

System Identification Based on On-line Variational Bayes Method and Its Application to Reinforcement Learning

Junichiro Yoshimoto^{1,2}, Shin Ishii^{2,1}, and Masa-aki Sato^{3,1}

¹ CREST, Japan Science and Technology Corporation

² Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara 630-0192, Japan

{juniti-y, ishii}@is.aist-nara.ac.jp

³ ATR Human Information Science Laboratories

2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0237, Japan

masa-aki@atr.co.jp

Published in *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*, Lecture Notes in Computer Science 2714, pp.123-131, 2003.

©Springer-Verlag Berlin Heidelberg 2003

Abstract. In this article, we present an on-line variational Bayes (VB) method for the identification of linear state space models. The learning algorithm is implemented as alternate maximization of an on-line free energy, which can be used for determining the dimension of the internal state. We also propose a reinforcement learning (RL) method using this system identification method. Our RL method is applied to a simple automatic control problem. The result shows that our method is able to determine correctly the dimension of the internal state and to acquire a good control, even in a partially observable environment.

1 Introduction

A state space model provides a fundamental tool for system identification and control. If state transition and observation are defined by linear systems disturbed by white Gaussian noises, the state space model is formulated as a Gaussian process with internal (hidden) state. This means that the probabilistic generative model belongs to the exponential family with hidden variables and its system parameters can be determined by the expectation-maximization (EM) algorithm [2, 6], within the maximum likelihood (ML) estimation.

According to the ML estimation, however, it is difficult to estimate the dimension of the internal state. A Bayesian approach can overcome the difficulty; the marginal likelihood provides the evidence of a model structure [3]. Although actual implementation of Bayes inference is often difficult, the variational Bayes (VB) method [1, 4] provides an efficient approximation algorithm as a natural extension of the EM algorithm.

In this article, we present an on-line VB method for the identification of linear state space models with unknown internal dimension. The learning method is implemented as alternate maximization of an on-line free energy [7], which can be used for determining the dimension of the internal state. Using this system identification method, we also propose a belief state reinforcement learning (RL) method. Our RL method is applied to a simple automatic control problem. The result shows that our method is able to estimate correctly the dimension of the internal state and system parameters, and to acquire a good control, even in a partially observable environment.

2 Probabilistic Model for Linear State Space Model

We consider a stationary linear state space model defined by

$$x_{t+1} = Ax_t + Bu_t + v_t; \quad y_t = Cx_t + w_t, \quad (1)$$

where $x_t \in \mathfrak{R}^N$ is an internal state. $y_t \in \mathfrak{R}^D$ and $u_t \in \mathfrak{R}^M$ denote an observable variable and a control variable, respectively. Suffix t indexes the discrete time. $A \in \mathfrak{R}^{N \times N}$, $B \in \mathfrak{R}^{N \times M}$ and $C \in \mathfrak{R}^{D \times N}$ are system parameters. $v_t \sim \mathcal{N}_N(v_t | 0, Q)$ and $w_t \sim \mathcal{N}_D(w_t | 0, R)$ are white Gaussian noises[†].

According to equation (1), the likelihood for a sequence of internal states and observation variables $(X_{1:T}, Y_{1:T}) \equiv \{(x_t, y_t) | t = 1, \dots, T\}$, under a given sequence of control variables $U_{1:T-1} \equiv \{u_t | t = 1, \dots, T-1\}$, is given by

$$p(X_{1:T}, Y_{1:T} | U_{1:T-1}, \theta) = \prod_{t=1}^T p(x_t | \tilde{x}_{t-1}, \theta) p(y_t | x_t, \theta) \quad (2)$$

$$p(x_t | \tilde{x}_{t-1}, \theta) = \begin{cases} \mathcal{N}_N(x_1 | \mu, S) & (\text{if } t = 1) \\ \mathcal{N}_N(x_t | \tilde{A}\tilde{x}_{t-1}, Q) & (\text{otherwise}) \end{cases}$$

$$p(y_t | x_t, \theta) = \mathcal{N}_D(y_t | Cx_t, R),$$

where $\tilde{A} \equiv (A, B)$ and $\tilde{x}_t \equiv (x_t', u_t')'$. A prime ($'$) denotes a transpose. μ and S are a mean vector and a precision matrix for the initial state, respectively. $\theta \equiv \{\mu, S, \tilde{A}, Q, C, R\}$ is the set of model parameters. For simplicity, $S \equiv \text{diag}(s_1, \dots, s_N)$, $Q \equiv \text{diag}(q_1, \dots, q_N)$ and $R \equiv \text{diag}(r_1, \dots, r_D)$ are assumed.

We assume that the prior distribution of the model parameter θ is given by

$$p(\theta | \xi) = p(\mu, S | \sigma) p(\tilde{A}, Q | \Phi, \chi) p(C, R | \Psi, \rho) \quad (3)$$

$$p(\mu, S | \sigma) = \mathcal{N}_N(\mu | 0, \gamma_{\mu 0} S) \prod_{n=1}^N \mathcal{G}(s_n | \gamma_{s 0} / 2, \gamma_{s 0} \sigma / 2)$$

[†] $\mathcal{N}_p(x | \mu, S) \equiv (2\pi)^{-p/2} |S|^{-1/2} \exp[-\frac{1}{2}(x - \mu)' S (x - \mu)]$ denotes the probability density function of the random variable x , which is a p -dimensional normal distribution with a mean vector μ and an precision (inverse covariance) matrix S .

$$p(\tilde{A}, Q|\Phi, \chi) = \prod_{n=1}^N \mathcal{N}_{\tilde{N}}(\tilde{a}_n | 0, \gamma_{a0} q_n \Phi) \mathcal{G}(q_n | \gamma_{q0}/2, \gamma_{q0} \chi/2)$$

$$p(C, R|\Psi, \rho) = \prod_{n=1}^D \mathcal{N}_N(c_n | 0, \gamma_{c0} r_n \Psi) \mathcal{G}(r_n | \gamma_{r0}/2, \gamma_{r0} \rho/2),$$

where $\tilde{N} \equiv N + M$, $\tilde{A} \equiv (\tilde{a}_1, \dots, \tilde{a}_N)'$, $\tilde{a}_n \in \mathfrak{R}^{\tilde{N}}$, $C \equiv (c_1, \dots, c_D)'$ and $c_n \in \mathfrak{R}^N$. $\mathcal{G}(x|\alpha, \beta)$ denotes the probability density function of the random variable x , which is a gamma distribution with parameters α and β^\ddagger . $\xi \equiv \{\sigma, \chi, \rho, \Phi \equiv \text{diag}(\phi_1, \dots, \phi_{\tilde{N}}), \Psi \equiv \text{diag}(\psi_1, \dots, \psi_N)\}$ is the set of variable hyper parameters that parameterize the prior distribution of the model parameter θ .

We also assume a hierarchical prior distribution for the hyper parameter ξ :

$$p(\xi) \equiv p(\sigma)p(\chi)p(\rho)p(\Phi)p(\Psi); \quad p(\sigma) = \mathcal{G}(\sigma | \gamma_{\sigma 0}/2, \gamma_{\sigma 0} \tau_{\sigma 0}^{-1}/2)$$

$$p(\chi) = \mathcal{G}(\chi | \gamma_{\chi 0}/2, \gamma_{\chi 0} \tau_{\chi 0}^{-1}/2); \quad p(\rho) = \mathcal{G}(\rho | \gamma_{\rho 0}/2, \gamma_{\rho 0} \tau_{\rho 0}^{-1}/2)$$

$$p(\Phi) = \prod_{n=1}^{\tilde{N}} \mathcal{G}(\phi_n | \gamma_{\phi 0}/2, \gamma_{\phi 0} \tau_{\phi 0}^{-1}/2); \quad p(\Psi) = \prod_{n=1}^N \mathcal{G}(\psi_n | \gamma_{\psi 0}/2, \gamma_{\psi 0} \tau_{\psi 0}^{-1}/2).$$

In the above prior distribution, all hyper parameters with suffix ‘ $_0$ ’ are constant.

3 On-line Variational Bayes Method

After observing $Y_{1:T}$ by giving $U_{1:T-1}$, the objective of Bayes inference is to obtain the posterior distribution of the unknown variables, $p(X_{1:T}, \theta, \xi | Y_{1:T}, U_{1:T-1})$. According to the Bayes theorem, the posterior distribution is given by

$$p(X_{1:T}, \theta, \xi | Y_{1:T}, U_{1:T-1}) = p(X_{1:T}, Y_{1:T} | U_{1:T-1}, \theta) p(\theta | \xi) p(\xi) / p(Y_{1:T} | U_{1:T-1}).$$

The normalization term $p(Y_{1:T} | U_{1:T-1})^\S$ is called the marginal likelihood, which provides the evidence of the model structure[¶] [3].

Due to the hierarchical prior distribution, exact calculation of the posterior distribution and the marginal likelihood is difficult; we use an approximation method. In the variational Bayes (VB) method [1], the posterior distribution is approximated by a tractable trial distribution $q(X_{1:T}, \theta, \xi)$. This approximation is executed by maximizing the free energy:

$$F[q] = \log p(Y_{1:T} | U_{1:T-1}) - \text{KL}(q(X_{1:T}, \theta, \xi) \| p(X_{1:T}, \theta, \xi | Y_{1:T}, U_{1:T-1})). \quad (4)$$

$\text{KL}(\cdot \| \cdot)$ denotes the Kullback-Leibler divergence between two distributions, which becomes minimum at zero when $q(X_{1:T}, \theta, \xi) = p(X_{1:T}, \theta, \xi | Y_{1:T}, U_{1:T-1})$.

[‡] $\mathcal{G}(x|\alpha, \beta) \equiv \beta^\alpha x^{\alpha-1} e^{-\beta x} / \Gamma(\alpha)$ and $\Gamma(\alpha) \equiv \int_0^\infty t^{\alpha-1} e^{-t} dt$.

[§] $p(Y_{1:T} | U_{1:T-1}) \equiv \int d\theta d\xi dX_{1:T} p(X_{1:T}, Y_{1:T} | U_{1:T-1}, \theta) p(\theta | \xi) p(\xi)$.

[¶] Here, the model structure corresponds to the dimension of the internal state, N .

After the maximization of the free energy, therefore, the trial distribution provides a good approximation for the true posterior distribution. Also, the free energy well approximates the (log) marginal likelihood.

The trial distribution is assumed to be factorized as

$$q(X_{1:T}, \theta, \xi) = q_x(X_{1:T})q_\theta(\theta)q_\xi(\xi), \quad q_x(X_{1:T}) = \prod_{t=1}^T q_t(x_t|x_{t-1}),$$

where $q_1(x_1|x_0) \equiv q_1(x_1)$. In this case, the free energy (4) can be rewritten as

$$F[q] = TL - H^\theta - H^\xi; \quad L = \frac{1}{T} \sum_{t=1}^T \left\langle \left\langle \log \frac{p(x_t|\tilde{x}_{t-1}, \theta)p(y_t|x_t, \theta)}{q_t(x_t|x_{t-1})} \right\rangle_{\theta} \right\rangle_x \quad (5)$$

$$H^\theta = \left\langle \log \frac{q(\theta)}{\langle p(\theta|\xi) \rangle_\xi} \right\rangle_\theta; \quad H^\xi = \left\langle \log \frac{q(\xi)}{p(\xi)} \right\rangle_\xi,$$

where $\langle f(X_{1:T}) \rangle_x = \int dX_{1:T} q_x(X_{1:T}) f(X_{1:T})$, $\langle f(\theta) \rangle_\theta = \int d\theta q_\theta(\theta) f(\theta)$ and $\langle f(\xi) \rangle_\xi = \int d\xi q_\xi(\xi) f(\xi)$. L corresponds to the expected mean log-likelihood. According to a batch VB algorithm [4], the free energy (5) is maximized with respect to q_x , q_θ , and q_ξ , alternately, after observing all times series. Here, we derive an on-line VB algorithm [7], in which the free energy at time τ is redefined by

$$F_\tau^\lambda[q] = T_0 L_\tau^\lambda - H^\theta - H^\xi \quad (6)$$

$$L_\tau^\lambda = \eta(\tau) \sum_{t=1}^{\tau} \left(\prod_{s=t+1}^{\tau} \lambda(s) \right) \left\langle \left\langle \log \frac{p(x_t|\tilde{x}_{t-1}, \theta)p(y_t|x_t, \theta)}{q_t(x_t|x_{t-1})} \right\rangle_{\theta} \right\rangle_x.$$

$\eta(\tau) = (\sum_{t=1}^{\tau} \prod_{s=t+1}^{\tau} \lambda(s))^{-1}$ is a normalization term, and T_0 is a constant corresponding to the confidence of the observed time series relative to a *priori* belief of the model parameter. $\lambda(s)$ ($0 < \lambda(s) < 1$) is a time-dependent discount factor for forgetting the effect of early inaccurate inference. By introducing the discount factor, the expected mean log-likelihood is modified into a weighted mean log-likelihood.

The on-line VB algorithm can be implemented as an alternate maximization process of the on-line free energy (6). We here consider the inference at time τ , where $q_{1:\tau-1} \equiv \{q_t|t = 1, \dots, \tau - 1\}$, q_θ and q_ξ has been determined from previously observed time series $Y_{1:\tau-1}$. After observing a new output y_τ , the on-line free energy is maximized with respect to q_τ while $q_{1:\tau-1}$, q_θ and q_ξ are fixed. This is the VB-Estep. In the next step, the VB-Mstep, the on-line free energy (6) is maximized with respect to q_θ while $q_{1:\tau}$ and q_ξ are fixed. In the last step, the VB-Hstep, the on-line free energy is maximized with respect to q_ξ while $q_{1:\tau}$ and q_θ are fixed. Although detailed procedure cannot be described for the lack of space, the main part is as follows.

1. VB-Estep

$$\tilde{x}_\tau \leftarrow \begin{cases} \langle \mu \rangle_\theta & \text{if } \tau = 1 \\ \langle A \rangle_\theta \hat{x}_{\tau-1} + \langle B \rangle_\theta u_{\tau-1} & \text{otherwise} \end{cases}$$

$$\bar{V}_\tau \leftarrow \begin{cases} \langle S \rangle_\theta^{-1} & \text{if } \tau = 1 \\ \langle Q \rangle_\theta^{-1} + \langle A \rangle_\theta \hat{V}_{\tau-1} \langle A' \rangle_\theta & \text{otherwise} \end{cases}$$

$$K \leftarrow \bar{V}_\tau \langle C' \rangle_\theta \left(\langle R \rangle_\theta^{-1} + \langle C \rangle_\theta \bar{V}_\tau \langle C' \rangle_\theta \right)^{-1}$$

$$\hat{x}_\tau \leftarrow \bar{x}_\tau + K (y_\tau - \langle C \rangle_\theta \bar{x}_\tau); \quad \hat{V}_\tau \leftarrow (I - K \langle C \rangle_\theta) \bar{V}_\tau \quad (7)$$

$$J \leftarrow \hat{V}_{\tau-1} \langle A' \rangle_\theta \bar{V}_\tau^{-1} \quad (8)$$

$$\hat{x}_{\tau-1} \leftarrow \hat{x}_{\tau-1} + J (\hat{x}_\tau - \bar{x}_\tau); \quad \hat{V}_{\tau-1} \leftarrow \hat{V}_{\tau-1} + J (\hat{V}_\tau - \bar{V}_\tau) J' \quad (9)$$

$$\langle\langle y_t y_t' \rangle\rangle \leftarrow (1 - \eta(\tau)) \langle\langle y_t y_t' \rangle\rangle + \eta(\tau) y_\tau y_\tau'$$

$$\langle\langle y_t x_t' \rangle\rangle \leftarrow (1 - \eta(\tau)) \langle\langle y_t x_t' \rangle\rangle + \eta(\tau) y_\tau \hat{x}_\tau'$$

$$\langle\langle x_t x_t' \rangle\rangle \leftarrow (1 - \eta(\tau)) \langle\langle x_t x_t' \rangle\rangle + \eta(\tau) (\hat{V}_\tau + \hat{x}_\tau \hat{x}_\tau')$$

$$\langle\langle \tilde{x}_{t-1} \tilde{x}_{t-1}' \rangle\rangle \leftarrow (1 - \eta(\tau)) \langle\langle \tilde{x}_{t-1} \tilde{x}_{t-1}' \rangle\rangle$$

$$+ \eta(\tau) \begin{pmatrix} \hat{V}_{\tau-1} + \hat{x}_{\tau-1} \hat{x}_{\tau-1}' & \hat{x}_{\tau-1} u_{\tau-1}' \\ u_{\tau-1} \hat{x}_{\tau-1}' & u_{\tau-1} u_{\tau-1}' \end{pmatrix} \quad (10)$$

$$\langle\langle x_t \tilde{x}_{t-1}' \rangle\rangle \leftarrow (1 - \eta(\tau)) \langle\langle x_t \tilde{x}_{t-1}' \rangle\rangle + \eta(\tau) \begin{pmatrix} \hat{V}_\tau J' + \hat{x}_\tau \hat{x}_{\tau-1}' \\ \hat{x}_\tau u_{\tau-1}' \end{pmatrix}. \quad (11)$$

Equations (8)-(11) are used if $\tau > 1$. I is an identity matrix and $\langle\langle \cdot \rangle\rangle$ is the weighted mean of sufficient statistics: $\langle\langle f(\cdot) \rangle\rangle \equiv \eta(\tau) \sum_{t=1}^{\tau} (\prod_{s=t+1}^{\tau} \lambda(s)) \times \int dX_{1:\tau} q_x(X_{1:\tau}) f(\cdot)$.

2. VB-Mstep

$$\Xi \leftarrow \left((T_0 - 1) \langle\langle \tilde{x}_{t-1} \tilde{x}_{t-1}' \rangle\rangle + \gamma_{a0} \langle \Phi \rangle_\xi \right)$$

$$\langle \tilde{A} \rangle_\theta \leftarrow ((T_0 - 1) \langle\langle x_t \tilde{x}_{t-1}' \rangle\rangle) \Xi^{-1}$$

$$\langle Q \rangle_\theta^{-1} \leftarrow \frac{\text{diag} \left((T_0 - 1) \langle\langle x_t x_t' \rangle\rangle - \langle \tilde{A} \rangle_\theta \Xi \langle \tilde{A}' \rangle_\theta + \gamma_{q0} \langle \chi \rangle_\xi \right)}{T_0 - 1 + \gamma_{q0}}$$

$$\Upsilon \leftarrow \left(T_0 \langle\langle x_t x_t' \rangle\rangle + \gamma_{c0} \langle \Psi \rangle_\xi \right); \quad \langle C \rangle_\theta \leftarrow (T_0 \langle\langle y_t x_t' \rangle\rangle) \Upsilon^{-1}$$

$$\langle R \rangle_\theta^{-1} \leftarrow \frac{\text{diag} \left(T_0 \langle\langle y_t y_t' \rangle\rangle - \langle C \rangle_\theta \Upsilon \langle C' \rangle_\theta + \gamma_{r0} \langle \rho \rangle_\xi \right)}{T_0 + \gamma_{r0}}$$

$$\langle \mu \rangle_\theta \leftarrow \hat{x}_1 / (1 + \gamma_{\mu 0})$$

$$\langle S \rangle_\theta^{-1} \leftarrow \frac{\text{diag} \left(\hat{V}_1 + \hat{x}_1 \hat{x}_1' + \gamma_{s0} \langle \sigma \rangle_\xi - \gamma_{\mu 0} \langle \mu \rangle_\theta \langle \mu' \rangle_\theta \right)}{1 + \gamma_{s0}}.$$

3. VB-Hstep

$$\langle \sigma \rangle_\xi \leftarrow \frac{N \gamma_{s0} + \gamma_{\sigma 0}}{\gamma_{s0} \text{Tr} [\langle S \rangle_\theta] + \gamma_{\sigma 0} \tau_{\sigma 0}^{-1}}$$

$$\langle \Phi \rangle_\xi \leftarrow \left(\frac{\text{diag} \left(\gamma_{a0} \langle \tilde{A}' \rangle_\theta \langle Q \rangle_\theta \langle \tilde{A} \rangle_\theta + N \Xi^{-1} + \gamma_{\phi 0} \tau_{\phi 0}^{-1} I \right)}{N + \gamma_{\phi 0}} \right)^{-1}$$

$$\begin{aligned} \langle \chi \rangle_\xi &\leftarrow \frac{N\gamma_{q0} + \gamma_{\chi 0}}{\gamma_{q0} \text{Tr}[\langle Q \rangle_\theta] + \gamma_{\chi 0} \tau_{\chi 0}^{-1}}; & \langle \rho \rangle_\xi &\leftarrow \frac{D\gamma_{r0} + \gamma_{\rho 0}}{\gamma_{r0} \text{Tr}[\langle R \rangle_\theta] + \gamma_{\rho 0} \tau_{\rho 0}^{-1}}. \\ \langle \Psi \rangle_\xi &\leftarrow \left(\frac{\text{diag} \left(\gamma_{c0} \langle C' \rangle_\theta \langle R \rangle_\theta \langle C \rangle_\theta + D\Upsilon^{-1} + \gamma_{\psi 0} \tau_{\psi 0}^{-1} I \right)}{D + \gamma_{\psi 0}} \right)^{-1}. \end{aligned}$$

4 Belief State Reinforcement Learning

If the true system is consistent with model (1) and we observe the internal state x_τ at every time, we can acquire a good control by applying a reinforcement learning (RL) method, which is formulated as a continuous Markov decision process (MDP). In most of RL methods, mapping from x_τ to u_τ is determined based on rewards received through experiences. In our situation, however, only partial information of the internal state x_τ , i.e., y_τ , is observable. If we try to determine mapping from y_τ to u_τ , which is referred to as a partially observable MDP (POMDP), it is known that the performance is poor [8]. Instead of that, we may consider a control policy that maps a belief state $p(x_\tau | Y_{1:\tau}, U_{1:\tau-1})$ to u_τ . This formulation is a belief state MDP, in which the sequence $\{p(x_\tau | Y_{1:\tau}, U_{1:\tau-1}), u_\tau | \tau = 1, \dots\}$ becomes a Markov process under a fixed control policy.

Under the formulation of the belief state MDP, our system identification method can be applied to automatic control problems. By applying equation (7) in the VB-Estep, we obtain the trial posterior distribution $q_x(x_\tau)$ as the normal distribution $\mathcal{N}_N(x_\tau | \hat{x}_\tau, \hat{V}_\tau^{-1})$, which approximates a current belief state $p(x_\tau | Y_{1:\tau}, U_{1:\tau-1})$. Therefore, the underlying POMDP is regarded as the MDP defined over state $\hat{s}_\tau \equiv (\hat{x}_\tau, \hat{V}_\tau)$ and control u_τ .

In order to solve the decision problem, we apply an actor-critic algorithm proposed by Konda [5]. Although it was derived for the general parameterized families of randomized stationary policies, we employ the following stochastic policy: $\pi_\vartheta(u_\tau | \hat{s}_\tau = \hat{s}) = \prod_{m=1}^M \mathcal{N}_1(u_{\tau m} | \sum_{k=1}^K \vartheta_{mk} \hat{s}_k, \vartheta_0)$. Here, $u_{\tau m}$ is the m -th element of u_τ and \hat{s}_k ($k = 1, \dots, K$) is the k -th element of \hat{s} . $\vartheta \equiv \{\vartheta_0, \vartheta_{mk} | m = 1, \dots, M; k = 1, \dots, K\}$ is the set of parameters that parameterize the control policy π_ϑ . The way to update parameter ϑ is referred to [5].

In the above belief state MDP, the determination of the dimension of the internal state is an important issue because it directly determines the dimension of the belief state \hat{s} and affects the performance. We determine it by the following procedure. We prepare several state space models, which have different dimensions with each other. An RL system is coupled with each state space model. At the beginning of each episode, an RL system corresponding to the model with the largest free energy is selected and trained using experiences in the episode. On the other hand, every state space model is trained using all time series. Figure 1 shows the architecture of our proposed method. Here, the arrow lines denote the stream of signals. The gray rectangle is a couple of the RL system and the state space model selected by maximum free energy criterion.

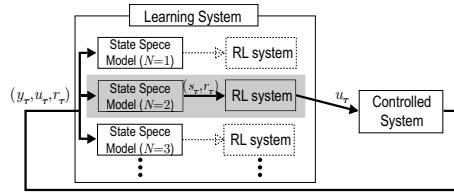


Fig. 1. Architecture of our RL method.

5 Experiment

Our RL method including the system identification is applied to an automatic control problem for a noisy linear system, which is defined by

$$\ddot{x} = 3u + v; \quad y = x + w; \quad v \sim \mathcal{N}_1(v|0, 0.1); \quad w \sim \mathcal{N}_1(w|0, 0.0009),$$

where initial states of x and \dot{x} in each episode are generated from a uniform distribution within ranges of $x \in [-10, 10]$ and $\dot{x} \in [-1, 1]$, respectively. The true dimension of the internal state is $N = 2$. Observation y is given with time interval 0.02 and a single episode consists of 1000 time steps. Control u is bounded within the range of $|u| \leq 1$. The reward function is defined by $r(x, \dot{x}, u) = -0.02x^2 - 0.1\dot{x}^2 - 0.001u^2$, which encourages the state to stay at the origin.

We prepared five state space models, whose dimensions varied from $N = 1$ to $N = 5$. Figure 2 shows the time courses of the free energy for the five state space models. The horizontal axis denotes the number of episodes. The vertical axes of left and right figures denote the absolute and relative free energies to the model structure with $N = 1$, respectively^{||}. The free energy of every model structure increases as the learning proceeds. Although the simplest model ($N = 1$) is selected at an early learning stage, the correct model ($N = 2$) has the largest free energy value after 50 episodes. In a later learning stage, the difference between $N = 1$ and $N = 2$ comes to be small because the system is stabilized around the origin, but they were not reversed even if the learning continued. This result shows that our system identification method is able to determine correctly the dimension of the target system.

Figure 3 shows the learning curve of our RL method and a naive actor-critic method without using belief state. The solid and dotted lines denote the time courses of the average cumulative rewards per 20 episodes acquired by our method and the naive method, respectively. Figure 4 shows the test control sequence by our RL system after learning. The left and right figures show the time series of observation y_t and control u_t , respectively. The control policy was gradually improved so that the system is sustained at the origin. This result shows our RL method coupled with the system identification by on-line VB algorithm is able to acquire a good control policy, while the underlying environment is an instance of POMDPs.

^{||} Although the left figure does not show the result of $N = 1$, one can know its value easily by seeing the left and right figures.

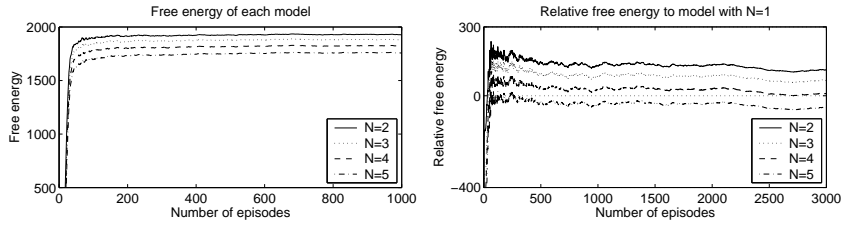


Fig. 2. Time courses of free energy by the five models.

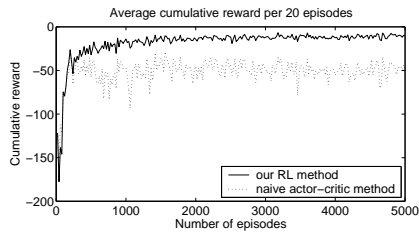


Fig. 3. Learning curve of our RL method and the naive actor-critic method.

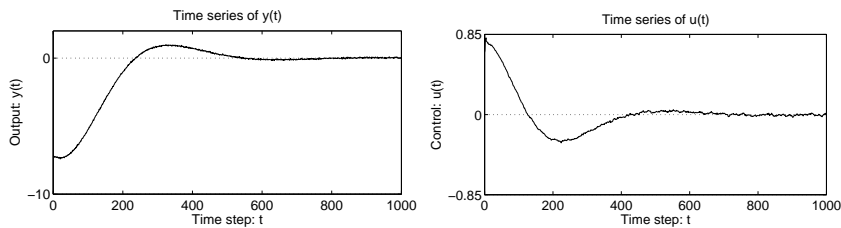


Fig. 4. A test control sequence by the trained system.

6 Conclusion

In this article, we presented an on-line VB algorithm for the identification of linear state space models. The method was able to estimate correctly system parameters, and the dimension of the internal state based on the free energy criterion. We also proposed an RL method using our system identification method, which can be applied to continuous POMDPs. We applied our RL method to a simple control problem. The result showed that our method was able to determine correctly the dimension of the internal state and to acquire a good control. Our near future work is to extend our system identification method to that for non-linear state space models.

Acknowledgement We would like to thank Dr. Doya for his valuable comments on this research. This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

1. Attias, H.: A variational Bayesian framework for graphical models, *Advances in Neural Information Processing Systems 12*, pp. 206–212 (2000).
2. Dempster, A. P. et al.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society B*, Vol. 39, pp. 1–38 (1977).
3. Frühwirth-Schnatter, S.: Bayesian model discrimination and Bayes factors for linear Gaussian state space models, *Journal of Royal Statistical Society B*, Vol. 57, pp. 237–246 (1995).
4. Ghahramani, Z. and Beal, M. J.: Propagation Algorithms for Variational Bayesian Learning, *Advances in Neural Information Processing Systems 13* (2001).
5. Konda, V. R.: *Actor-Critic Algorithms*, PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2002).
6. Roweis, S. and Ghahramani, Z.: A Unifying Review of Linear Gaussian Models, *Neural Computation*, Vol. 11, pp. 305–345 (1999).
7. Sato, M.: Online model selection based on the variational Bayes, *Neural Computation*, Vol. 13, No. 7, pp. 1649–1681 (2001).
8. Singh, S. P. et al.: Learning without state-estimation in partially observable Markovian decision processes, *Proceedings of the 11th International Conference on Machine Learning*, pp. 284–292 (1994).