F-004

複素価値関数を用いた強化学習に関する基礎的検討

A fundamental study of Q-learning with complex value function

澁谷 長史 †　　濱上 知樹 †
Takeshi SHIBUYA　Tomoki HAMAGAMI

1 はじめに

強化学習において，エージェントは環境から与えられる報酬に基づいて行動政策の獲得を行う．強化学習は，環境との相互作用をもとに自律的に学習を行うため，自律移動ロボットなどへの応用に向いている．

一方で，不完全知覚問題や次元の呪いなどをはじめとする多くの本質的問題も存在し [1]，重要な課題となっている．これら本質的問題を解決するような研究もさまざまに行われている [2][3]．しかし，多くの方法は環境に対してサブゴールなど一定の仮定を導入することで環境への依存性が高まってしまう．

本研究では，従来法とは異なり，仮定を用いずにこれらの問題の解決を図り，汎用性のある手法を検討している．特に本稿では価値関数に複素数を用いる手法について基礎的な検討を行った．複素価値関数による強化学習を複素強化学習と呼ぶことにする．具体例として本稿では Q-learning における Q 値を複素数として扱う．これを $\dot{Q}$-learning と呼ぶことにする．計算機実験によって複雑なアルゴリズムを用いることなく不完全知覚問題のある環境においてに提案手法が有効であることを示す．

2 提案手法

2.1 方針

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の情報を含ませる．

本稿では，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 $\dot{Q}$ 値を予測する．図 1 は予測された $\dot{Q}$ 値（$\dot{R}$）と選択可能な行動に対応する $\dot{Q}$ 値（$\dot{Q}_1$，$\dot{Q}_2$）の関係を示したものである．$\dot{R}$ が異なると，原点と各 $\dot{Q}$ 値からの垂線の足とを結ぶ線分の長さも異なる．このことを利用すると，選択確率を文脈に基づいて動的に変化させることができる．Q 値を実数として扱う場合には単純に大きさの比較を行うことしかできないが，複素数である $\dot{Q}$ 値を用いることで位相差を含んだ比較が可能となる．これによって複雑なアルゴリズムを使わなくても文脈に応じた行動選択をすることができるようになる．

2.2 更新アルゴリズムの定式化

状態 s_i から行動 a_i をとって状態 s_{i+1} へと遷移し報酬 r を受け取ったときの，$\dot{Q}$ 値の更新則を以下のように定義する．

$$\dot{Q}(s_{i-k}, a_{i-k}) \leftarrow (1-\alpha)\dot{Q}(s_{i-k}, a_{i-k}) + \alpha(r + \gamma \dot{Q}_{\max}^{s_i \to s_{i+1}})\dot{u}(k) \quad (1)$$

† 横浜国立大学大学院工学府

k ステップ前の状態，行動をそれぞれ s_{i-k}，a_{i-k} とする．

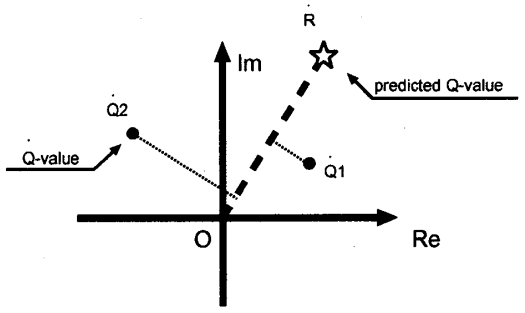

図 1　ある状態における $\dot{Q}$ 値を複素平面上に示した図

ここで，k ステップ前の状態，行動をそれぞれ s_{i-k}，a_{i-k} とする．$\dot{u}(k)$ は形式上の適格度トレースであり，式 2 のように定義する．

$$\dot{u}(k) = \dot{\beta}^k \quad (2)$$

式 1 の適用は，予め定めた整数 N_e を用いて，$0 \le k \le N_e$ の範囲で行う．ただし，$\dot{\beta}$ は絶対値が 1 以下の複素数である．$\dot{Q}_{\max}^{s_i \to s_{i+1}}$ は，式 3 のように定義する．

$$\dot{Q}_{\max}^{s_i \to s_{i+1}} = \max_{a'} \dot{Q}(s_{i+1}, a')\overline{\dot{R}_i} \quad (3)$$

ここで，予想される $\dot{Q}$ 値（$\dot{R}_i$）は次のように定義する．

$$\dot{R}_i = \dot{Q}(s_i, a_i)/\dot{\beta} \quad (4)$$

2.3 行動選択アルゴリズムの定式化

本稿では，Max-Boltzmann 選択を用いる．すなわち，状態 s_i に居るエージェントは，確率 $1 - P_{\max}$ で Boltzmann 選択を行い，確率 $P_{\max}$ で greedy 方策を行うことにする．

状態 s_i，行動 a_i に対応する $\dot{Q}$ 値を $\dot{Q}(s_i, a_i)$ とする．また，状態 s_i における行動 a の Boltzmann の選択確率を $Prob(s_i, a)$ とする．状態 s_i における行動集合を $A(s_i)$，直前の状態と行動に対応する $\dot{Q}$ 値を $\dot{Q}(s_{i-1}, a_{i-1})$，Boltzmann 選択の温度パラメータを T とするとき，$Prob(s_i, a)$ を次式のように定める．

$$Prob(s_i, a) = \frac{\exp\left(Re[\dot{Q}(s_i, a)\overline{\dot{R}_{i-1}}]/T\right)}{\sum_{a' \in A(s_i)} \exp\left(Re[\dot{Q}(s_i, a')\overline{\dot{R}_{i-1}})]/T\right)} \quad (5)$$

ただし，$Re[\cdot]$ は複素数の実部を表す．

greedy 方策は $\arg\max_a Prob(s_i, a)$ を選択することにする．

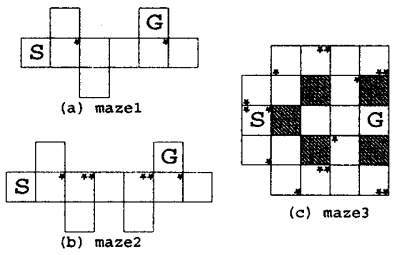

図 2　実験環境．S はスタートを，G はゴールをそれぞれ表す．*や**は不完全知覚の影響のある状態を表す．

表 1　パラメータ設定

	trials	α	$1-P_{\max}$
phase 1	1 to 20	0.05	0.1
phase 2	21 to 80	(100-try)/400	(100-try)/1600
phase 3	81 to 100	0	0

3　計算機実験

図 2 のような簡単なグリッドワールドにおける迷路問題を対象として計算機実験を行い，提案手法の有効性を確認する．

3.1　状態空間と行動集合

エージェントが観測可能な情報は，東西南北周囲 4 マスの壁の有無のみとし，この情報を直接状態として割り当てることにする．すなわち観測可能な状態数は $2^4 = 16$ となる．

これらの環境において不完全知覚の影響のある状態が存在する．例えば，*においてはそれぞれにおいて選択すべき行動が異なり，**においては同じ行動をとらなければならない．

エージェントが任意の状態において選択することができる行動は，壁のない方向に進むのみとする．すなわち，行動集合 $A = \{$東, 西, 南, 北$\}$ の空集合でない部分集合とする．

3.2　パラメータ設定と実験結果

エージェントは，ゴールにたどり着くと環境から報酬 $r = 100$ を受け取り，初期状態であるスタートに再配置されるものとした．エージェントの行動 1 ステップごとに負の報酬を与えることや，ゴールにたどり着くのにかかったステップ数に応じて報酬を変えることなど，早くゴールにたどり着く学習を助長するような報酬の与え方はしない．

試行数 100 を 3 つのフェーズに分け，それぞれについてパラメータの設定を行った．ステップごとに変化するパラメータついては表 1 のように設定し，それ以外のパラメータについては各フェーズにおいて共通とし $\beta = 0.9\exp(j\pi/6)$, $\gamma = 0.999$, $T = 3000$, $N_e = 1$ とした．ただし，$j^2 = -1$ である．

計算機実験の結果を図 3 に示す．この結果は 100 試行を 1 学習として 100 学習行い，収束したものに関しての平均である．

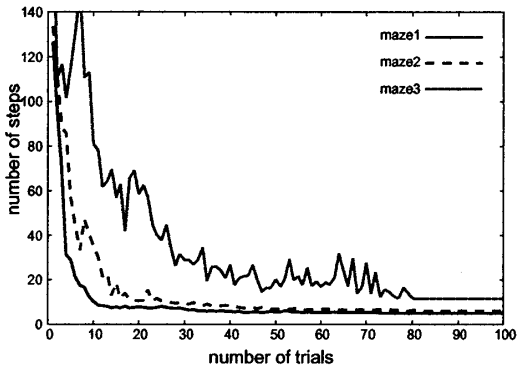

図 3　実験結果

4　考察

maze1，maze2 においては 100% が収束し，maze3 においては 95% が収束した．本稿では maze1，maze2 において 100% が最短経路を実現するような方策を獲得した．

maze1 では獲得した一連の行動について，ある $\dot{Q}$ 値は次の行動に対応する $\dot{Q}$ 値と β の偏角だけずれるような位相をもつよう学習がされた．

maze2 においては maze1 のように一連の行動に対する $\dot{Q}$ 値は単純な関係にはならないかった．しかし，$\dot{Q}$ 値の位相を自律的に調整することで，不完全知覚問題を解決している様子が観察された．

maze3 では，最短経路を実現するような方策を学習しなかったが，環境中を一部往復する行動をとることで自律的に環境を多重化して不完全知覚問題に対処している様子が観察された．

5　おわりに

強化学習の価値関数に複素数を用いる手法を提案した．本稿では，その具体例として Q-learning の Q 値を複素数とする $\dot{Q}$-learning を検討しその有効性を示した．今後は，同じ行動に対して複数の $\dot{Q}$ 値を割り当てることを行うなど，さまざまな観点からさらなる性能の向上を目指す．

参考文献

[1] Kaelbling.and Littman and Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, pp. 237–285, 1996.

[2] T.Hamagami, S.Koakutsu, and H.Hirata. Reinforcement learning to compensate for perceptual aliasing using dynamic additional parameter: Motivational value. In *Proc. of IEEE Int' conf. on SYSTEMS (SMC2002)*.

[3] 山城, 上野, 武田. 遅れ報酬に基づく遺伝的アルゴリズムによる部分観測マルコフ決定問題の解決手法. 信学論誌 (D-I), Vol. J84-D-I, No. 12, pp. 1635–1647, 2001.