The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08),
B-Con Plaza, Beppu, Oita, Japan, January 31-February 2, 2008

Continuous Internal-State Controller for a Partially Observable Linear Dynamical System

Yuki Taniguchi, Takeshi Mori
Graduate School of Information Science
Nara Institute of Science and Technology
Takayama 8916-5, Ikoma, Nara, 630-0192, Japan

Shin Ishii
Graduate School of Informatics
Kyoto University
Gokasho, Uji, Kyoto, 611-0011,Japan

Abstract

In this study, in order to control a partially observable linear dynamical system, we propose a novel framework, called continuous state controller (CSC). The CSC incorporates an auxiliary "continuous" state variable, called on internal state, whose stochastic process is Markov. The parameters of the transition probability of the internal-state are adjusted properly by a policy gradient-based reinforcement learning, and then the dynamics of the linear dynamical system can be extracted. Computer simulations show that the good control of the partially observable linear dynamical system is achieved by our CSC.

1 Introduction

Many reinforcement learning (RL) techniques have been successfully applied to completely observable environments [4], which are often formulated as Markov decision processes (MDPs). However, many real-world problems such as autonomous acquisition of the robot's control cannot be formulated as MDPs, because noises or obstacles, which are introduced to the robot's sensors, prevent the agents from observing all of the state variables accurately. Such problems can be formulated as partially observable Markov decision processes (POMDPs) [3], and some RL methods employing belief states have been developed for solving POMDPs [6]. However, those techniques suffer from several difficulties even with effective approximations [2]; the model of the environment should be known, and even if we know the model, the dimensionality of the belief space is usually high. These difficulties make the POMDP-RL with belief states infeasible in real-world problems.

Recently, a policy-gradient RL algorithm with finite state controllers (FSCs) has been proposed for solving POMDPs, called IState-GPOMDP [1]. The FSC is a probabilistic policy which incorporates an internal state as an input, and the transition probability of the internal state is identified by the policy-gradient RL algorithm, together with the optimization of the policy. The essential dynamic characters of the target state space can be extracted by learning of the transition of the internal state, which is performed in an irrelevant manner to the underlying dimensionality of the target state space. Because effective dimensionality for controls of a high-dimensional system is often much smaller than the dimensionality of the whole state space, as the dependence between the state variables increases, feature extraction by the IState-GPOMDP can be more effective than directly reconstructing the true state space as done by the belief state-based methods. We have shown the effectiveness of the IState-GPOMDP with the FSCs when applied to a partially observable multi-agent system through computer simulations [5].

Although the IState-GPOMDP shows good performance when applied to discrete-state dynamical systems, a direct application to more realistic problems whose state space is continuous should be intractable. In other word, essential features in such a continuous-state dynamical system may have continuous dynamics, which cannot be extracted by the naive IState-GPOMDP with the FSCs.

To overcome this difficulty, in this study, we propose a novel framework by introducing a continuous internal-state transition model, called a continuous state controller (CSC), instead of the previous finite-state alternative (FSC). In this new framework, the parametric transition model of the internal state in the CSC can be learned by the IState-GPOMDP together with the policy optimization. The parameters of the transition model are adjusted so as to maximize the average reward, and then the continuous dynamics of the essential features can be extracted in the framework of the reward maximization. We apply this algorithm to a control problem of a linear dynamical system under the assumption that some

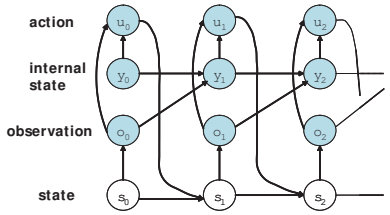

Figure 1: Graphical model of a POMDP with an internal state. The current action u_t and the following internal state y_{t+1} depend on the current internal state y_t and the observation o_t.

state variables cannot be observed. Computer simulation shows that our continuous internal-state transition model could successfully extract the dynamical characters of the missing variables, and simultaneously, the IState-GPOMDP could achieve good control for the dynamical system.

2 POMDPs with internal-state transition models

We consider a finite POMDP, consisting of a set of true states $\mathcal{S}$, a set of observations $\mathcal{O}$, a set of actions $\mathcal{U}$, each of which is the output of the controller, and a set of scalar rewards $\mathcal{R}$. At time t, the true state $s_t \in \mathcal{S}$ cannot be perceived directly by the agent, but instead, the observation o_t, observed by the agent, is drawn from the probability $P(o_t|s_t)$ conditioned on the state $s_t \in \mathcal{S}$. The observation o_t is, for example, the true state variable lacking some dimensionalities or those stained by observation noise. So, it is impossible to realize an optimal control based only on such insufficient observations. Then, we employ the model in which an internal state $y_t \in \mathcal{Y}$ is added to the POMDP definition above as an additional input to the policy. In the case of using the discrete internal state, such a policy is called a finite state controller (FSC) and has been successfully used to solve POMDPs [5]. The stochastic process over the internal state is prescribed by the transition probability $P(y_{t+1}|y_t, o_t)$, where y_t and y_{t+1} are the internal states at time t and $t+1$, respectively. In this case, the policy is the mapping from a pair of observation o_t and internal state y_t to the output u_t. Figure 1 represents the graphical model of a POMDP employing an internal state. The transition probability of the internal state $P(y_{t+1}|y_t, o_t)$ embedded in the FSC is identified by the policy gradient algorithm together with the optimization of the policy $P(u_t|y_t, o_t)$. The important dynamic characters of the target state space are extracted by learning of the transition probability of the internal-state.

3 Continuous state controllers for IState-GPOMDP

Although the FSCs can achieve good performance in controlling partially observable discrete-state dynamic systems, the variables in the state space are often continuous in the real world, and then the essential dynamic characters can also have continuous dynamics. In this section, we propose a novel framework called continuous state controller (CSC), which has a continuous internal state and can express the features with continuous dynamics. Then, we explain how to apply the IState-GPOMDP, which was formally proposed by Aberdeen and Baxter [1], to our CSCs.

The IState-GPOMDP is a policy gradient-based RL method which does not seek to estimate the value function, but adjusts the policy parameters θ and the internal-state transition parameters ϕ directly to maximize the average reward:

$$\eta(\phi, \theta) := \lim_{T\to\infty} \frac{1}{T} E_{\ \theta} \left[\sum_{t=1}^{T} r_t \right], \tag{1}$$

where $E_{\ \theta}$ denotes the expectation with respect to the trajectory $(s_0, y_0, o_0, u_0), (s_1, y_1, o_1, u_1), \ldots,$ prescribed by the parameters ϕ and θ. The internal state and the output of the controller are assumed to be drawn from the parameterized Gaussian distributions as

$$\begin{aligned} p(y_{t+1}|y_t, o_t) &:= \mathcal{N}(y_{t+1}|f(y_t, o_t; \phi), \sigma^2), \\ p(u_t|y_t, o_t) &:= \mathcal{N}(u_t|g(y_t, o_t; \theta), \sigma^2_{\theta}), \end{aligned} \tag{2}$$

where $f(y_t, o_t; \phi)$ and $g(y_t, o_t; \theta)$ are functions of (y_t, o_t) parameterized by ϕ and θ, respectively.

The following table shows the pseudo-code of the IState-GPOMDP applied to a continuous dynamical system to achieve good control.

```
0: while
1:   until the terminal condition, i.e., while t < T
2:     Observe o_t from p(o_t|s_t).
3:     Draw y_{t+1} from p(y_{t+1}|y_t, o_t, φ) and u_t from
       p(u_t|y_t, o_t, θ).
4:     Update the estimation of the policy gradient
       Δ_t with respect to φ and θ.
5:     Control the system by u_t.
6:     t++.
7:   end
8:     Δ_t is added to the parameters φ and θ,
     where    is the learning rate.
9: end
```

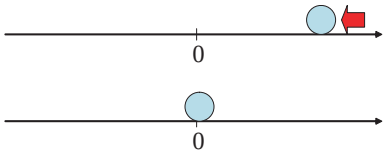

Figure 2: An LQR task, a simple example of the linear dynamical system. The goal of this task is to make the ball stop around the origin by controlling the force applied to the ball.

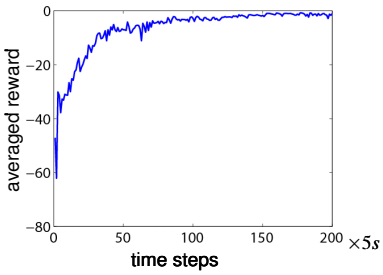

Figure 3: The learning process

4 Computer simulation

In this section, we apply the CSC to a partially observable linear dynamical system and evaluate its performance.

4.1 Linear dynamical system

A linear dynamical system which we apply our method to is the linear-quadratic regulator (LQR) task as Figure 2. The goal of this task is to make the ball stop around the origin. We define the dynamics of this system as

$$p(\mathbf{s}_{t+1}|\mathbf{s}_t, u_t) = \mathcal{N}(\mathbf{s}_{t+1}|\mathbf{A}\mathbf{s}_t + \mathbf{B}u_t, \mathbf{\Sigma}), \qquad (3)$$

where $\mathbf{s}_t = (x_t, v_t)^T$ denotes the state vector composed of the position and the velocity of the ball, u_t denotes the force applied to the ball at time t, and $\mathbf{\Sigma}$ denotes the state transition noise; $\mathbf{\Sigma} = \mathrm{diag}(1,1) \times 10^{-3}$. The matrices $\mathbf{A}$ and $\mathbf{B}$ denote the system parameters as follows:

$$\mathbf{A} = \begin{matrix} 1 & \\ 0 & 1 \end{matrix}, \mathbf{B} = \begin{matrix} 0 \\ \end{matrix},$$

where the time constant is set at $= 1/60s$. The agent observes the state and takes an action every time step.

The rewards are given by $r(\mathbf{s}_t, u_t) = -(\mathbf{s}_t^T \mathbf{Q} \mathbf{s}_t + Ru_t^2)$, where $\mathbf{Q} = \mathrm{diag}(0.025, 0.01)$ and $R = 0.01$.

In our experimental setting, a single learning episode continues in 10 seconds, and each parameter is updated every learning episode. If the absolute value of the position of the ball exceeds 10, the learning episode is terminated and the parameter is immediately updated. In this case, the position and the velocity are set to be around 0 randomly.

4.2 Partially observable linear dynamical system and policy parameterization

To evaluate the performance of the CSC, we apply it to the partially observable linear dynamical system. In this system, the velocity of the ball, which is necessary to control the system, cannot be observed. It is impossible to control the system smoothly only by the position, so we introduce the continuous internal states as to compensate the missing velocity.

More concretely, in the completely observable LQR, it is easily controlled by learning the parameters (θ_1, θ_2) of the policy:

$$P(u_t|o_t) = \mathcal{N}(u_t|\theta_1 x_t + \theta_2 v_t, \sigma_\theta^2). \qquad (4)$$

In the partially observable LQR, on the other hand, the velocity cannot be observed, so we replace it with the internal state of the policy:

$$P(u_t|o_t, y_t) = \mathcal{N}(u_t|\theta_1 x_t + \theta_2 y_t, \sigma_\theta^2), \qquad (5)$$

and the internal state transition is modeled as

$$P(y_{t+1}|o_t, y_t) = \mathcal{N}(y_{t+1}|\phi_1 x_t + \phi_2 y_t, \sigma^2), \qquad (6)$$

where both σ_θ^2 and σ^2 are 0.1^2. The action at time t and the internal state at time $t+1$ are drawn from these distributions, respectively. The parameters , and $\boldsymbol{\theta}$ are learned by the IState-GPOMDP.

4.3 Simulation result

We applied the IState-GPOMDP with our CSC to the controlling problem of the partially observable linear dynamical system above. The initial values of ϕ_1 and ϕ_2 are $0 + \epsilon$ ($\epsilon \sim \mathcal{N}(0, 10^{-2})$), and those of θ_1 and θ_2 are $-20 + \epsilon$. Figure 3 shows the learning process, which is averaged over 100 runs and smoothed over every 300 time step (5 seconds). The parameters were successfully learned so as to increase the averaged reward.

Figure 4 shows the time-series of the position, the velocity and the internal state, and Figure 5 shows those of the acceleration and the internal state, where panel(a) shows the time-series before learning and the panel(b) shows those after learning. The acquired

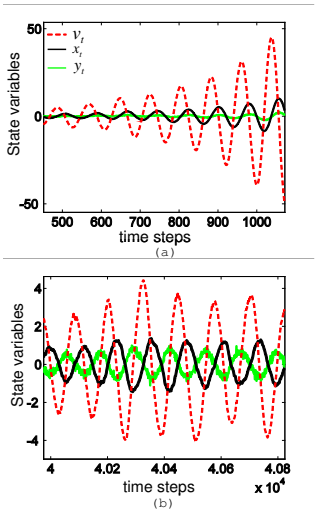

Figure 4: (a) The time-series of the position, velocity and internal state before learning. (b) The time-series after learning.

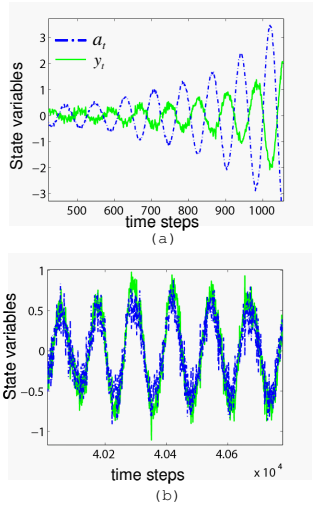

Figure 5: (a) The time-series of the acceleration and internal state before learning. (b) The time-series after learning.

internal-state transition seems to imitate the acceleration rather than the velocity. Because the average reward increased as shown in Figure 3, the acceleration seems to be an important feature for controlling this system.

5 Discussion

In our experimental result, good control of the linear dynamical system in the partially observable environment, in which the velocity cannot be observed, could be achieved by our method. As a result, the parameters are converging and the reward is getting close to 0, which shows the reliability of our method. However, the essential feature extracted by the internal state transition model, which is assumed to be necessary information to control the system, the acceleration rather than the velocity. A further analysis of such a result is interesting, but remains as our future work.

6 Concluding Remarks

In this article, we proposed the continuous state controller (CSC), which is a probabilistic policy incorporating continuous-state variables, and applied it to controlling a simple partially observable linear dynamical system. As a result, it could achieve good control by compensating a missing state variable with the continuous internal state. The CSC we proposed here is, however, a one-dimensional model, so it is necessary to extend it to be applicable to more complex systems or nonlinear-dynamical systems. Such extension of out CSC will be shown in our future study.

References

[1] Aberdeen, D., Baxter, J., "Scaling Internal State Policy-Gradient Methods for POMDPs," Proceedings of the 19th International Conference on Machine Learning, pp. 3-10, 2002.

[2] Hauskrecht, M., "Value-function approximations for partially observable Markov decision processes," *Journal of Arti cial Intelligence Research*, **13**, pp. 33-99, 2000.

[3] Kaelbling, L.P., Littman, M.L, Cassandra, A.R., "Planning and acting in partially observable stochastic domains," *Arti cial Intelligence*, **101**, pp. 99-134, 1998.

[4] Sutton, R., Barto, A., An introduction to reinforcement learning, MIT Press, 1998.

[5] Taniguchi, Y., Mori, T., Ishii, S., "Reinforcement Learning for Cooperative Actions in a Partially Observable Multi-Agent System," *Lecture Notes in Computer Science*, Vol. 4668, pp. 229-238, 2007.

[6] Thrun, S., "Monte Carlo POMDPs," *Advances in Neural Information Processing Systems*, Vol. 12, pp. 1064-1070, 2000.