2A1-3

# 強化学習のためのParticle Filterを用いた連続行動空間表現

A Continuous Action Space Representation by Particle Filter for Reinforcement Learning

柏村 洋平 Yohei KASHIMURA　上野 敦志 Atsushi UENO　辰巳 昭治 Shoji TATSUMI

大阪市立大学大学院 工学研究科 電子情報系専攻
Department of Physical Electronics and Infomatics, Graduate School of Engineering, Osaka City University

Reinforcement Learning is a kind of machine learning. We know the past learning algorithms handled many applications, and most of them have states and actions in discrete spaces. But, more applications in real world have gotten a lot of attention recently. In such application, learning algorithms often have to handle continuous state and action spaces. In a simple way, we can handle continuous spaces by discretization. However, the simple discretization causes some problems which make learning over a finite period of time become hard or impossible. We propose a method which handles continuous action spaces efficiently. The method presents stochastic policy with particle filter. We demonstrate it through a task of swinging up a pendulum.

## 1. はじめに

強化学習（RL: Reinforcement Learning）とは，ある環境におけるエージェントが，その環境と相互作用しながら学習を行う機械学習の一種である [2]．強化学習は，有限の状態と有限の行動を持つゲームの戦略などの有限マルコフ決定過程における問題を解くために適用されてきたが，近年では，より複雑な状態と行動を持つ実世界へと，その適用範囲を広げる研究が進められている．特に，ロボットの運動制御や自律制御の分野での研究が活発になっており，歩行の獲得を代表とする複雑な制御問題にも適用されている．しかし，実世界における問題は，連続した時間が流れていることや，状態や行動の数が無限であること，環境に多くのノイズが含まれることで，有限マルコフ決定過程における問題に比べると，非常に難しい問題であるといえる．本論では, 連続空間における強化学習問題を対象とし，その特徴を挙げ，問題点とアプローチを説明する．特に連続行動空間の扱い方について，Particle Filter を用いることで，エージェントの経験を基に表現された確率分布から行動を選択する手法を提案する．提案手法は，空間が連続であることを考慮したアルゴリズムとなっており，従来のテーブル形式で行動の評価値を扱う学習法に比べ，高速な強化学習法である．この提案手法の性能評価を行うために，連続空間を扱う学習タスクとして，振子の倒立安定化問題について実験行った．

## 2. 連続空間における強化学習

過去に提案されてきた多くの強化学習法は，離散状態空間，および離散行動空間を扱ってきたが，近年では，連続空間を扱うことに研究の関心が高まっている．特にロボットなどの実世界の環境を考えると，センサ入力やアクチュエータへの出力などは実数値を持つことが多く，連続空間を扱う強化学習法が必要とされる．

### 2.1 連続空間を扱う場合に生じる問題

従来の強化学習法では，離散状態空間および，離散行動空間を扱う場合，状態，あるいは状態行動対の１つに１エントリが対応するようなテーブル形式での推定価値関数を扱ってきた．このようなテーブル形式を必要とする強化学習法を，連続空間に適用する最も単純な方法は，連続空間の離散化であるが，離散化することで，様々な問題が生じることが分かっている．連続状態空間と連続行動空間の離散化で共通して問題になる点は，

- テーブルを格納するためのメモリ量
- テーブルを正確に埋め尽くすために必要な時間とデータ量
- 量子化雑音，精度の消失

である．ここで，三つの点を挙げたが，一つ目のメモリ量の問題は，ハードウェアの問題であり，移動ロボットなどで強化学習を行う場合には，問題になる可能性があるが，近年のハードウェア，特に記憶媒体の発展をみると，それほど案ずる必要は無いかもしれない．といえるのも，二つ目の点の方が，ずっと問題であるからだ．テーブルの大きさが大きくなればなるほど，学習にかかる時間が増加する．例えば，価値関数は，各状態を繰り返し経験することで真の価値に近づくが，テーブル上のある１エントリの状態に極めて低い確率でしか到達しない場合，そのエントリを十分に信頼できるものにするためには，他の状態をその何倍もの回数経験しなければならない．次に，三つ目の精度の問題は，計算機により離散化を行う際に必ず生じる問題である．連続値を離散化するため，状態空間の場合は，異なる特徴をもつ状態を混同することがあり，不完全知覚問題 [1] の一つ原因となる．また，行動空間の場合は，微小な出力の差により，その結果から得られる報酬が大きく異なることがあるかもしれない．この問題は先の２つの問題とトレードオフの関係にあり，単純に離散化の粗さを変えることで，どちらかを優先して解決しようとすると他方の問題が顕著になる．

### 2.2 連続空間を扱う強化学習法

強化学習で連続状態空間を扱う手法として，関数近似を用いることが考えられている．状態を，その特徴を基に類似した状態をまとめて扱うことで，学習に用いる状態数を減らすことができる．また，強化学習で連続行動空間を扱う場合，離散行動空間を扱う場合と異なり，学習エージェントの政策が無限に存在する連続値の中から行動を選択しなければならない．連続状態空間上の入力値は受動的なものであるが，連続行動空間上の出力値は，エージェントが自ら探索して決定する必要がある．

連絡先: 柏村 洋平, 大阪市立大学大学院 工学研究科 電子情報系専攻 知識情報処理工学研究室, 〒558-8585 大阪市住吉区杉本, TEL/FAX:06-6605-2778, yohei@kdel.info.eng.osaka-cu.ac.jp

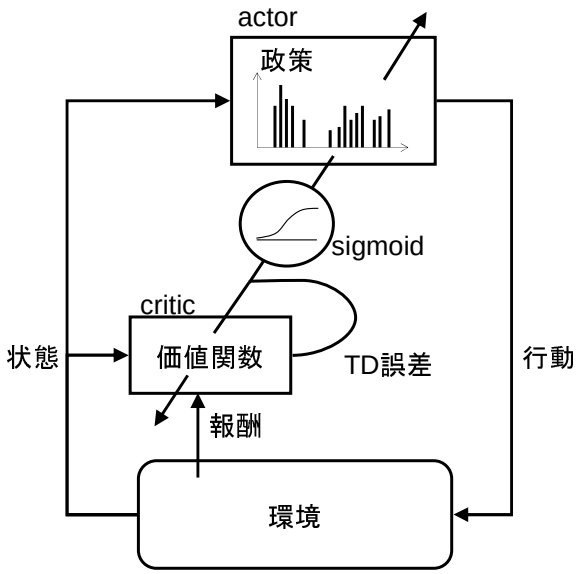


図 1: 提案手法の枠組み

あらかじめ決められた多数の行動の選択肢を効率よく学習するために，行動間の距離を考慮に入れたアルゴリズムとして，確率的二分木を用いる手法 [4] などが提案されているが，そのような手法では，学習速度は改善されるが，選択肢の間にある，より価値の高い行動を選択することが不可能である．そこで，連続行動空間を扱う強化学習では，その政策を確率分布関数とする手法が用いられる [5]．これまで提案されてきた確率的政策には，正規分布を用いたものが多く，正規分布の重ね合わせによる政策 [6, 7] や，複数の正規分布を重み付けして選択する手法 [3] が存在する．

## 3. 提案手法

提案手法では，TD 学習の一種である **Actor-critic 法**を用いる．Actor-critic 法は，状態の価値関数と独立に政策表現の構造を持つ手法である．この政策部分が行動選択に用いられることから “actor” と呼ばれ，actor が選択した行動を評価して価値関数を推定する部分は “critic” と呼ばれる．critic は価値推定を行い，行動の実行結果が期待されたものよりも良かったかを示す TD 誤差を出力する．提案手法では，TD 誤差を受け取って Particle Filter による確率的政策を構築する actor を提案する．Particle Filter による連続行動空間表現とは，行動 $a$ を実行したとき，その評価値を基に算出した重み $w$ の particle を，行動空間上の $a$ に配置することである．状態ごとに複数の particle で分布を形成することで，エージェントの確率的政策を表現する．図 1 にその枠組みを示した．

### 3.1 初期化

各状態 $s$ に対して，以下の手順で Particle Filter を構築する．リストの長さ（particle の個数）は，あらかじめ決めておき $n$ とする．particle のリスト $\{(a^i, w^i)\}_{i=1}^n$ を生成し，次式で初期化する．

$$a^i = random(a_{min}, a_{max}) \quad (1)$$

$$w^i = 0.5 \times \gamma_w^{n-i} \quad (2)$$

ただし，$random(a, b)$ は範囲 $[a, b]$ の一様乱数返す関数とし，$a_{min}, a_{max}$ は，それぞれ，状態 $s$ でとりうる行動の最小値と最大値とする．また $\gamma_w (0 < \gamma_w \leq 1)$ は，古い particle への割引率である．

### 3.2 価値関数の更新

critic による価値関数の更新方法は，一般的な TD 学習の形式を用いる．状態 $s_t$ で，行動 $a_t$ を実行したときに，報酬 $r_{t+1}$ を受け取り，次の状態 $s_{t+1}$ に遷移したとすると，TD 誤差 $\delta_t$ は，以下のように定義される．

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s) \quad (3)$$

この値を用いて価値関数は次式で更新される．

$$V(s_t) \leftarrow V(s_t) + \alpha \delta_t \quad (4)$$

ここで，$\gamma$ は割引率，$\alpha$ はステップサイズパラメータである．

### 3.3 行動選択

状態 $s_t$ における particle のリスト $\{(a^i, w^i)\}_{i=1}^n$ から，その重み $w^i$ に従ってルーレット選択により $(a^k, w^k)$ を選択する．次に，以下の式に従い，出力 $a_t$ を決定する．

$$\epsilon = \frac{\sigma}{w^k}(a_{max} - a_{min}) \quad (5)$$

$$a_l = max(a_{min}, a^k - \epsilon) \quad (6)$$

$$a_r = min(a_{max}, a^k + \epsilon) \quad (7)$$

$$a_t = random(a_l, a_r) \quad (8)$$

ただし，$\sigma$ は，初期値 0.5 とし，時間で減少する定数である．ここで，$\epsilon$ は探索範囲であり，重み $w^k$ が大きいほど探索範囲が狭くなり，逆に $w^k$ が小さいほど探索範囲が広くなる．また，時間経過による $\sigma \rightarrow 0$ により，探索範囲も $\epsilon \rightarrow 0$ となる．

### 3.4 政策の更新

Particle Filter により表現された政策は，critic により出力された TD 誤差 $\delta_t$ を用いて，以下の手順で更新される．まず，$\delta_t$ をシグモイド関数に入力し，その出力を重みとする．

$$w_t = \frac{1}{1 + exp(-\beta \delta_t)} \quad (9)$$

ただし，$\beta$ は定数である．シグモイド関数を用いる事で，出力の値域が $(0, 1)$ になる．次に，状態 $s_t$ に対応するリストの各 particle の重みを減少させる．

$$w^i \leftarrow w^i \times \gamma_w \quad (10)$$

最後に，particle のリストから先頭の particle を除去し，新しい particle $(a_t, w_t)$ をリストの末尾に追加する．

## 4. 実験と考察

シミュレーションにより，提案手法の特徴と性能を評価した．まず，状態数 1 の，報酬が行動の価値に一致する単純な環境で提案手法の確率的政策の変化を検証した．次に，振子の倒立安定化問題について従来の固定の離散化方式との比較を行った．

### 4.1 単純な問題を用いた特徴の検証

提案手法により Particle Filter を用いて表現された確率的政策が，連続行動空間上でどのような分布を持ち，学習が進むとどのように変化するかを調べるために，状態数 1 で，状態遷移が起こらない（ループバックする）最も単純な環境で実験を

表 1: テスト関数の学習に用いたパラメータ

| | |
|---|---|
| 価値関数のステップサイズパラメータ（$\alpha$） | 0.4 |
| TD 誤差更新時の割引率（$\gamma$） | 0.9 |
| particle リストのサイズ（$n$） | 50 |
| $\sigma$ の割引率（$\gamma_\sigma$） | 0.99 |
| シグモイド関数への入力値を調整する定数（$\beta$） | 1.0 |

行った．報酬は，行動からの射影として，次式で与えられるテスト関数を用いた．ただし，とりうる行動は $[-50, 50]$ とした．

$$f(x) = \begin{cases} 0 & \text{if } x < -9 \\ x^2 & \text{if } -9 \leq x \leq 10 \\ 0 & \text{if } 10 < x \end{cases} \tag{11}$$

この関数は，図 2 で示されとおり，以下の 3 つの特徴を持つ．

- 局所最適解が 2 つ存在する
- 最適解の点で不連続（すぐ隣で値 0 になる）
- 行動空間の大部分で報酬 0 が連続する．

一つ目の特徴は，山登り法を用いると，$x = -9$ 付近の行動を学習してしまう可能性がある問題であるといえる．それに加えて二つ目の特徴により，単純な正規分布の当てはめが困難であるといえる．また，三つ目の特徴は，探索範囲中で報酬が得られる部分が限られているため，局所的に集中して探索することが有効であるといえる．

表 1 で示される学習パラメータを用いて，1000 ステップの学習を行った結果，図 3 に示される学習曲線が得られた．得られる報酬の最大値は，$x = 10$ のときで，$f(10) = 100$ となることから，その付近の解が学習されたことが分かる．また，開始時刻 $t = 0$ から 100 ステップごとに Particle Filter で表現されている行動の確率分布を算出し，その時間変化を図 4 に示した．その結果のグラフより確率的政策が以下のように変化していく様子が確認できる．

1. 開始時刻 $t = 0$ では，選択確率が一様分布になっており，偏りなく行動が選択される．
2. $t = 100, 200$ で報酬を獲得できない領域の選択確率が低くなり，報酬獲得につながる領域に集中する．
3. $t = 300, 400, 500$ で二つの局所解に，それぞれ選択確率が集中する．
4. $t = 600, 700$ で価値の高い解へと収束する．

このように，提案手法の確率的政策の，知識利用と環境探索のトレードオフに対するアプローチを確認できた．つまり，単純な山登りを行わずに，良い結果が期待される行動を重点的に探索する様子が確認できた．

### 4.2 倒立振子の安定化問題

提案手法の性能を評価するために，振子の倒立安定化問題の実験を行った．実験は図 5 のような振子についてシミュレーションで行った．振子の運動方程式は以下のようになる．

$$m\ddot{x} = -mgl\sin x - k\dot{x} + T \tag{12}$$

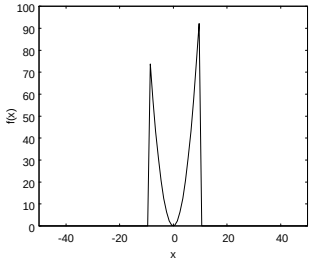


図 2: テスト関数

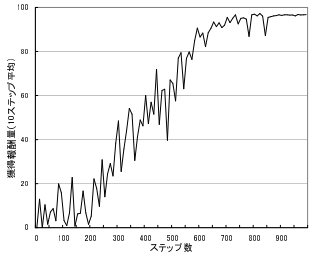


図 3: テスト関数の学習曲線

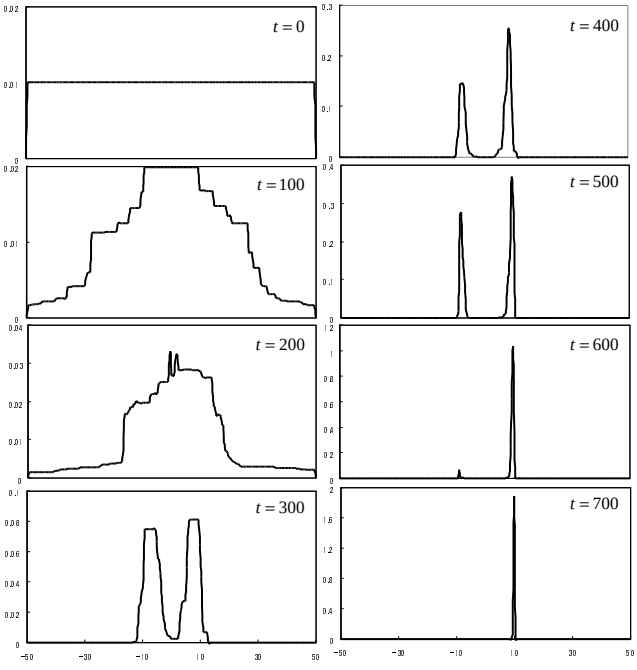


図 4: 確率的政策の変化

ただし，回転軸に重さ $0(kg)$，長さ $l(m)$ のリンクがつながり，その先に重さ $m(kg)$ の重りが付いているもとする．振子の角度を $x(rad)$ とし，その角速度を $\dot{x}(rad/s)$，角加速度を $\ddot{x}(rad/s^2)$ とする．$k$ は速度にかかる抵抗の係数とし，$T(N)$ は回転軸でリンクに加えるトルクである．実験では，$m = 1(kg)$，$l = 1(m)$，$k = 0$ とし，重加速度 $g = 9.8(m/s^2)$ とした．状態は，振子の角度 $x(-\pi < x \leq \pi)$ と角速度 $\dot{x}(-3\pi \leq \dot{x} \leq 3\pi)$ であり，出力はトルク $T(-20 \leq T \leq 20)$ である．エージェントは 0.2 秒毎に，現在の状態に対する報酬を受け取り，次の 0.2 秒間に出力するトルクを変更できる．シミュレーションは，Runge-Kutta 法（4 次）を用いて，エージェントの 1 ステップあたりに 4 回，つまり 0.05 秒毎に運動方程式を逐次計算して行った．学習の目標は，出力トルクの量を抑えつつ，振子を倒立状態で安定させることであり，この目標を達成させるために，時刻 $t$ での報酬を次式で設定した．

$$r_t = \frac{|x|}{\pi} - 0.5\frac{|T|}{20} \tag{13}$$

これは，振子の角度が倒立状態（$x = \pi$）に近いほど高い報酬を与え，トルクの出力量に応じた罰（負の報酬）を与えるものである．この環境で初期状態 $x = 0$，$\dot{x} = 0$ から 20 秒間を 1 エピソードとして繰り返し学習させた．この実験では，簡単

表 2: 倒立振子の実験に用いた Q-Learning のパラメータ

| | |
|---|---|
| ステップサイズパラメータ（$\alpha$） | 0.25 |
| 割引率（$\gamma$） | 0.9 |
| ボルツマン選択の初期温度 | 1.0 |
| ボルツマン選択の収束温度 | 0.001 |

表 3: 倒立振子の実験に用いた提案手法のパラメータ

| | |
|---|---|
| 価値関数のステップサイズパラメータ（$\alpha$） | 0.4 |
| TD 誤差更新時の割引率（$\gamma$） | 0.9 |
| particle リストのサイズに（$n$） | 30 |
| $\sigma$ の割引率（$\gamma_\sigma$） | 0.99999 |
| シグモイド関数への入力値を調整する定数（$\beta$） | 250 |

のため，状態空間の扱い方について，角度 $x$ と角速度 $\dot{x}$ についてそれぞれの定義域を等間隔に 50 分割することで離散化を行った．

各状態に 30 個の particle を持つ提案手法と，離散環境で優れた性能を持つ強化学習法である Q-Learning を，行動空間を等分割により，5 分割，30 分割，100 分割して適用した手法とで実験を行った．実験は，学習が十分に収束するように 20000 エピソード繰り返し行い，それぞれの学習を 10 回行ったものの平均を評価した．実験で示した Q-Learning には，ボルツマン選択を用い，その温度の割引率を，各分割数の学習が収束したときの 1 エピソードあたりの報酬量が大きく，かつ出来るだけ早く収束したものを選んだ．それぞれのパラメータを表 2 に示す．実験に用いた提案手法のパラメータを，表 3 に示す．

その結果，図 6 で示される学習曲線が得られた．

これより，行動を等分割してテーブル形式で持つ Q-Learning では，分割数を大きくすると，より細かい離散化を行うため，最終的に得られる報酬獲得量が大きくなるが，学習の立ち上がりが非常に遅くなっている．これは，テーブルの大きさが非常に大きくなるため，そのエントリが十分に信頼できるものになるまでに時間がかかりすぎているためであると考えられる．それに比べ，提案手法は学習の立ち上がりが早い点と，収束した 1 エピソードあたりの獲得報酬量の両方の点で，従来の Q-Learning より優れているといえる．これは，提案手法が連続行動空間の特徴をうまく利用して探索を進めることができているため，学習の立ち上がりが早く，また，選択可能な行動が連続空間上に無限に存在するため，固定の離散化の隙間に存在する価値の高い行動を実行できていると考えられる．

## 5. おわりに

実世界の環境で強化学習を行うためには，学習速度が非常に重要な要素であり，従来の強化学習法で連続空間上にある学習問題を扱う際に，いくつかの問題点が生じることを指摘し，連続行動空間における行動を学習するための手法として Particle Filter を用いた確率的政策表現によるアプローチを提案した．そして，連続空間を扱う問題に提案手法を適用し，その特徴を調べ，性能評価を行い，先に指摘した学習速度や探索能力において提案手法が優れていることを示した．今後の課題は，正規分布を用いた他の確率的政策を持つ手法との性能比較や，連続状態空間を関数近似で扱う手法との組合せ，多次元行動空間への適用など，様々な環境の実験を試行することが挙げられる．また，学習パラメータの理論的な設定方法や，適格度トレースの導入などにより，提案手法を改良することが考えられる．

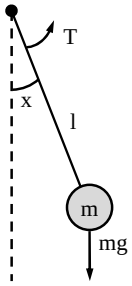


図 5: 振子

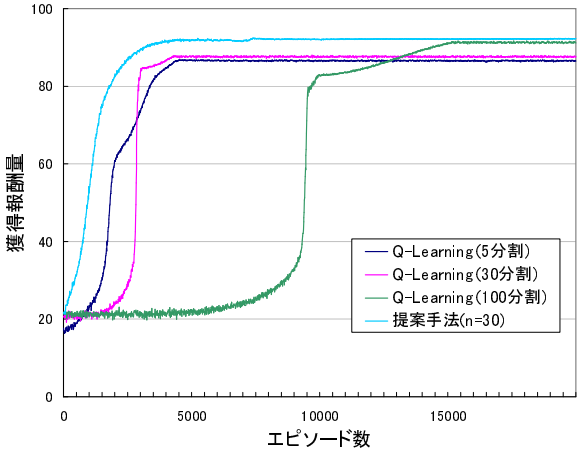


図 6: 倒立振子の安定化問題での性能比較

## 参考文献

[1] Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *National Conference on Artificial Intelligence*, pp. 183–188, 1992.

[2] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[3] 木村元, 荒牧岳志, 小林重信. 重み付けされた複数の正規分布を用いた政策表現. 人工知能学会論文誌, Vol. 18, No. 6, pp. 316–324, 2003.

[4] 木村元, 小林重信. 確率的 2 分木の行動選択を用いた antor-critic アルゴリズム - 多数の行動を扱う強化学習 -. 計測自動制御学会論文集, Vol. 37, No. 12, pp. 1147–1155, 2001.

[5] 木村元, 小林重信. ロボットの強化学習における状態-行動空間の汎化. 日本ロボット学会誌, Vol. 22, No. 2, pp. 161–164, 2004.

[6] 森本淳, 銅谷賢治. 強化学習を用いた高次元連続状態空間における系列運動学習：起き上がり運動の獲得. 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理, Vol. 82, No. 11, pp. 2118–2131, 1999.

[7] 吉本潤一郎, 石井信, 佐藤雅昭. 連続力学システムの自動制御のためのオンライン em 強化学習法. システム制御情報学会論文誌, Vol. 16, No. 5, pp. 209–217, 2003.