部分観測環境における線形力学システムの強化学習を用いた自動制御

谷口 雄基 (PY), 森 健, 石井 信
Yuki Taniguchi(PY), Takeshi Mori, and Shin Ishii
京都大学情報学研究科
{yuki-t, tak-mori}@sys.i.kyoto-u.ac.jp
ishii@i.kyoto-u.ac.jp

Abstract— In order to control partially observable Markov decision processes, we propose a novel framework called continuous state controller (CSC). The CSC incorporates an auxiliary "continuous" state variable, called an internal state, whose stochastic process is Markov. Computer simulations show that good control of partially observable linear dynamical systems is achieved by our CSC.

Keywords— **Reinforcement Learning, Partially Observable Environments, Linear Dynamical Systems, Internal State, Policy Gradient Method**

1 はじめに

実世界の多くの問題は, 様々な障害やノイズなどの影響から, マルコフ決定過程 (Markov Decision Process; MDPs) として定式化できない. このような最適制御問題は部分観測マルコフ決定過程 (Partially Observable Markov Decision Process; POMDPs) として定式化できることがある.

近年, POMDP を解く強化学習法として, 有限状態コントローラ (Finite State Controllers; FSCs) を用いた方策勾配法である, IState-GPOMDP, Exp-GPOMDP が提案された [1]. FSC は内部状態を持つ確率的方策であり, その内部状態遷移確率は, 行動選択確率とともに方策勾配法によって学習される. その際, 真の状態空間の大きさとは無関係に, 状態空間から重要な特徴を内部状態遷移として抽出することができる.

FSC は離散状態のダイナミカルシステムには適用ができる一方で, 現実にしばしば見られるような状態空間が連続な問題には直接適用することができない. 連続状態を持つダイナミカルシステムでは, 抽出すべき特徴も連続であることが多いので, FSC ではそれを表現しきれない.

本研究ではこれを克服するために, 連続状態コントローラ (Continuous State Controllers; CSCs) という連続な内部状態遷移モデルを持つ新しい枠組みを提案する. CSC の学習には IState-GPOMDP と Exp-GPOMDP を用いる. このアルゴリズムを, 環境の要素の一部が観測できないような状況での線形力学システムの自動制御問題に用いる. 計算機実験により連続な内部状態遷移モデルがうまく力学的特徴を抽出できるように獲得されることを示す.

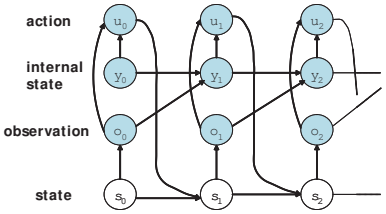

図 1: 内部状態を含む POMDP のグラフィカルモデル

2 内部状態遷移モデルと POMDP

POMDP は, 真の状態 $s \in \mathcal{S}$, エージェントが選択できる行動 $a \in \mathcal{A}$, 観測 $o \in \mathcal{O}$, 報酬 $r \in \mathcal{R}$ の 4 つの変数により記述される. 時刻 t における真の状態 $s_t \in \mathcal{S}$ は観測できないが, 代わりに s_t を条件とした確率 $P(o_t|s_t)$ によりサンプリングされる o_t を観測できる. この観測過程により不確実性が導入されるので, 観測のみに基づく方策では最適な行動を取ることはできない. よって, 付加入力として内部状態 $y_t \in \mathcal{Y}$ を加えた方策を用いることで上記の POMDP を解くことを考える. このような方策 (コントローラ) を FSC と言う [1]. 内部状態の確率過程は, 時刻 $t, t+1$ における内部状態 y_t, y_{t+1} を用いて $P(y_{t+1}|y_t, o_t)$ で表され, 行動選択過程は, $P(a_t|y_t, o_t)$ で表される. 図 1 に内部状態を含む POMDP のグラフィカルモデルを示す.

3 連続状態コントローラ

ここでは内部状態を連続に拡張した CSC の提案を行い, IState-GPOMDP, Exp-GPOMDP を CSC に適用する手法を示す.

3.1 CSC の学習

IState-GPOMDP は方策勾配ベースの強化学習法であり, 方策パラメータ θ と内部状態遷移パラメータ ϕ を以下の長期平均報酬が最大となる方向に直接調節する.

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

ここで $\mathbb{E}_{\phi,\theta}$ はパラメータ θ, ϕ により規定されるサンプル系列 $(s_0, y_0, o_0, c_0, a_0), (s_1, y_1, o_1, c_1, a_1), \ldots$ に関する

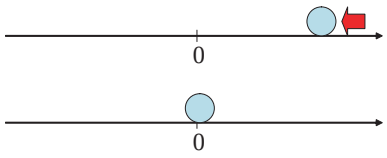

図 2: LQR タスク．このタスクの目的はボールを原点付近にとどめておくことである．

期待値を表す．内部状態と行動の出力はそれぞれ以下のガウス分布からサンプリングされるものとする．

$$p(y_{t+1}|y_t, o_t) := \mathcal{N}(y_{t+1}|f(y_t, o_t; \phi), \sigma_\phi^2)$$
$$p(u_t|y_t, o_t) := \mathcal{N}(u_t|g(y_t, o_t; \theta), \sigma_\theta^2) \quad (2)$$

ここで $f(y_t, o_t; \phi)$ と $g(y_t, o_t; \theta)$ はそれぞれ ϕ, θ でパラメトライズされた (y_t, o_t) の関数である．ϕ, θ は方策勾配法にしたがい，それぞれ

$$\phi \leftarrow \phi + \alpha \nabla_\phi \eta(\phi, \theta)$$
$$\theta \leftarrow \theta + \alpha \nabla_\theta \eta(\phi, \theta)$$

のように学習される．ここで α は学習係数である．

3.2 Exp-GPOMDP を用いた CSC の学習

前節の IState-GPOMDP を用いた CSC の学習では，各時刻において行動と内部状態をサンプリングしていたが，内部状態のサンプリングは，現時刻の内部状態の確率分布を伝搬させることにより省略することができる．この手法はサンプリングノイズを取り除くことができるので，より効率よく方策勾配を推定することができる．時刻 t の内部状態 y_t の分布が $p(y_t) = \mathcal{N}(y_t; \mu, \sigma_t^2)$ であるとき，y_{t+1} の分布は

$$\begin{aligned} p(y_{t+1}) &= \int_{y_t} p(y_{t+1}|y_t, o_t; \phi) p(y_t) dy_t \\ &= \int_{y_t} N(y_{t+1}; f(y_t, o_t; \phi), \sigma_\phi^2) N(y_t; \mu, \sigma_t^2) dy_t \end{aligned}$$

となり，同様に行動 u_t の選択確率分布は

$$\begin{aligned} p(u_t) &= \int_{y_t} p(u_t|y_t, o_t; \theta) p(y_t) dy_t \\ &= \int_{y_t} N(u_t; f(y_t, o_t; \theta), \sigma_\theta^2) N(y_t; \mu, \sigma_t^2) dy_t \end{aligned}$$

となる．これらの計算は $f(y_t, o_t; \phi)$ と $f(y_t, o_t; \theta)$ がともに線形であるとき解析的に行うことができる．

4 計算機実験

4.1 線形力学システム

図 2 に示すような LQR タスクに対して CSC を適用する．ダイナミクスは以下で定義する．

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

$\mathbf{s}_t = (x_t, v_t)^T$ は時刻 t におけるボールの位置と速度からなるベクトルであり，u_t はボールにかかる力 (制御信号) を表している．$\mathbf{\Sigma}$ は状態遷移確率分布の分散であり，$\mathbf{\Sigma} = \mathrm{diag}(1, 1) \times 10^{-3}$ である．$\mathbf{A}$ と $\mathbf{B}$ は以下とする．

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

時定数は $\tau = 1/60s$ と設定する．報酬は $r(\mathbf{s}_t, u_t) = -(\mathbf{s}_t^T \mathbf{Q} \mathbf{s}_t + R u_t^2)$, $\mathbf{Q} = \mathrm{diag}(0.025, 0.01)$, $R = 0.01$ で与えられる．

4.2 部分観測の線形力学システム

前節の LQR を部分観測問題にし，CSC を評価する．ここではボールの速度を観測できないようにし，連続な内部状態にそれを補う役割を期待する．

完全観測の LQR では，以下の方策パラメータ (θ_1, θ_2) を学習することで容易に制御できる．

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

一方部分観測 LQR では速度は観測できないので，速度を内部状態に置き換えた方策をとる．

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

内部状態遷移確率は以下のようになる．

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

ここでは σ_θ^2 と σ_ϕ^2 はともに 0.1^2 とした．

4.3 実験結果

上記の部分観測 LQR に IState-GPOMDP と Exp-GPOMDP を適用して学習を行った結果，その学習曲線は図 3 のようになった．図 3(b) はパラメータ収束点

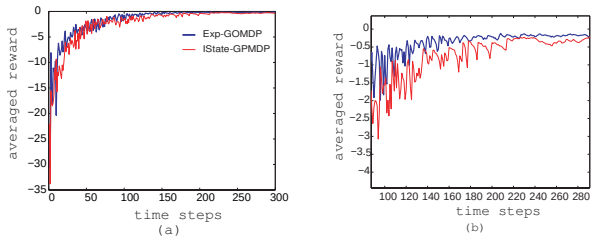

図 3: IState-GPOMDP と Exp-GPOMDP それぞれによる学習曲線

付近の学習曲線を示している．サンプリングノイズが乗らない分，Exp-GPOMDP の方が若干良い結果を示している．

5 まとめ

本研究では，連続な内部状態を持つ CSC を提案し，簡単な部分観測線形力学システムに適用することで，有効性を確認した．今後，非線形なシステムに適用できるように改良していきたい．

参考文献

[1] Aberdeen, D. and Baxter, J. (2002) ICML