reward maximization simultaneously. The simultaneous approach can avoid a problem of dependency between the policy and other parameters. The maximizer of the MWL function is proven to converge to the maximizer of the reward function. Also, we propose an EM-like alternating inference algorithm which we refer to as *factorized asymptotic Bayesian policy search* (FABPS) by extending recently-developed *factorized asymptotic Bayesian* (FAB) inference ([Fujimaki and Morinaga, 2012], [Fujimaki and Hayashi, 2012], [Hayashi and Fujimaki, 2013], [Eto *et al.*, 2014], [Liu *et al.*, 2015], [Hayashi *et al.*, 2015]). By taking advantage of the FAB hidden state selection mechanism, FABPS simultaneously determines both the dimensionality of hidden states and the complexity of the policy functions through a single run of EM alternating optimization, and it finds, as well, the parameter that maximizes expected total rewards. Our experiments, on simulation and helicopter data, show that FABPS outperforms state-of-the-art POMDP model selection methods about both model selection and total rewards. Because of the space limitation, we omit all proofs of theorems, information on optimal parameters, and detailed derivation processes of wFIC and FABPS algorithms.

2 Preliminary

2.1 Notation for POMDPs

A POMDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, r, p, p_o)$, where $\mathcal{S}$ is a discrete state space, $\mathcal{A}$ is an action space, and $\mathcal{O}$ is an observation space. Let $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $o \in \mathcal{O}$ denote a state, an action, and an observation, respectively. A POMDP considers a discrete state space with finite K elements, i.e., $\mathcal{S} = (s_1, \dots, s_K)$ and $|\mathcal{S}| = K$. Both the action space and the observation space can be either continuous or discrete. A transition function $p : \mathcal{A} \times \mathcal{S} \times \Lambda \times \mathcal{S} \to [0, 1]$ is denoted by $p(s'|a, s, \lambda) := \Pr(s'|s, a, \lambda)$ where Λ is a transition parameter space and λ is its element, i.e., $\lambda \in \Lambda$. Let λ be decomposed into $\{\lambda_k\}_k$ and λ_k effects the transition to the k-th state. An observation function $p_o : \mathcal{S} \times \mathcal{O} \to [0, 1]$ is denoted by $p_o(o|s)$ with a given s. Let M be model information for characterizing POMDPs, and $\mathcal{M}$ be a set of possible models. For example, M contains K which is the dimensionality of the state space for POMDPs and $\mathcal{M} = \{1, 2, 3, \dots\}$.

To describe how agents of POMDPs behave, a belief distribution $b : \mathcal{S} \to [0, 1]$ is introduced. Let $\mathcal{B}$ be a set of b. Since the state is partially observable, the agent possesses the belief distribution and updates it in each period. POMDP agents determine its action on the basis of its state and a policy function $\pi : \mathcal{B} \times \mathcal{A} \times \Theta \to [0, 1]$ is denoted by $\pi(a|b, \theta) := \Pr(a|b, \theta)$. Here, Θ is a policy parameter space and $\theta \in \Theta$ is the policy parameter. The policy takes the belief b as an argument, since b is a sufficient statistic for the history of the state transition. Let θ be decomposed into $\{\theta_k\}_k$ and θ_k effects the policy function related to the k-th state. For example, a stochastic policy as a mixed normal distribution as $\pi(a|b, \theta) = \sum_{k=1}^{K} b(s_k)\mathcal{N}(a|\theta_{k,1}, \theta_{k,2})$ with $\theta_k = (\theta_{k,1}, \theta_{k,2})$ where $\theta_{k,1}$ denotes a mean and $\theta_{k,2}$ denotes a variance of the normal distribution. From a pair of a state and an action that the agent takes, it gains a reward through a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is denoted by $r(s, a)$. Here, we assume $r(s, a)$ is known through this paper.

We define a sequence of the variables to analyze its dynamics. Let A be a sequence of the action a with length T, i.e., $A := (a_1, \dots, a_T)$. Similarly, $O := (o_1, \dots, o_T)$, $B := (b_1, \dots, b_T)$ and $S := ((s_1^k)_{k=1}^K, \dots, (s_T^k)_{k=1}^K)$, where $s_t^k = 1$ means the agent remains the k-th state s_k at time t, and $s_t^k = 0$ otherwise. Also, we assume that we have independent and identically distributed n sequences. We describe A^i as the i-th sequence for $i = 1, \dots, n$, and $a_{i,t}$ as the t-th element of A^i. Also, let $A_n := (A^1, \dots, A^n)$ be a set of the n sequences. We also define O_n, S_n and B_n in the same way. For simplicity, this paper assumes the distribution of the initial state s_0 and initial belief b_0 to be known though it would be straightforward to estimate them from data.

2.2 Purposes of This Study

Purpose I. Model Identification and Parameter Estimation We aim to identify model as well as to estimate parameters. In other words, we seek to estimate the dimensionality of latent variables (and complexity of observation functions). Although this is more challenging than the estimation of distribution, identifying appropriate model complexity usually gives us smaller generalization error and also better understanding and interpretation of the estimated model. Suppose the true data generating process is written by

$$(A_n, O_n) \sim P(A, O|\theta^*, \lambda^*, M^*) \tag{1}$$

where θ^* and λ^* are true parameters and $M^* \in \mathcal{M}$ is a true model. Our first purpose is to determine M^* as well as to estimate $P(A, O|\theta^*, \lambda^*, M^*)$ (note that many models can exist in $\mathcal{M}$ which achieve the smallest KL divergence between the estimated model and the true model as M^*.)

Purpose II. Exploration of Reward Maximized Policy We aim to directly explore a policy that maximizes the expectation of the sum of future rewards, defined as follows:

$$\iiint \frac{1}{T} \sum_{t=1}^{T} \beta^{t-1} r(s_t, a_t) p(S, A, O|\theta, \lambda, M) dSdAdO, \tag{2}$$

where $\beta \in [0, 1)$ is a discount factor. It is well known, in the context of DPS, that a good parameter estimator (i.e., achieve a small KL divergence value) does not necessarily produce a high reward. We aim to achieve both at the same time. From a view of the exploration-exploitation trade-off, this paper focuses on the batch exploration of the policy based on given data. Sequential exploration-exploitation raises another significant challenge, i.e., online model selection of POMDPs, which is out-of-scope for this paper.

3 Simultaneous Inference and Policy Search

This section proposes a *marginalized weighted likelihood* (MWL) approach which serves an estimator for bridging Bayesian inference which is a natural way of model selection (Purpose I) and direct policy search which is a natural way to maximize reward (Purpose II).

3.1 Weighted Likelihood

Since direct maximization of the expectation of rewards using the first order condition of (2) is not tractable, we consider its lower bound as follows:

$$\begin{aligned} &\log \iiint p(S, A, O|\theta, \lambda) R_T^\beta dSdAdO \qquad (3) \\ &\geq \iiint q(S, A, O) \log \frac{p(S, A, O|\theta, \lambda) R_T^\beta}{q(S, A, O)} dSdAdO, \end{aligned}$$

where $R_T^\beta := R(S, A) := \frac{1}{T}\sum_{t=1}^{T} \beta^{t-1} r(s_t, a_t)$. It is easy to confirm that the lower bound is maximized when $q(S, A, O) \propto$

$p(S)p(A,O|\theta,\lambda)R_T^\beta$ holds. Then, a first order condition for maximizing the lower bound with respect to θ is as follows:

$$\iiint p(S)p(A,O|\theta,\lambda) \times \left(\sum_{t=1}^{T} \nabla_\theta \log \pi(a_t|b_t,\theta)\right) R_T^\beta dAdOdS = 0. \quad (4)$$

To obtain the solution of (4), we extend weighted likelihood for a DPS method of MDPs by [Ueno *et al.*, 2012] as follows:

$$\begin{aligned} & p^*(S,A,O|\theta,\lambda,M) \quad (5) \\ & := \prod_{t=1}^{T} \pi(a_t|b_t,\theta)^{Q_t^\beta} p_o(o_t|s_t)p(s_t|s_{t-1},a_{t-1},\lambda), \end{aligned}$$

where $Q_t^\beta := \sum_{j=t}^{T} \beta^{t-j} r(s_j,a_j)$. Here, Q_t^β is regarded as the discount sum of the future rewards, and it has the substantial effect on the reward maximization (it is also important for FABPS in Section 4.2). The weighted likelihood function has a weight Q_t^β as an exponent and it assures that the derivative of the weighted likelihood has a similar role to the first order condition (4). Along with discussion in [Ueno *et al.*, 2012], the maximizer of the weighted likelihood with respect to parameters asymptotically converges to the solution of (4). Theorem 1 in the following supports the claim.

Theorem 1. *Given the dataset with some λ' and M' and $p(S)$, a maximizer of $\int p(S)\log p^*(S,A,O|\theta,\lambda',M')dS$ with respect to θ converges to the solution of* (4) *as $T \to \infty$.*

Since the weighted likelihood plays important roles as a likelihood function and a reward maximizer, the weighted likelihood can construct an estimator for the model and parameters and bridge reward maximization and statistical inference (maximum likelihood).

3.2 Marginalized Weighted Likelihood

In order to take advantage of the weighted likelihood and integrate it with Bayesian inference, we propose *marginalized weighted log-likelihood* (MWL) as follows.

Definition 1. *Marginalized Weighted Log-Likelihood (MWL) for POMDPs with dataset* (A_n,O_n)

$$\begin{aligned} & MWL(A_n,O_n|M) \\ & := \max_q \int q(S_n) \log\left(\frac{p^*(S_n,A_n,O_n|M)}{q(S_n)}\right) dS_n. \quad (6) \end{aligned}$$

Here, the marginalized weighted likelihood $p^*(S_n,A_n,O_n|M)$ is given by marginalizing with the prior $\phi(\theta,\lambda)$ and $q(S_n)$ denotes the variational distribution. The model is selected by maximizing MWL as:

$$\tilde{M} := \arg\max_{M\in\mathcal{M}} MWL(A_n,O_n|M). \quad (7)$$

As explained above, individual $p^*(S_n,A_n,O_n|\theta,\lambda)$s converge to the maximum total reward solutions given S_ns, and thus the maximizer of (6) is also expected to achieve high total reward (Purpose II) with an appropriately chosen $q(S_n)$. Further, as many Bayesian model selection studies have already shown ([Konishi and Kitagawa, 2008]), the marginalization over unknown parameters provides regularization effects supported by the Bayesian learning theory, which matches to our Purpose I, i.e., model selection of POMDPs.

4 Algorithm : FAB Policy Search

In practice, MWL is not tractable. This section proposes algorithms to achieve Purpose I and Purpose II by approximately maximizing MWL.

4.1 Weighted FIC for MWL

We follow the idea of factorized information criterion (FIC) ([Fujimaki and Morinaga, 2012]) which has been empirically and theoretically proven to be an asymptotically accurate approximation of marginal log-likelihood and to outperform the other state-of-the-art model selection methods for many models ([Fujimaki and Hayashi, 2012], [Hayashi and Fujimaki, 2013], [Eto *et al.*, 2014], [Liu *et al.*, 2015], [Hayashi *et al.*, 2015]). We apply the idea of FIC to MWL and propose *weighted factorized information criteria (wFIC)* which is derived as follows:

$$\begin{aligned} & wFIC(A_n,O_n,M) \quad (8) \\ & := \max_{q,\theta,\lambda} \sum_{S_n} q(S_n) \Bigg[\log p^*(S_n,A_n,O_n|\theta,\lambda,M) - \log q \\ & \utilde{- \sum_{k=1}^{K} \frac{D_{\lambda_k}}{2} \log\left(\sum_{i,t=1}^{n,T} s_{i,t}^k\right)} \underline{\underline{- \sum_{k=1}^{K} \frac{D_{\theta_k}}{2} \log\left(\sum_{i,t=1}^{n,T} Q_{i,t}^\beta s_{i,t}^k\right)}} \Bigg], \end{aligned}$$

where $Q_{i,t}^\beta := \sum_{j=t}^{T} \beta^{t-j} r(s_{i,j},a_{i,j})$ and D_z denotes the dimensionality of z. Roughly speaking, the formation of wFIC is derived by the Laplace's approximation method and the variational approximation.

We observe interesting terms of wFIC in their regularization terms. The double underline in (8) depends on $Q_{i,t}^\beta$ and $Q_{i,t}^\beta$ is the sum of rewards with the discount factor β. The terms means that wFIC more actively penalizes latent states with smaller rewards and encourage those with larger rewards, i.e., it enables to eliminate latent states (model selection) by keeping high-rewarded ones (reward maximization). We will discuss the effect of the regularizers in Section 4.2.

The following theorem follows [Fujimaki and Morinaga, 2012] and guarantees asymptotic accuracy of wFIC.

Theorem 2.

$$MWL(A_n,O_n|M) = \textit{wFIC}(A_n,O_n,M) + O(1).$$

4.2 FAB Policy Search

In this section, we propose an algorithm to optimize the model and the policy which maximizes wFIC and refer to it as *Factorized Asymptotic Bayesian Policy Search (FABPS)*. The FABPS algorithm performs an EM-like alternating maximization of q and (θ,λ).

The FABPS algorithm is constituted by following three steps: E-step, M-step, and a shrinkage step. The E-step and the M-step follow the EM-algorithm and update q and (θ,λ) from initial values. The shrinkage step is performed after each E-step, and it eliminates irrelevant latent variables. At the beginning of the FAB algorithm, $(q^{(0)},\theta^{(0)},\lambda^{(0)})$ are randomly initialized. Let the superscript (ℓ) stand for an ℓ-th iteration. Also, we set the initial model M to a large number. We summarize the overview of the algorithm in Algorithm 1.

E-step An E-step optimizes q by fixing $\theta = \theta^{(\ell-1)}$ and $\lambda = \lambda^{(\ell-1)}$ as well as the belief distribution $\{b_{i,t}\}_{i,t}$. Although there exists no closed-form update equation, q can be efficiently updated by the forward-backward algorithm as follows:

$$q^{(\ell)}(s_{i,t}^k) = f_{i,t}^k b_{i,t}^k, \tag{9}$$

where

$$f_{i,t}^k = \begin{cases} \frac{1}{\zeta_{i,1}}\tilde{p}(a_{i,1}, o_{i,1}|s_{i,1}) \quad (\text{ if } t = 1) \\ \frac{1}{\zeta_{i,t}}\tilde{p}(a_{i,t}, o_{i,t}|s_{i,t}^k) \sum_{k'=1}^K f_{i,t-1}^{k'} \pi(a_{i,t-1}|b_{i,t-1}^{k'}, \theta) \\ \times p(s_{i,t}^k|s_{i,t-1}^{k'}, a_{i,t-1}, \lambda) \quad (\text{otherwise}) \end{cases}$$

$$b_{i,t}^k = \begin{cases} 1 \quad (\text{ if } t = T) \\ \frac{1}{\zeta_{i,t+1}} \sum_{k'=1}^K b_{1,t-1}^{k'} \tilde{p}(a_{i,t-1}, o_{i,t-1}|s_{i,t-1}^k, \theta) \ , \\ \times p(s_{i,t}^{k'}|s_{i,t-1}^k, a_{i,t-1}, \lambda) \quad (\text{otherwise}) \end{cases}$$

and $\zeta_{i,t}$ is a normalization constant for $\sum_{k=1}^K f_{i,t}^k = 1$. Here, $\tilde{p}$ is defined by:

$$\tilde{p}(a_{i,t}, o_{i,t}|s_{i,t}^k, \theta) = p_o(o_{i,t}|s_{i,t}^k)\pi(a_{i,t}|b_{i,t}^k, \theta_k)^{Q_{i,t}^\beta} \delta_{i,t,k}^\pi \delta_{i,t,k}^p, \tag{10}$$

where $\delta_{i,t,k}^\pi$, and $\delta_{i,t,k}^p$ are defined as follows:

$$\delta_{i,t,k}^\pi := \frac{1}{\Delta_t^\pi} \exp\left(\frac{-D_{\theta_k} Q_{i,t}^\beta}{2\sum_{i,t=1}^{n,T} q_{i,k,t}^{(\ell-1)} Q_{i,t}^\beta}\right), \tag{11}$$

$$\delta_{i,t,k}^p := \frac{1}{\Delta_t^p} \exp\left(\frac{-D_{\theta_{\lambda_k}}}{2\sum_{i,t=1}^{n,T} q_{i,k,t}^{(\ell-1)}}\right), \tag{12}$$

where $q_{i,k,t}^{(\ell-1)} := q^{(\ell-1)}(s_{i,t}^k)$. The terms $\Delta_{i,t}^\pi$ and $\Delta_{i,t}^p$ are normalization constants to make $\sum_{k=1}^K \delta_{i,t,k}^\pi = 1$ and $\sum_{k=1}^K \delta_{i,t,k}^p = 1$, respectively.

The terms $\delta_{i,t,k}^\pi$ and $\delta_{i,t,k}^p$ come from the regularization terms in (8). First, $\delta_{i,t,k}^p$ also appears in FAB for HMMs ([Fujimaki and Hayashi, 2012]) and has an effect to eliminate small hidden states (i.e., $\sum_{i,t=1}^{n,T} q_{i,k,t}^{(\ell-1)}$ is small) through EM iterations. It has been well-studied that FAB algorithms automatically select hidden state dimensionality by this "shrinkage" effect. The term $\delta_{i,t,k}^\pi$ is unique in FABPS. For $\delta_{i,t,k}^\pi$, it eliminates hidden states having less expected rewards from the model. Therefore, this regularization term has an effect of removing hidden states having poor policies.

M-step An M-step optimizes λ and θ by fixing $q = q^{(\ell)}$. The transition parameter λ is updated by:

$$\lambda^{(\ell)} = \arg\min_\lambda \sum_{i,t,k,k'=1}^{n,T,K,K} KL\big[p(s_{i,t}^k|s_{i,t-1}^{k'}, a_{i,t-1}, \lambda) \\ \|q(s_{i,t}^k|s_{i,t-1}^{k'}, a_{i,t-1})\big], \tag{13}$$

where $KL[\cdot\|\cdot]$ is Kullback-Leibler divergence. The transition distribution is calculated by the elements of the Forward-Backward algorithm as $q(s_{i,t}^k|s_{i,t-1}^{k'}, a_{i,t-1}) = \frac{1}{\zeta_{i,t}} f_{i,t}^k f_{i,t-1}^{k'} \tilde{p}(a_{i,t}, o_{i,t}|s_t^k, \theta) p(s_{i,t}^k|s_{i,t-1}^{k'}, a_{i,t-1}, \lambda) b_{i,t}^k$.

Algorithm 1 FABPS algorithm for selecting K

```
Given $(A_n, O_n), \mathcal{M}, \pi(a|b,\theta), p_o(o|s), \epsilon > 0, r(s,a)$
Initialize $K \Leftarrow \arg\max_K \mathcal{M}$
Initialize $\{q^{(0)}\}_{i,t,k=1}^{n,T,K}, \theta^{(0)}, \lambda^{(0)}$
for $\ell = 1, 2, \ldots$ do
    /* E-step */
    $\theta \Leftarrow \theta^{(\ell-1)}, \lambda \Leftarrow \lambda^{(\ell-1)}$
    $\{(\delta_{i,t,k}^\pi, \delta_{i,t,k}^p)\}_{i,t,k=1}^{n,T,K} \Leftarrow$ (11) and (12)
    $\{\tilde{p}(a_{i,t}, o_{i,t}|s_{i,t}^k, \theta\}_{i,t,k=1}^{n,T,K} \Leftarrow$ (10)
    $\{q_{i,t,k}^{(\ell)}\}_{i,t,k=1}^{n,T,K} \Leftarrow$ (9)
    /* Shrinkage step */
    while $\sum_{i,t=1}^{n,T} q_{i,t,K} < \epsilon$ do
        $K \Leftarrow K - 1$   //Shrinkage operation
    end while
    /* M-step */
    $\{q_{i,t,k}\}_{i,t,k} \Leftarrow \{q_{i,t,k}^{(\ell)}\}_{i,t,k=1}^{n,T,K}$
    $\lambda^{(\ell)} \Leftarrow$ (13), $\theta^{(\ell)} \Leftarrow$ (14)
    /* Check convergence */
    if $\{q_{i,t,k}^{(\ell)}\}_{i,t,k}, \theta^\ell$ and $\lambda^\ell$ converge then
        break
    end if
end for
$\tilde{M} \Leftarrow K, \theta \Leftarrow \theta^{(\ell)}, \lambda \Leftarrow \lambda^{(\ell)}$
return $\tilde{M}, \theta, \lambda$
```

We update the policy function by solving the following problem:

$$\theta^{(\ell)} = \arg\max_\theta \left[\sum_{i,t,k=1}^{n,T,K} q(s_{i,t}^k) \log \pi(a_{i,t}|b_{i,t}^k, \theta) \right. \\ \left. - \sum_{k=1}^K \frac{D_{\theta_k}}{2} \log\left(\sum_{i,t=1}^{n,T} q(s_{i,t}^k)\right)\right]. \tag{14}$$

In the optimization problem for θ, the penalty term works as a l_0-penalty, and therefore it automatically controls complexity of the policy function as feature selection.

Shrinkage step As explained in the part of E-step, redundant and poorly-rewarded states are strongly regularized by (12). In this step, we perform a shrinkage operation which eliminates redundant states. More specifically, we check whether $\sum_{i,t} q^{(\ell)}(s_{i,t}^k) < \epsilon$ with sufficient small $\epsilon > 0$ for any k. If such k exists, we eliminate the k-th state from the model. The elimination operation is called *shrinkage*, and it enables the FAB algorithm to select the model automatically.

5 Experiments

5.1 Visual Demonstration: Heterogeneous Policy

One of the advantages of FABPS is a capability to select different complexities of policy functions in (14). This section

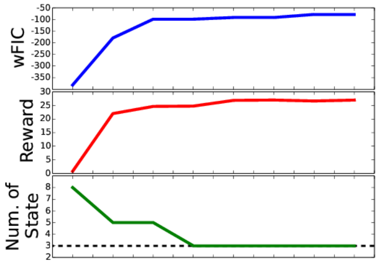

Figure 1: Improvement of wFIC, reward and K during FABPS.

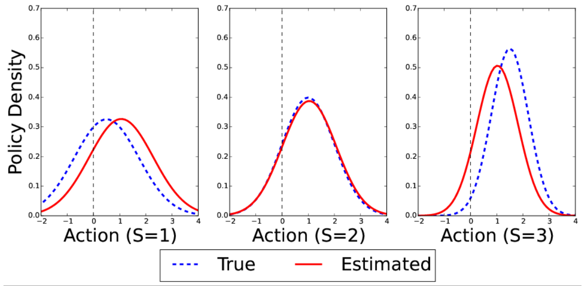

Figure 2: Estimated density of stochastic policy function.

visually demonstrates that FABPS practically discovers such heterogeneous policy functions. We set a POMDP model as follows by following a similar manner with the data generation of MDP models in [Ueno *et al.*, 2012]. The latent state space is $\mathcal{S} = \{1, 2, \ldots, K\}$ and the true K is 3. The action takes any real number : $\mathcal{A} = \mathbb{R}$, and sets the transition probability as follows: $p(s|s', a') = \lambda_k$ where s is the j-th nearest to $s' + a'$ with $\sum_{j=1}^{K} \lambda_j = 1$. The reward function is $r(s, a) = -s^2 - a^2 + 10$. We allow the policy to be heterogeneous : $\pi(a|b, \theta) = \sum_{k=1}^{K} b(s_k)\mathcal{N}(a|\theta_{k,1}, \theta_{k,2}^2)$, where $\mathcal{N}$ is a density function of Gaussian distribution. We set the synthetic data are generated from the policy function, and true parameter values are $(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}, \theta_{3,1}, \theta_{3,2}) = (0.5, 1.0, 1.5, 1.5, 1.0, 0.5)$. The data size is set to $N \times T = 1500$.

Fig. 1 illustrates how values of wFIC, total reward, and state dimensionality changed over FAB iterations. As can be observed, wFIC value monotonically increased as the theory guarantees. Although the total reward slightly decreased at the last stage of FAB iterations, the estimator maximizing wFIC almost maximized the total rewards. Further, the true state dimensionality was recovered by the regularization in wFIC without any prior knowledge and through only single path optimization (not grid search like cross validation).

In addition, Fig. 2 illustrates the estimated stochastic policy functions for each state $s = 1, s = 2$ and $s = 3$. FABPS successfully recovered the true parameter of policy functions for all hidden states (i.e., $S = 1$, $S = 2$, and $S = 3$.) This result demonstrated a unique simultaneous model selection capability of FABPS on hidden state dimensionality and policy function complexity.

5.2 Model Selection

We next evaluated model selection and reward maximization performance of wFIC, FIC[1], BIC ([Schwarz, 1978]) and iPOMDP ([Doshi-Velez *et al.*, 2015]) which is non-parametric Bayesian model selection for POMDPs. We generated data by following a similar manner with [Ueno *et al.*, 2012]. We set the hidden state space $\mathcal{S} = \{1, 2, \ldots, K\}$. The observation function is Gaussian with mean $s : p_o(o|s) = \mathcal{N}(s, 1)$. Also the action setting takes any real number : $\mathcal{A} = \mathbb{R}$, and the transition probability setting is as follows: $p(s|s', a') = \lambda_k$ where s is the k-th nearest to $s' + a'$ with $\sum_{k=1}^{K} \lambda_k = 1$. The reward function is $r(s, a) = (-s^2 - a^2 + 10)/10$, and the policy function is Gaussian mixture: $\pi(a|b, \theta) = \sum_{k,j=1}^{K,J} b(s_k)\mathcal{N}(a|\theta_{k,1}^j, \theta_{k,2}^2)$. True parameter values are $(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}, \theta_{3,1}, \theta_{3,2}) = (0.5, 1.0, 1.5, 1.5, 1.0, 0.5)$. We set the maximum value of J as 5 and that of K as 8. We randomly generated training data sequences with length 30 ($T = 30$) and different n. All of the values are means in 10 trials.

We evaluated two types of model selections: **(1)** Selection of hidden state dimensionality K, **(2)** Selection of order of policy function J.

Select hidden state dimensionality

In this evaluation, we set $J = 1$ for BIC and iPOMDP since they cannot choose J and M simultaneously. For FABPS and FAB, we set the maximum value of J as 5 and that of K as 8, and both J and K are selected automatically. Note that this setting is rather advantageous for BIC and iPOMDP since the oracle value of J is given to them.

Table 1 shows the estimated number of the hidden states. FABPS almost perfectly estimated the true number of hidden states. FAB also performs well, but FABPS outperformed it when the data size is smaller. BIC was strongly over-fitted and significantly overestimated the number of hidden states. iPOMDP performed fairly well but still was inferior to FABPS.

Table 2 shows the reward obtained by each method. We operate 100 agents to make an action for 100 times with the estimated policy and compute the sum of the rewards with the discount factor. wFIC achieved the highest reward which was slightly better than that of FIC. The other methods were significantly inferior. For BIC, it significantly over-fitted as it was in Table 1 and even the policy diverged with a large number of samples. For iPOMDP, it is not DPS and maximizes a different objective function. Eventually, its reward was less.

Select order of policy function

In this evaluation, we set $K = 3$ for BIC. Note that iPOMDP cannot select J so we excluded it from this evaluation. Fig. 3 shows estimated J values. As with the case of selecting K, wFIC performed the best and FIC also performed fairly well. BIC was again over-fitted and selected much larger order J than the true value.

Table 3 shows estimated J values, with the means of 10 trials. The estimation by wFIC is closer to the true value than that of the other methods. BIC performed badly due to the

[1] Although as standard FIC for POMDP has not been proposed, it can be derived in a similar manner to wFIC.

$n \times T$	wFIC	FIC	BIC	iPOMDP
1500	**3.0**	3.2	8.0	3.98
3000	**3.0**	3.1	7.0	3.85
6000	**3.0**	**3.0**	7.0	3.98
9000	**3.0**	**3.0**	7.0	4.05
12000	**3.0**	**3.0**	6.9	3.76

Table 1: Selected the number of states K. True value is 3.

$n \times T$	wFIC	FIC	BIC	iPOMDP
1500	**3.16**	3.06	1.79	1.69
3000	**2.93**	2.56	2.03	2.14
6000	**3.12**	2.32	0.0	2.11
9000	**3.01**	2.26	0.0	2.02
12000	**3.02**	2.94	0.0	2.53

Table 2: Simulated reward with estimated K and parameter.

$n \times T$	wFIC	FIC	BIC	iPOMDP
1500	**1.6**	2.1	4.8	N/A
3000	**1.5**	2.3	3.4	N/A
6000	**1.9**	2.0	3.4	N/A
9000	**1.3**	**1.3**	3.8	N/A
12000	**1.3**	2.0	4.2	N/A

Table 3: Select the order of policy function J. The true value is 1.

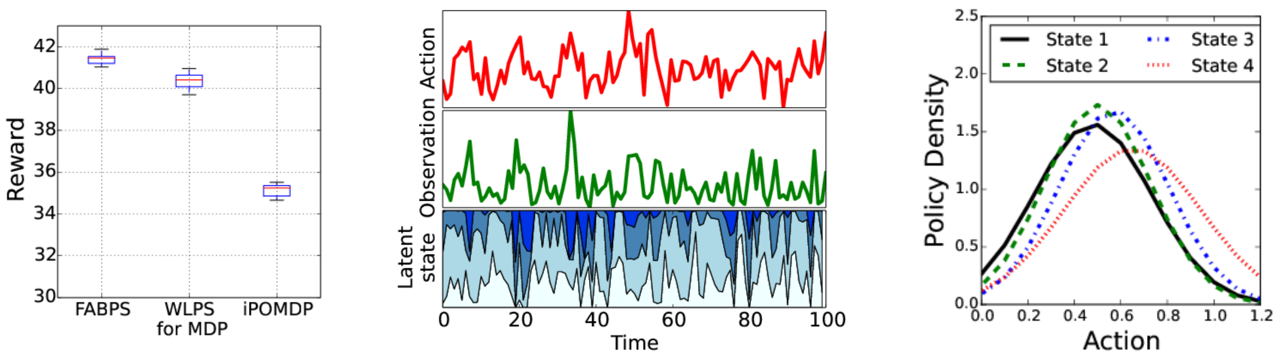

Figure 3: Reward by Helicopter data (left). Action and Observation in the Helicopter data, and the latent distribution estimated by FABPS (middle). Estimated policy densities for Helicopter data (right).

high effect of the likelihood term, and iPOMDP was unable to estimate the value.

For each of the experiments, wFIC and FIC take around 20 seconds to derive the solution, and BIC takes 30 seconds, and iPOMDP takes around 4 seconds.

5.3 Helicopter Data

We used data for helicopter control, provided by [Abbeel *et al.*, 2010]. With the data, we searched the policy for controlling stability of the helicopter. The data included that for the manipulation by an operator and the angle of the helicopter at given times. We set the angle as observations that the driver received, and also configure the manipulation with a control stick as actions. With given observation and the action, we tried to determine a latent environmental model from the training data and tried to maximize reward by estimated policies. The angle data is contained in $[-1,1]^3$, and the manipulation is in $[-1,1]^4$. Thus we use the norm of each data as variables for POMDPs. Also we set the reward as $r(s,a) = -s^2 - a^2 + 10$. We set the observation function and the policy function as follow : $p_o(o|s) = 0.5\exp(-0.5(o-s))$ and $\pi(a|b,\theta) = \sum_{k=1}^K b(s_j)\mathcal{N}(a|\theta_{k,1},\theta_{k,2}^2)$. We suppose the latent state variable takes discrete K values : $\mathcal{S} = \{1,2,\dots,K\}$.

We compared performances of wFIC, iPOMDP, and WLPS for MDP by [Ueno *et al.*, 2012]. WLPS for MDP is not for POMDP. Thus WLPS treated the observation as the state. iPOMDP requires the action and the observation to take discrete value. Thus we reformed the action space and the observation space to spaces with 4 elements. For the analysis with iPOMDP, the transition function and the observation distribution were also reformed to fit the discrete spaces.

The estimated values of K for wFIC and iPOMDP were 4 and 4.02, respectively. Since we do not know the true model (and the true model is not a POMDP), it is hard to say which methods outperformed with the others from the model selection point of view. However, one advantage of wFIC is again that it does not require any prior knowledge to determine K and J. Further, the left figure in Fig. 3 shows total reward of each method. As with the subsection 5.2, we operate 100 agents to make action for 100 times with the estimated policy, and compute the sum of the rewards with the discount factor. In addition to automatic and simultaneous model selection capability, wFIC achieved the highest reward. This indicates that the model selection of wFIC worked appropriately to determine appropriate model complexity to maximize reward as we expect.

We plotted the action and the observation in the data in the middle figure of Figure 3. Furthermore, the middle panel of Figure 3 contains an estimated distribution of the latent variable at the bottom, where the light blue shows the latent state with stable ($s = 1$), and the dark color ($s = 2, 3, 4$) represents unstable states. The result describes that the latent variables capture the effect of the action and observation; when the state is dark, the observation is unstable, and the manipulation also behaves severely. The estimated policy densities are plotted in the right figure in Figure 3. The result shows that FABPS can identify the heterogeneous policy in each state from the helicopter data.

Acknowledgements

M.Imaizumi is supported by Grant-in-Aid for JSPS Fellows (17J03208).

References

[Abbeel *et al.*, 2010] Pieter Abbeel, Adam Coates, and Andrew Y Ng. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 2010.

[Åström, 1965] Karl J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

[Busa-Fekete *et al.*, 2014] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, 2014.

[Cai *et al.*, 2009] Chenghui Cai, Xuejun Liao, and Lawrence Carin. Learning to explore and exploit in pomdps. In *Advances in Neural Information Processing Systems*, pages 198–206, 2009.

[Capitan *et al.*, 2013] Jesus Capitan, Matthijs TJ Spaan, Luis Merino, and Anibal Ollero. Decentralized multi-robot cooperation with auctioned pomdps. *The International Journal of Robotics Research*, 32(6):650–671, 2013.

[Debbi *et al.*, 2013] Hichem Debbi, Mustapha Bourahla, and Aimad Debbi. Medical treatment analysis using probabilistic model checking. *International Journal of Biomedical Engineering and Technology*, 12(4):346–359, 2013.

[Deisenroth and Rasmussen, 2011] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning*, pages 465–472, 2011.

[Doshi-Velez *et al.*, 2015] Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):394–407, 2015.

[Eto *et al.*, 2014] Riki Eto, Ryohei Fujimaki, Satoshi Morinaga, and Hiroshi Tamano. Fully-automatic bayesian piecewise sparse linear models. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2014.

[Fujimaki and Hayashi, 2012] Ryohei Fujimaki and Kohei Hayashi. Factorized asymptotic bayesian hidden markov models. In *Proceedings of the 29th International Conference on Machine Learning*, pages 799–806, 2012.

[Fujimaki and Morinaga, 2012] Ryohei Fujimaki and Satoshi Morinaga. Factorized asymptotic bayesian inference for mixture modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 400–408, 2012.

[Hayashi and Fujimaki, 2013] Kohei Hayashi and Ryohei Fujimaki. Factorized asymptotic bayesian inference for latent feature models. In *Advances in Neural Information Processing Systems*, pages 1214–1222, 2013.

[Hayashi *et al.*, 2015] Kohei Hayashi, Shinichi Maeda, and Ryohei Fujimaki. Rebuilding factorized information criterion: Asymptotically accurate marginal likelihood. In *Proceedings of the 32th International Conference on Machine Learning*, page 1358–1366, 2015.

[Howard, 1960] Ronald A Howard. *Dynamic programming and Markov processes*. MIT Press, 1960.

[Irissappane *et al.*, 2014] Athirai A Irissappane, Frans A Oliehoek, and Jie Zhang. A pomdp based approach to optimally select sellers in electronic marketplaces. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1329–1336, 2014.

[Konda and Tsitsiklis, 2000] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[Konishi and Kitagawa, 2008] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[Levine and Koltun, 2013] Sergey Levine and Vladlen Koltun. Variational policy search via trajectory optimization. In *Advances in Neural Information Processing Systems*, pages 207–215, 2013.

[Liu *et al.*, 2015] Chunchen Liu, Lu Feng, Ryohei Fujimaki, and Yusuke Muraoka. Scalable model selection for large-scale factorial relational models. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1227–1235, 2015.

[Peters and Schaal, 2007] Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, pages 745–750. ACM, 2007.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Schwarz, 1978] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[Spaan *et al.*, 2010] Matthijs TJ Spaan, Tiago S Veiga, and Pedro U Lima. Active cooperative perception in network robot systems using pomdps. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4800–4805, 2010.

[Ueno *et al.*, 2012] Tsuyoshi Ueno, Kohei Hayashi, Takashi Washio, and Yoshinobu Kawahara. Weighted likelihood policy search with model selection. In *Advances in Neural Information Processing Systems*, pages 2357–2365, 2012.

[Watanabe, 2009] Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009.

[Young *et al.*, 2010] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, 2010.