1D3-02

連続値入力に対応したProfit Sharingに基づく強化学習システム

Reinforcement Learning Systems Based on Profit Sharing in Continuous State Spaces

宮崎和光[*1] 小林重信[*2]
Kazuteru Miyazaki Shigenobu Kobayashi

[*1]独立行政法人大学評価・学位授与機構 National Institution for Academic Degrees and University Evaluation
[*2]東京工業大学 Tokyo Insititute of Technology

Reinforcemen Learning is a kind of machine learning. We know Profit Sharing [Miyazaki 94], Rational Policy Making algorithm (RPM)[Miyazaki 99] and PS-r* [Miyazaki 03] to guarantee the rationality in a typical class of Partially Observable Markov Decision Processes. However they cannot treat continous state spaces. In this paper, we describe a solution how to adapt them to the environment that has continous sensory inputs. We give RPM a mechanism to treat continous sensory inputs. We show the effectiveness in numerical examples.

1. はじめに

著者らはこれまでいくつかの Profit Sharing(PS) に基づく強化学習システムを提案してきた [宮崎 94, 宮崎 99, 宮崎 03]. そこでは, つねに離散の入力を前提としている. しかし実世界には状態間に位相が仮定できる問題が多く存在する. 本稿では, そのような問題に対し, PS に基づく強化学習システムを拡張するための方法を提案する. これにより,PS に基づく強化学習システムの実問題への応用可能性を広げるものと考える.

2. 問題設定

通常, 計算機上では, 連続値で入力された情報は, 何らかの形で離散化が施される. 離散化された入力の種類のことを**状態数**と呼ぶ. 粗い離散化が行われた場合, 状態数は少なくて済むが, いわゆる**不完全知覚問題** [Chrisman 92] が生じやすくなる. 一方, 細かい離散化がなされれば, 不完全知覚問題は生じにくくなるが, 状態数が多くなる. このことは, 一般に, 学習により多くの時間を要することを意味する. よって, 理想的には, 入力信号は不完全知覚問題が生じない程度に粗く離散化されることが望ましい.

本稿では, 環境からの**感覚入力**に対し, **行動**を選択し, 実行に移す学習器を想定する. 一連の行動に対して, 環境から**報酬**が与えられる. 以下では, 正の報酬のみを扱い, 負の報酬である**罰**の存在は考えない. また報酬は 1 種類のみとする. これは,Profit Sharing (PS) [宮崎 94], 合理的政策形成アルゴリズム (RPM) [宮崎 99], PS-r* [宮崎 03] などの報酬設定と一致する. 時間は認識-行動サイクルを 1 単位として離散化される. 感覚入力は離散的な属性の種類ごとに連続値で与えられる. 離散的な属性の種類を**次元数**と呼ぶ. 行動は離散的なバリエーションの中から選ばれる.

学習の目的は,「報酬が得られないループに陥らない」こととする. ここで, **報酬が得られないループ**とは [宮崎 94] において定義されている**迂回系列**のことであり, 迂回系列に陥らないことは, [宮崎 94] における**合理性**の追求と同じことを意味する. これは, たとえ, 不完全知覚が生じていたとしても合理性が維持されているならば, それでよしとする立場である.

以上より, 本稿の目的は,「(状態間に位相が仮定できる) (連続値で入力される) 感覚入力を合理性が維持されるように離散化する」手法を提案することとなる.

連絡先: 宮崎和光, 独立行政法人 大学評価・学位授与機構, 〒187-8587 東京都小平市学園西町 1-29-1, TEL 042-353-1834, FAX 042-353-1861, teru@niad.ac.jp

3. 連続値入力に対応したPSに基づく強化学習システムの提案

3.1 感覚入力の離散化

ある時間ステップ t ごとに, 連続値で与えられる感覚入力を離散化することを考える. 以下では, 離散化された感覚入力を**状態**と呼ぶ. t は状態数の上限値を目安に決めるパラメータである[*1].

まず, 時刻 t での感覚入力を S_t とすると, S_t から S_{t+1} に遷移した時点で, S_t を中心とする n 次元正規分布を張る (図 1 参照). その正規分布の主軸方向を $\overrightarrow{S_{t+1}} - \overrightarrow{S_t}$ で定義する (図 1 の d_1 軸). 主軸以外の方向は, 各々が直交するようにグラムシュミットの直交化法などを用いて生成する (図 1 の $d_2, , , d_n$ 軸). 主軸の裾野の広さは, $3\sigma_1 = |\overrightarrow{S_{t+1}} - \overrightarrow{S_t}|$, 主軸以外の裾野の広さは, $3\sigma_i = \frac{|\overrightarrow{S_{t+1}} - \overrightarrow{S_t}|}{3\sqrt{n}}$ $(i = 2, 3, ..., n)$ とする. ここで主軸以外の方向の裾野の広さに $\frac{1}{\sqrt{n}}$ を乗じているのは [Kita 98] の知見を利用したためであるが, $\frac{1}{3}$ は経験的に裾野の広さを狭くするために乗じている.

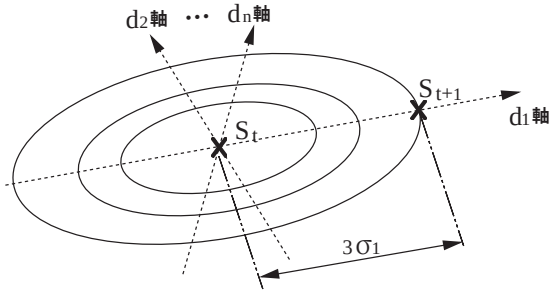

図 1: 感覚入力の離散化.

最後に正規分布の高さの最大値が 1 になるように変換する. n 次元空間内の各軸上の値が $d_i (i = 1, 2, ..., n)$ で与えられたとき,

*1 4 章の実験では意思決定の時間ステップと同一とした.

求める $f(d_1, d_2, ..., d_n)$ は, 以下の式で与えられる [統計学 92].

$$f(d_1, d_2, ..., d_n) = e^{-\frac{1}{2}(\frac{d_1^2}{\sigma_1^2} + \sum_{i=2}^{n} \frac{d_i^2}{\sigma_i^2})}, i = 2, 3, ..., n. \quad (1)$$

3.2 連続値入力に対応した RPM

(1) 式を利用して連続値入力を離散化することを考える. 本稿では, その一例として, RPM との組み合わせを与えるが, PS や PS-r*との組み合せも比較的容易に行える. また, 報酬と罰が混在するより一般的な場合に対しては, 罰回避政策形成アルゴリズム [宮崎 01] 等との組合せで対応可能である.

RPM は, 新規に経験した状態ではランダムに行動を選択し, その選択した行動を次々メモリ上の各状態に対応した場所に上書きする手法である. そして, 報酬を得た時点で, メモリ上の各行動を政策に登録し, 政策が存在する状態では, 必ず, その政策が示す行動を選択する.

RPM との組合せを考えた場合, 最初の報酬を得るまでと, 報酬を得た後では処理が異なる. そこで, 以下では, これらを分けて説明する.

3.2.1 最初の報酬を得るまでの処理

最初の報酬を得るまでは, 通常の RPM 同様, 行動をランダムに選択する. そして, 選択した行動を次々メモリ上の各状態に対応した場所に上書きする. この際, どの状態に上書きすべきかを決定しなければならない. ここでは, この決定に, 感覚入力と記憶している状態 (の分布) との間の距離を利用する. その際, 各状態の分布には, その分布の影響の及ぶ範囲を限定するためのパラメータ **f_para** (0.0 <f_para≤ 1.0) を付与し[*2] 近さの判定に活用する (図 2 参照). 図 2 からもわかる通り, f_para が大きくなると, その状態がカバーする範囲が狭まる.

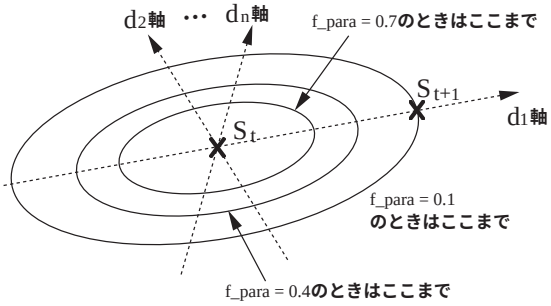

図 2: f_para の意味.

まず, 観測されている感覚入力の各分布上に対する (1) 式の値を計算し, その値が f_para 以上である分布の中で最も値が大きい状態のメモリ上に, そのとき選択した行動を上書きする[*3]. そして, 報酬を得たときには, 通常の RPM 同様, 政策を形成する.

図 3 を用いて感覚入力を既知の状態とマッチングさせる具体例を示す. まず, 中央の「x」という座標に相当する感覚入力 (連続値) が n 次元ベクトルで与えられたとする. このとき, この座標が既知の 2 種類の状態 (状態 A および B) の分布と交わる座標の (1) 式の値は, それぞれ,$f_A = 0.4$ および $f_B = 0.5$ であった. 今, 状態 A の f_para は 0.2, 状態 B の f_para は 0.4

*2 4 章の実験では f_para の初期値は 0.1 とした.

*3 状態数を縮約するという観点では, 「f_para 以上である分布の中で最も値が小さい状態のメモリ上に, そのとき選択した行動を上書きする」という方針も考えられる.

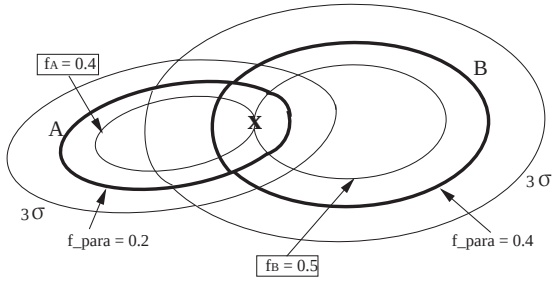

図 3: 得られた感覚入力を既知の状態とマッチングさせた具体例.

だとすれば, この f_A および f_B はともに f_para より大きな値であるので, 単純に f_A と f_B の大小関係を比較し, より大きな値を示している状態 B に「x」は属すると判断する. 一方, 状態 B の f_para が 0.6 であった場合には, $f_B = 0.5$ は f_para よりも小さな値となり, 「x」は状態 B のカバーする範囲外とされ, 状態 A に属すると判断される.

3.2.2 報酬を得た後の処理

報酬を得た後には, より確実に合理性を保証するために, 極力, ランダムな行動選択を排除する. そのためには, まず, 感覚入力を得た後に, その感覚入力が既知のいずれの状態に最も近いかの判定が必要となる. この判定には, 3.2.1 節で述べた方法と同様に, f_para を参照し, 各状態の分布から返される (1) 式の値で判定する.

行動は, 近いとされた状態 (f_para 以上である分布) の中に政策が存在すれば, その政策に従った行動を実行する. 近い状態がない, もしくは, 政策が存在しない場合には, ランダムに行動を選択する. この場合, 政策の形成方法は,3.2.1 節と同様である.

もし, 現在得られている政策で報酬が得られるならば, 合理性が維持されていると言えるので, その政策を変える必要はない. 一方, 報酬が得られない (ある行動数[*4] 経過しても報酬が得られない) ならば, 合理性が維持されていないことを意味するので, ループに陥っていると判断し, ループ内の各状態に関する分布の f_para を大きくする[*5]. この処理により, 不完全知覚が減少する. ここで, 記憶している各状態に対応した分布が小さくなるわけではなく, 分布の利用範囲を決めるパラメータ f_para が大きくなることで, その状態の利用範囲が狭まる点に注意されたい.

3.3 その他考え得る工夫

- 状態数の削減
  例えば, 各状態の使用 (経験) 頻度をモニターし, 頻度の低い状態を削除することで状態数の削減は可能である.

- より積極的な不完全知覚への対応
  PS-r*的な考えを導入し, より積極的に不完全知覚に対応することも可能である. 但し, この場合, 遷移先状態に関する情報を記憶する必要がある. 具体的には, 状態数を S, 行動の種類を A とすれば, 記憶容量が, 現状の $O(AS)$ から $O(AS^2)$ に変化する.

*4 4 章の実験では最初に報酬を得たときの行動数の 5 倍の行動数.

*5 4 章の実験では +0.1 とした.

4. 数値例

4.1 実験環境

提案手法の有効性を調べるために, 図 4 に示した 2 次元 (x および y) 平面環境に適用した. 座標 $(x, y) = (0.0, 0.0)$ が始点, $(0.9, 0.9)$ から $(1.0, 1.0)$ の部分が終点である. 終点に到達すると報酬が与えられ始点に戻される. 外枠および黒い部分には進入不可である (行動実行前に存在していた地点に強制的に戻される).

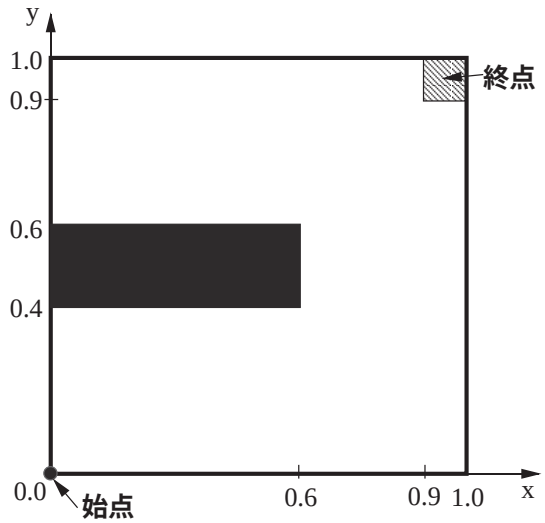

図 4: 実験で用いた環境.

行動は「停止, 上へ移動, 右へ移動, 下へ移動, 左へ移動」の 5 種類の中から選ばれる. 各方向への移動量は「0.1 の基本量」+「$-0.01/M$ から $0.01/M$ の一様乱数」で決定される. ここで,M は,∞(決定的な状態遷移),10, 4, 1 の 4 種類の実験を行った. すなわち,M を変えることで環境の不確実さをコントロール可能な環境になっている. また, 希望していない次元方向へは「-0.5/1000 から 0.5/1000 の一様乱数」で決まる量だけ移動するものとする. これは,「停止」行動でも動く可能性があることを意味する.

4.2 実験結果

始点から終点に至るまでの行動数を調べた. 実験は, 乱数の種を変えて 100 回行った. 始点から終点までの行動数の変化を図 5 に示す. 横軸は終点への到着回数, 縦軸は始点から終点までの行動数の平均値である. 図より, 環境の非決定性に対応しつつ, 政策を改善していることが見てとれる. しかしながら,$M = 1$ や $M = 4$ の場合, 結果にかなりのばらつきがみられる. これは最初に得られた報酬に強く依存した政策が形成される RPM の特徴を反映したものである. このことにより最適性を犠牲にして, その分, 素早く合理性を保証することに成功している.

このような学習の素早さを確認するために, 表 1 に終点到着回数が 1〜5 および 4000 回のときの始点から終点までの行動数の平均と標準偏差を示す. 各 M ごとに上段が平均, 下段が標準偏差である. 報酬獲得回数が 1 の場合の結果が,「学習なし」, すなわちランダムに行動を選択したときの結果であることから, 少ない報酬獲得回数の段階から急速に学習しているこ

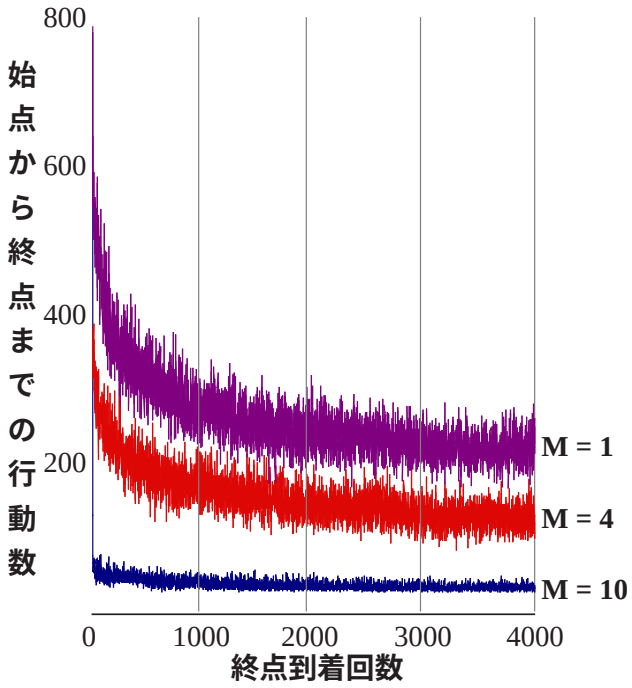

図 5: 始点から終点までの行動数の変化.

とがわかる. これは RPM の持つ学習の高速性が適切に引き継がれている結果であると考えられる.

次に状態数の変化を確認する. $M = 10$,4,1 の場合の状態数の変化をそれぞれ図 6,7,8 に示す. 横軸は終点への到着回数, 縦軸は状態数である.

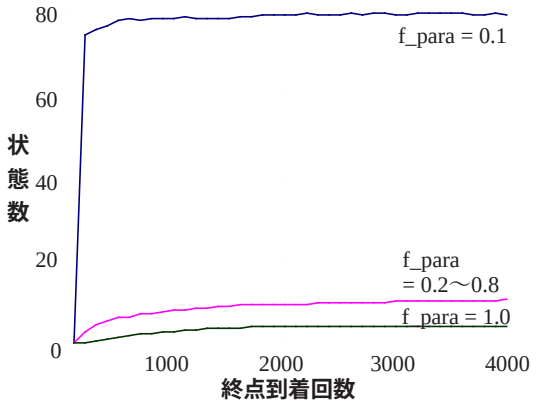

図 6: $M = 10$ の場合の状態数の変化.

各図で f_para=0.1 とある曲線は, 最初に作られたまま変化していない状態の個数を意味する. ここには, よくマッチし使われる状態と, 全く使われずに放置されたままのものが混在している. 状態の使用頻度をモニターすれば, これらの切り分けが可能になると考える.

次に,f_para=0.2〜0.8 とある曲線は, 何度かマッチしたことにより, 利用範囲が変化した状態の個数を意味する. これは, 環境の不確実さに合わせて適応的に変化した部分である. この部

表 1: 始点から終点に至る行動数の平均および標準偏差. 各 M ごとに上段が平均, 下段が標準偏差を表す.

終点到着回数	1	2	3	4	5	4000
$M = \infty$	774.11	24.42	24.42	24.42	24.42	24.42
	568.88	3.87	3.87	3.87	3.87	3.87
$M = 10$	740.62	140.34	109.36	100.10	112.79	32.14
	561.53	285.50	254.82	239.33	279.50	35.20
$M = 4$	780.74	445.98	379.31	365.91	294.63	136.53
	611.43	418.27	417.86	397.38	365.12	233.07
$M = 1$	771.08	597.12	659.08	635.01	538.95	250.48
	564.10	398.37	396.79	409.21	420.67	250.36

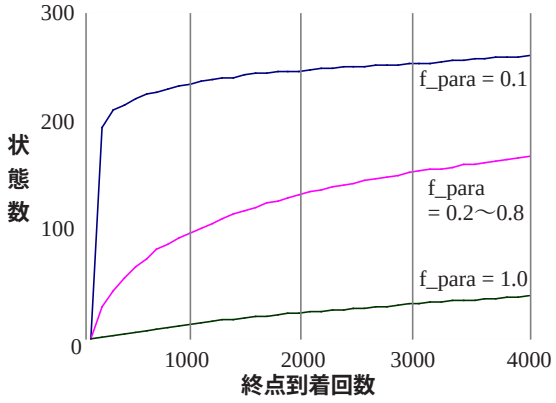

図 7: $M = 4$ の場合の状態数の変化.

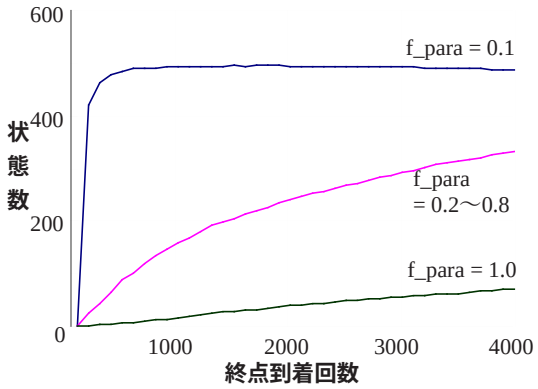

図 8: $M = 1$ の場合の状態数の変化.

分に関しても, 状態の使用頻度をモニターすれば, 重要なものと, そうでないものの切り分けが可能になると考える.

最後に,f_para=1.0 とある曲線は, f_para が 1.0 まで変化してしまった状態の個数を意味する. これは, 始点 (0.0,0.0) 以外はまず使われるのことのない状態であると考えられる. したがって, 始点以外はすべて削除可能であると思われる.

これらの図から, 環境の非決定性に応じ, 適宜, 状態が形成されているのがわかる. また,f_para=0.1 である状態の個数が比較的初期の段階で安定している. これは, 提案手法が不用意に次々状態を生成してしまう手法でないことを意味する. よって, 状態の削減機能を付加することでより効率的な状態保持が可能になると思われる.

5. おわりに

本稿では, 著者らがこれまで提案してきた PS に基づく強化学習システムのひとつである RPM を連続値入力に対応させるための手法を提案した. これにより,PS に基づく強化学習システムの実問題への応用可能性を広げるものと考える.

今後は, ランダム選択以外の他手法との比較, 倒立振子などのより現実的な問題への適用, 報酬と罰が混在する環境への拡張などを順次行う予定である.

参考文献

[Chrisman 92] L. Chrisman: Reinforcement Learning with perceptual aliasing: The Perceptual Distinctions Approach, Proc. of the 10th National Conference on Artificial Intelligence, pp. 183-188 (1992)

[統計学 92] 東京大学教養学部統計学教室編: 基礎統計学 III 自然科学の統計学, (1992)

[Kita 98] Kita, H., Ono, I. and Kobayashi, S.: Theoretical Analysis of the Unimodal Normal Distribution Crossover for Real-coded Genetic Algorithm, Proc. 1998 IEEE Int. Conf. on Evolutionary Computation, pp.529-534 (1998).

[宮崎 94] 宮崎和光, 山村雅幸, 小林重信: 強化学習における報酬割当ての理論的考察, 人工知能学会誌, Vol. 9, No. 4, pp. 580-587 (1994)

[宮崎 99] 宮崎和光, 荒井幸代, 小林重信: POMDPs 環境下での決定的政策の学習, 人工知能学会誌, Vol. 14, No. 1, pp. 148-156 (1999)

[宮崎 01] 宮崎和光, 坪井創吾, 小林重信: 罰を回避する合理的政策の学習, 人工知能学会誌, Vol. 16, No. 2, pp. 148-156 (2001)

[宮崎 03] 宮崎和光, 小林重信: Profit Sharing の不完全知覚環境下への拡張 : PS-r*の提案と評価, 人工知能学会論文誌, Vol.18, No.5, pp.286-296 (2003).