LF-007

複素評価値を用いたProfit Sharingに関する研究

Reinforcement Learning by Profit Sharing with Complex Evaluation Value

島田 慎吾 † **澁谷 長史** † **濱上 知樹** †
Shingo SHIMADA Takeshi SHIBUYA Tomoki HAMAGAMI

1 はじめに

強化学習を自律移動ロボット等の実環境で応用しようとした場合，ノイズやセンサ精度などの制約から異なる状態を同じ状態とみなしてしまい，学習が進まなくなるという問題が生じる．このような問題は不完全知覚問題と呼ばれている．過去，この問題の解決を目指した種々の手法が提案されている．しかし，アルゴリズムが環境に依存し複雑になる，多くのリソースが必要となるなどの問題が残されている [1] [2].

筆者らはこの問題に対する新たな手法として，価値関数を複素数で表す複素強化学習を検討している．この具体的アルゴリズムとして，Q-learning[3] の Q 値を複素数として扱う $\dot{Q}$-learning を提案した [4]. $\dot{Q}$-learning を用いることにより，価値の大きさだけでなく，位相情報を用いて文脈を表すことが可能になる．先の報告では，不完全知覚問題を含む迷路問題において $\dot{Q}$-learning が有効に機能することを示した．その一方で，学習パラメータがヒューリスティックスに依存する，学習速度が遅いなどの問題点が残されていた.

本稿は，これらの問題を背景に，Q-learning と比べて学習パラメータが少なく，学習速度が速いなどの利点を持つ Profit Sharing(PS) [5][6] に複素強化学習の原理を応用した手法を提案する．そして，計算機シミュレーションにより不完全知覚問題のある環境において提案手法が有効に働くことを示す.

2 複素強化学習

本研究では，不完全知覚問題に対して有効な新しいアプローチとして複素強化学習を提案している．複素強化学習では，評価値に複素数を用いることで，不完全知覚問題のある環境でも適切な動作の獲得を目指す.

これまでの強化学習では，評価値 v が実数で定義されていた．基本的により大きな評価値 v をもつルールが有効なルールとして選択されていた.

これに対し複素強化学習では，評価値 $\boldsymbol{v}$ を複素数で定義し，さらに，複素数で定義された内部状態 $\boldsymbol{i}$ を導入する．これにより，評価値 $\boldsymbol{v}$ の絶対値の大きさだけでなく評価値 $\boldsymbol{v}$ の位相と内部状態 $\boldsymbol{i}$ の位相との関係も評価してルールを選択することが可能になる．例えば，内部状態 $\boldsymbol{i}$ と各評価値 $\boldsymbol{v}$ の内積をとり，その内積値がより大きな評価値 $\boldsymbol{v}$ を持つルールを有効なルールとする．このとき，評価値 $\boldsymbol{v}$ と内部状態 $\boldsymbol{i}$ の位相関係が図 1 のような場合は，評価値 $\boldsymbol{v}_2$ より評価値 $\boldsymbol{v}_1$ を持つルールが有効なルールとして選択される．エージェントは内部状態 $\boldsymbol{i}$ に応じて行動を選択することで適切な動作を獲得できる.

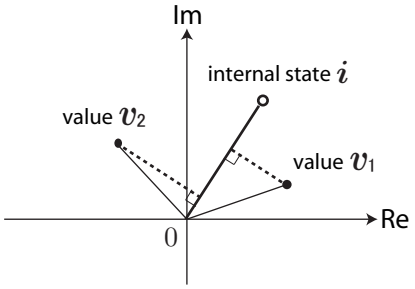


図 1 複素強化学習の基本原理

3 $\dot{Q}$-learning

Q-learning における Q 値を複素数で定義し，Q 値が複素数であることを明示的に $\dot{Q}$ 値と表す．$\dot{Q}$ 値の更新式において遷移先の状態に関連する $\dot{Q}$ 値との荷重平均を取る際に，位相回転を加えて時系列情報を含ませる．そして，直前の行動の $\dot{Q}$ 値との位相差を考慮して，次のステップで選択されるであろう $\dot{Q}$ 値を予測しながら行動を選択することで不完全知覚問題のある環境でも適切な動作を獲得する.

すでに，シミュレーション実験の結果，不完全知覚問題を含む迷路問題において $\dot{Q}$-learning が有効に機能することを確認している．しかし，学習速度が遅いことや学習パラメータ設定が環境に強く依存するという問題が残されている.

4 提案手法

提案手法は，従来の PS の評価値を複素数で定義することにより不完全知覚問題のある環境でもエージェントに適切な動作の獲得を可能にさせる．以下に，提案手法のもととなった従来の PS の概要と提案手法の詳細を示す.

4.1 Profit Sharing

PS は，初期状態から n ステップ目において，状態 s_n を観測し，行動 a_n をとり，報酬 r を得たときにエピソード単位で学習を行う手法である．このとき，各状態行動対 (s_n, a_n) に付加された評価値 $v(s_n, a_n)$ に強化値 f_n を加える．各評価値 $v(s_n, a_n)$ に加える強化値 f_n を決定する関数を強化関数と呼び，式 1 で定義する．W はエピソードの長さである．この強化値 f_n を用い，エピソード中に経験した状態行動対 (s_n, a_n) の評価値 $\boldsymbol{v}(s_n, a_n)$ を式 2 に従い更新する.

$$f_n = \begin{cases} r & (n = W-1) \\ \gamma f_{n+1} & (n = W-2, W-3, \cdots, 0) \end{cases} \quad (1)$$

$$v(s_n, a_n) \leftarrow v(s_n, a_n) + f_n \quad (n = W-1, W-2, \cdots, 0) \quad (2)$$

† 横浜国立大学大学院工学府

一般的に PS では行動選択方法としてルーレット選択を用いる．ルーレット選択では，状態 s_n における行動 a の選択確率 $P_R(s_n, a)$ を式 3 で定義する．

$$P_R(s_n, a) = \frac{v(s_n, a)}{\sum_{a' \in A(s_n)} \left(v(s_n, a') \right)} \tag{3}$$

ただし，状態 s_n における行動集合を $A(s_n)$ とする．

4.2 複素評価値を用いた Profit Sharing

本研究では，Q-learning と比較して学習パラメータが少なく，学習速度が速いなどの利点を持つ PS に複素強化学習の原理を応用する．従来の PS において実数で表されていた評価値を複素数で定義する方法を提案する．具体的には，価値の大きさを複素数の絶対値として表し，時系列情報を複素数の位相部分に持たせる．さらに，エージェントに複素数で定義された内部状態を内部変数として持たせることで不完全知覚問題のある環境でもエージェントに適切な動作を獲得させる．

4.2.1 評価値の学習

従来の PS と同様に報酬 r を受け取りエピソード単位で評価値 $\boldsymbol{v}(s_n, a_n)$ を更新する場合について考える．このとき，得られた報酬 r と強化関数に基づき強化値 $\boldsymbol{f}_n$ を決定する．このとき用いられる強化関数を式 4 に，位相回転の伝播の様子を図 2 のブロック図に示す．また，図 3 に示すように，一般的な PS の強化値 f_n は実数軸上で減少するのに対し，複素評価値を用いた PS の強化値 $\boldsymbol{f}_n$ は複素平面上で螺旋状に回転しながら減少する．この強化値 $\boldsymbol{f}_n$ を用い，エピソード中に経験した状態行動対 (s_n, a_n) の評価値 $\boldsymbol{v}(s_n, a_n)$ を式 5 に従い更新する．

$$\boldsymbol{f}_n = \begin{cases} r & (n = W-1) \\ \gamma e^{j\omega_{n+1}} \boldsymbol{f}_{n+1} & (n = W-2, W-3, \cdots, 0) \end{cases} \tag{4}$$

$$\boldsymbol{v}(s_n, a_n) \leftarrow \boldsymbol{v}(s_n, a_n) + \boldsymbol{f}_n \quad (n = W-1, W-2, \cdots, 0) \tag{5}$$

4.2.2 内部状態更新方法

複素強化学習では，エージェントに文脈を保持する複素数として内部状態 $\boldsymbol{i}(n)$ を持たせる．$n+1$ ステップ目において行動選択に用いる内部状態 $\boldsymbol{i}(n+1)$ は，n ステップ目の内部状態 $\boldsymbol{i}(n)$ に $-\omega_{n+1}$ の位相回転を加えた値とする．ただし，初期状態 s_0 における内部状態 $\boldsymbol{i}(0)$ の位相は初期位相 θ_0 とする．ここで，初期位相 θ_0 は，初期状態 s_0 における状態行動対集合の中で評価値の絶対値が最も大きい状態行動対の評価値の位相とする．よって，n ステップ目における内部状態 $\boldsymbol{i}(n)$ は式 6 で表すことができる．ただし，$j^2 = -1$ である．

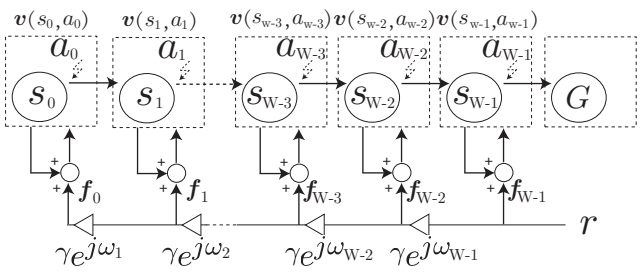


図 2　複素評価値更新のブロック図

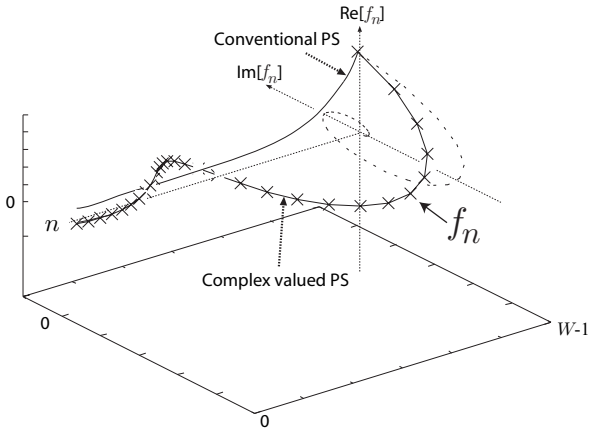


図 3　強化関数の比較

$$\boldsymbol{i}(n) = \begin{cases} \exp(j\theta_0) & (n = 0) \\ \exp\left(j\left(\theta_0 + \sum_{k=1}^{n} (-\omega_k) \right) \right) & (n = 1, 2, \cdots) \end{cases} \tag{6}$$

4.2.3 行動選択方法

本稿では，行動選択手法として複素強化学習のためのルーレット選択を用いる．この選択手法は，内部状態と評価値の内積値をもとに行動を決定する．状態 s_n における行動 a の選択確率 $P_C(s_n, a)$ を式 7 で定義する．ここで，$V(s_n, a)$ は，状態行動対の評価値 $\boldsymbol{v}(s_n, a)$ と，エージェントが持つ内部状態の複素共役 $\bar{\boldsymbol{i}}(n)$ を用いて式 8 で定める．

$$P_C(s_n, a) = \frac{V(s_n, a)}{\sum_{a' \in A(s_n)} \left(V(s_n, a') \right)} \tag{7}$$

$$V(s_n, a) = \mathrm{Re}\left[\boldsymbol{v}(s_n, a) \bar{\boldsymbol{i}}(n) \right] \tag{8}$$

ただし内部状態との位相差が $\pi/2$[rad] より大きい，つまり $V(s_n, a) < 0$ となる評価値の行動は，選択候補から外すものとする．また，行動集合 $A(s_n)$ のすべての評価値で $V(s_n, a) < 0$ となる場合は，ランダムに行動を選択する．

5 シミュレーション実験

本手法の有効性を確認するためにグリッドワールドにおける棒運びタスクを用いて実験を行った．

エージェントの目標は，棒の中心を持ったままスタート領域から移動し，ゴール領域まで棒を運ぶことである．エージェントが選択可能な行動は，東西南北へ 1 マス移動するか，その場にとどまったまま $\pm\frac{\pi}{6}$ 回転するかの 6 通りである．

図 4 に示すように，エージェントは移動しながら適切に回転し，棒が壁にぶつからないよう隙間を通過する必要がある．このとき，エージェントは自分の座標のみ観測可能で，現在の棒の回転角については観測できないも

表 1　パラメータ設定

	r	γ	ω_n[rad]
Conventional PS	100.0	0.5	-
Complex valued PS	100.0	0.5	$\pi/18$

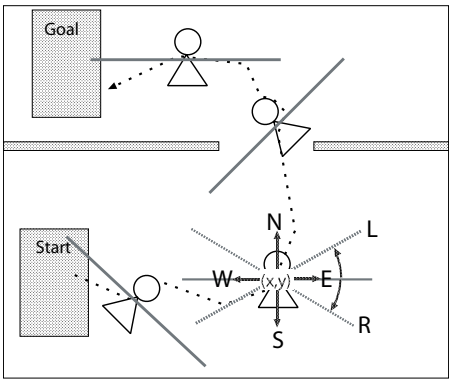


図 4　棒運びタスクとエージェントの行動

のとする．すなわち，棒の角度については不完全知覚が生じており，その多重度は 12 通りになる．

スタート領域からゴール領域にたどり着くまでを 1 試行とし，1 試行あたり最大 1000 ステップとした．それぞれの実験において試行数 1000 回を 1 学習とし，1000 学習行った．実験で用いたパラメータを表 1 に示す．ここで，ω_n は，ステップによらず一定の値をとるものとした．

6　結果および考察

シミュレーション結果を図 5 に示す．従来の PS では，試行数を重ねても学習が進んでいない．これに対して提案手法である複素評価値を用いた PS では，従来の PS を大きく上回る学習結果が得られた．

また，提案手法を用いたエージェントが獲得した典型的な軌跡を図 6 に示す．図 6(a)，図 6(b) はそれぞれ 19 ステップ，20 ステップでタスクを達成した場合の軌跡である．どちらの場合でも棒の回転を適切に行い壁の隙間を通過する行動を獲得できた．提案手法を用いたエージェントは内部状態に応じた評価値を選択することで，不完全知覚問題がある環境でも適切な動作を獲得していることが確認できた．

従来の PS では，棒の角度について生じている不完全知覚のためにエージェントは適切な回転量を判断できずタスクの達成が困難となる．

一方，提案手法では，棒の回転と移動の時系列情報を評価値の位相に含ませ，タスクを時間方向に分割することで適切な行動を獲得できた．

7　おわりに

複素強化学習の新たなアルゴリズムとして複素評価値を用いた PS を提案した．提案手法は，不完全知覚問題がある環境において従来の PS を大きく上回る学習能力

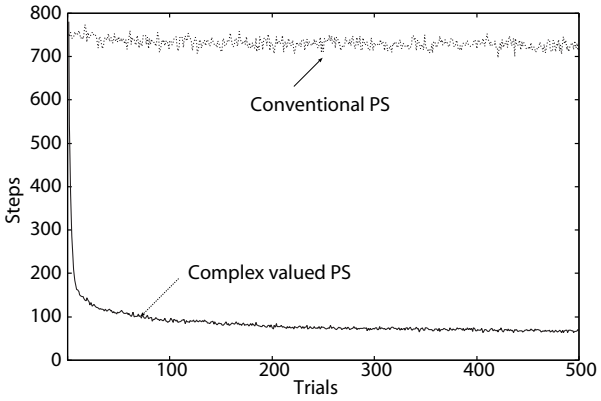
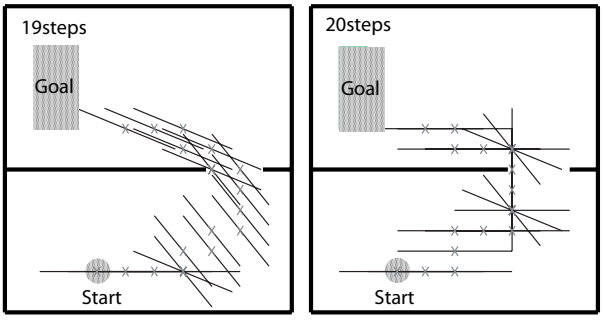


図 5　シミュレーション結果



(a) 19 steps で運んだ軌跡　　(b) 20 steps で運んだ軌跡

図 6　ゴールまでの軌跡

を有することが確認できた．

今後は，ω_n によって変化する学習の効率を調べるとともに，本アルゴリズムによって解ける問題のクラスを多重度などをもとに分析する予定である．さらに，$\dot{Q}$-learning との学習能力の比較も行う予定である．

参考文献

[1] 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，Jan.1999.

[2] Georgios Theocharous, Leslie Pack Kaelbling, “Approximate Planning in POMDPs with Macro-Actions, ” Advances in Neural Information Processing Systems 16, Vancouver, 2004

[3] C.J.C.H.Watkins and P.Dayan, “Technical note:Q-learning,” Machine Learning, Vol.8, pp.279-292, 1992.

[4] 澁谷長史，濱上知樹，“複素価値関数を用いた強化関数に関する基礎的検討”，第 4 回情報科学技術フォーラム 一般講演論文集 第 2 分冊，pp.197-198, Sep.2005.

[5] Grefenstette, J. J. , “Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms,” Machine Learning, 3, pp.225-245, 1988.

[6] 宮崎和光，木村元，小林重信，“Profit Sharing に基づく強化学習の理論と応用”，人工知能学会誌，Vol.14，No.5，pp.1-8，Sep.1999.