機械学習手法を用いた頑健かつ効率的な行動制御

理学専攻 情報科学コース
齋藤 碧

1 はじめに

システムの状態推定やその制御において，不安定で複雑な対象に対しても頑健性や効率性のある処理が求められている．このことから，近年，システムの制御に機械学習手法を用いることが増えている．本研究では，パーティクルフィルタを用いることにより，非線形で複雑な状態空間の推定を可能にし，環境外乱に対して頑健性のある制御手法を確立した．また，学習知識の転移において，新しいタスク上で効率的な学習を行うことを目的とし，スパースコーディング [1] を用いた学習知識の選択手法を提案し，その性能を比較した．

2 パーティクルフィルタを用いた頑健な行動制御

2.1 提案手法

システム制御において，実時間内に複雑な制御対象の複数の情報を取得，近似計算をするには制御速度の問題がある．そのため提案手法では，倒立振子の安定化制御において振り子の傾く角度とそれに応じた制御量には依存関係があると仮定し，そのような場合に，1つの観測可能な値から他の値の推定ができるとした．また，一般的なパーティクルフィルタは，観測可能な値に対する追従を行うが，提案手法では複数のパーティクルフィルタを用いることにより観測不可能な値を推定する．これにより，適用するパーティクルフィルタの粒子の集合を $X_t = \{\hat{\boldsymbol{x}}_t^{(k)}, \hat{\boldsymbol{z}}_t^{(k)}, \pi_t^{(k)}\}_{k=1}^{K}$ として $\boldsymbol{x}$ と $\boldsymbol{z}$ が依存関係にある場合，$\boldsymbol{x}$ のみの観測から $\boldsymbol{z}$ の値を推定する．提案手法のアルゴリズムを示す．また，$\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{z}}_t$ はそれぞれ $\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{z}}_t$ に代入されたパーティクルの値を表している．

step1 初期設定：ランダムな K 個の $\hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{z}}_{t-1}$ を生成する
step2 予測：$\hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{z}}_{t-1}$ にそれぞれノイズを乗せる
step3 ソート：$\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{z}}_t$ をそれぞれ昇順にソートする
step4 観測値取得：観測可能な $\boldsymbol{x}$ を取得，$\boldsymbol{z}$ は観測値なし
step5 尤度計算：重み $\pi_t^{(k)} = p(\boldsymbol{x}_t|\hat{\boldsymbol{x}}_t^{(k)})(1 \leq k \leq K)$ を計算
step6 尤度共有：粒子のインデックスに重みを依存した形で $\hat{\boldsymbol{z}}$ にも $\hat{\boldsymbol{x}}$ と同じ尤度 π を与える
step7 リサンプリング：π_t に比例した確率で $\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{z}}_t$ を K 個抽出する
step8 時間更新：$t \to t+1$ として step2〜step7 を繰返す

step1〜step2 においては，通常のパーティクルフィルタと同じ操作を行う．step3 において，$\hat{\boldsymbol{x}}$，$\hat{\boldsymbol{z}}$ それぞれの値は step1 により初期パーティクルがランダムに生成されているため，それぞれのパーティクルの値を昇順に並べ替える．ここでソートを行うことにより，step6 において尤度を共有する際に，π_t を $\hat{\boldsymbol{x}}$ のインデックスに依存する形で $\hat{\boldsymbol{z}}$ の重みとして共有が可能となる．step7 以降は，$\hat{\boldsymbol{x}}$，$\hat{\boldsymbol{z}}$ 共に通常のパーティクルフィルタと同様である．このように，2つのパーティクルフィルタを用いて，その尤度を共有することにより，観測不能な $\boldsymbol{z}$ の推定値 $\hat{\boldsymbol{z}}$ を $\boldsymbol{x}$ から求めることが可能となる (図 1)．

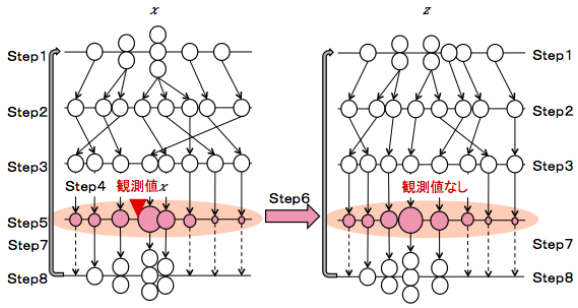

図 1: 尤度を共有する2連のパーティクルフィルタ

2.2 実験

2.2.1 実験設定

実験では先行研究 [2] を参考に，倒立振子を設定した．倒立振子は，上部の回転軸と下部の台車から構成されており，長さ $l = 0.5$(m), 質量 $m = 0.3$(kg) の回転軸を振り子とみなし，台車の質量 $m_c = 3.0$(kg) とする．振り子の角度を ϕ(rad) とし，台車の移動距離を p(m) とする．また，外乱を表現するために倒立振子の台車の初期位置付近に一辺 10.0(mm) の立方体を設置した．倒立振子が外乱に影響される場合においても，制御量 f は，振り子の直立状態からの角度 ϕ に依存する．角度 ϕ と台車にかける力 f を推定するパーティクルをそれぞれ 500 個用意する．パーティクルの集合を $X_t = \{\hat{\phi}_t^{(k)}, \hat{f}_t^{(k)}, \pi_t^{(k)}\}_{k=1}^{K}$ と設定した．f は，一般的には与えられた ϕ から計算して求めるものであるが，提案手法を用いて観測可能な ϕ から，観測不能である f の推定値 $\hat{f}$ を導く．推定するにあたり，$\hat{\phi}$ と，$\hat{f}$ の初期値の範囲を以下の様に設定する．

$$0.0 \leq \{\hat{\phi}_0^{(k)}\}_{k=1}^{K} \leq 0.5 \tag{1}$$

$$0.0 \leq \{\hat{f}_0^{(k)}\}_{k=1}^{K} \leq 10.0 \tag{2}$$

このように得られた推定値 $\hat{f}$ を $f \propto \hat{f}$ として出力する．倒立振子の対称性を考慮して，振り子が傾く方向により制御量 $\hat{f}$ を正負の2つに場合分けをする．

$$f = \begin{cases} r\hat{f} & \text{if } (p < 0.0) \\ -r\hat{f} & \text{if } (p \geq 0.0) \end{cases} \tag{3}$$

上記，$\hat{\phi}$，$\hat{f}$ の初期値と $r = 5.0$ は経験的に求められた値を用いた．

2.2.2 実験結果および考察

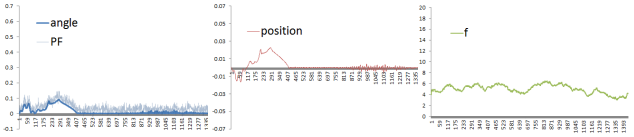

図 2: 角度 ϕ　図 3: 振り子位置　図 4: 推定値 $\hat{f}$

図 2 は，濃い太線が実際に観測された ϕ の値，薄い細線はパーティクルフィルタによる ϕ の推定値を表した．これより，ϕ の推定が大きく外れていないことがわかる．振り子の位置を示す図 3 から，外乱により振り子が大きく傾いたが，外乱を乗り越えて最終的に安定化が図れたことが伺える．図 4 は，$\hat{\phi}$ と尤度を共有して得られた $\hat{f}$ の絶対値である．これにより，実際に観測されない値も提案手法を用いることにより推定結果を導出できることが示された．

3 次元圧縮による転移を用いた効率的な行動学習

3.1 提案手法

新しい環境において，エージェントの学習回数の削減を目指す為に，転移学習が行われる [3]．しかし，強化学習で獲得した知識を転移する際に，どの知識を転移させるかを定義するのは難しい．そこで，本研究ではソースタスクで強化学習により獲得した転移知識の選択手法にスパースコーディングを導入し，効率的な転移を目指す．スパースコーディングの最適化式は以下で示される．

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1. \quad (4)$$

式 (4) において $\boldsymbol{y}$ は入力信号，$\boldsymbol{D}$ は辞書と呼ばれる基底の集合，$\boldsymbol{x}$ は $\boldsymbol{y}$ を基底の線形和で表現した際のそれぞれの基底に対応する係数行列である．ソース・ターゲットタスク共に, 5 色のコスト付き (白:0, 青: -2, 緑:-3, 赤:-5, 黒:-10) の迷路 (縦:30, 横:30) をタスクとした．まず最初に, 複数のソースタスクで Q-learning を行い, それぞれのタスクにて 900 マス分のマス目コストと Q 値を獲得した．以下に提案手法の手順を示す．

step1 ターゲットタスクの現在探索しているマス目を含む周囲 25 マスのマス目コストを取得し, $\boldsymbol{y}$ に代入する

step2 $\boldsymbol{D}$ にも同様にソースタスクの 25 マス分のマス目コストを 1 基底として代入し, ターゲットの状態と類似しているソースの状態をスパースコーディングで算出する

step3 step2 の結果, $\boldsymbol{x}$ が非 0 の基底の添字に対応する部分の Q 値を $\boldsymbol{x}$ の割合で総和をとる

step4 step3 で求めた Q 値を, step1 で現在探索しているマス目へ転移する Q 値とする

step1〜step4 をターゲットタスク上の探索で繰り返すことにより (図 5), オンラインで Q 値を転移し, 学習効率の向上を図る．

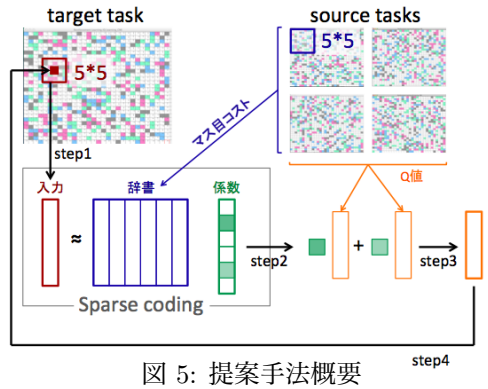

図 5: 提案手法概要

3.2 実験

3.2.1 実験設定

本実験では, ソース・ターゲットタスク共に, 900 マスのコスト付き迷路を実験対象とし, スタートを左上, ゴールを右下に設定した．また, エージェントは上下左右に 1 マスずつ移動できるものとした．それぞれのタスクでは, マス目コストをランダムに再配置し, 新しい環境を用意した．このようなターゲット環境で, 100 エピソード繰り返し, それぞれのエピソードに必要としたステップ数とコスト量を評価した．Q-learning と提案手法 (ソースタスク 4 個分，10 個分，k-means を用いて 10 個分の基底数をを 4 個分にしたもの) を比較した．

3.2.2 実験結果および考察

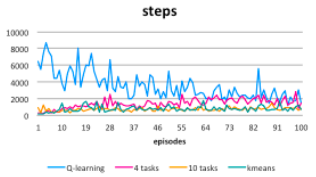

図 6: ステップ数

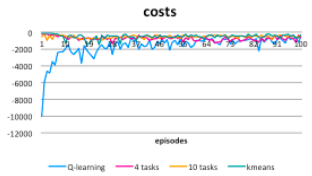

図 7: コスト量

図 6, 7 に結果を示した．ここで，縦軸はそれぞれ 1 エピソードにかかったステップ数とマス目コスト, 横軸は 100 回分のエピソードを示している．青のグラフが Q-learning，赤が 4 個分，橙が 10 個分，緑が k-means の結果を示している．これより，提案手法により Q-learning よりも探索回数とマス目コストを抑え，効率の良い学習が行えることが分かった．また，辞書が大きい程転移精度が向上し，k-means を用いて基底数を制限しても転移精度を維持できたことを確認した．

4 おわりに

本研究では，複数のパーティクルフィルタを用いたシステム制御手法を提案し，倒立振子安定化制御を通して，運動方程式による制御では解析的に近似困難な環境外乱に対し，提案手法が頑健であることを示した．また，強化学習で獲得した転移知識の選択手法にスパースコーディングを導入する手法を確立し，新しいタスクで学習しなおすよりも転移学習の効率化に成功した．実験では，辞書の大きさによる比較を行い，その中で k-means 法を用いた辞書基底数の制限をしたが，転移精度を維持することができた．今後の課題として，提案した手法の精度向上を目指したいと考えている．

参考文献

[1] Olshausen, B.A, and Field, D.J. : Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381:607-609, 1996.

[2] D.Stahl and J.Hauth, : PF-MPC:Particle Filter-Model Predictive Control, Berichte des Fraunhofer ITWM, Nr.201, 2011.

[3] Haitham B. Ammar, Karl Tuyls, Matthew E. Taylor, Kurt Driessens, and Gerhard Weiss, : Reinforcement Learning Transfer via Sparse Coding, AAMAS 2012, 4-8, 2012.