F-001

複素強化学習において行動価値を多重化する効果について

An effect of multiplied action values in the complex-valued reinforcement learning

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

1 はじめに

自ら行動し経験を重ねることで振る舞いを獲得する枠組みとして強化学習 [1] が知られている. この枠組みでは，エージェントとよばれる学習主体はある環境のなかで観測・行動・状態遷移を繰り返し（試行錯誤），望ましい状態になった場合には特別な信号（報酬）を受け取る. エージェントはなるべく多くの報酬が得られるような振る舞いの獲得を目指す.

アプリケーションによってはセンサの種類・数・精度は制約をうけ，エージェントが観測によって自身の状態を一意に識別できない場合がある. すなわち，複数の異なる状態を同じ状態としてみなすという問題が発生する. この問題は不完全知覚問題 [2] とよばれ，強化学習を実環境に適用する際のボトルネックとなっている.

この不完全知覚問題に対して，筆者らは複素強化学習とよばれる枠組みを提案している [3][4]. 提案する枠組みにおいて，複素化された価値関数は価値の大きさだけではなく位相情報を表現することができる. 先の報告では複素強化学習が，不完全知覚問題を含む小規模な迷路タスクを達成できることを明らかにした [3].

しかし，従来の複素強化学習では，ある観測に対して，ある行動をとることが複数回求められるようなタスクでは，十分な学習が行えない場合があった. そこで本稿では，ひとつの行動に複数の行動価値を割り当てる，行動価値の多重化を提案する.

2 複素強化学習

文脈依存な行動価値を実現するためにふたつの複素数を導入する. ひとつめは，複素数で表現された価値（複素価値）である. 複素価値の絶対値で従来の価値の大きさを，複素価値の位相で時系列情報を表すことにする. もうひとつは，内部参照値とよばれるエージェントの文脈を保持する変数である.

複素行動価値と内部参照値との相互作用について，以下の仮定を設ける.

> **仮定 1** 複素行動価値の絶対値が大きいほど，その行動は選ばれやすい
> **仮定 2** 複素行動価値の位相と内部参照値の位相が近いほど，その行動は選ばれやすい

ひとつめの仮定は，ある状態において将来の期待収益が大きい行動ほど選ばれやすいという従来の強化学習の考え方を踏襲する仮定である. ふたつめの仮定は，内部参照値を文脈の相として活用するための仮定である.

† 横浜国立大学大学院工学府

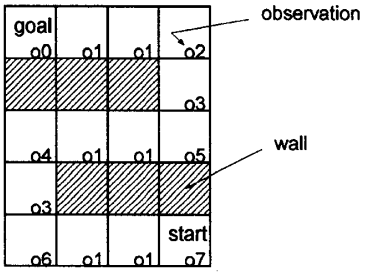

図 1　実験に用いた迷路

筆者らは，複素価値関数を Q-learing に適用した Q̇-learning を提案している.[3]

3 提案手法

複素強化学習は，文脈を獲得することで不完全知覚問題の解決を図っており，既に簡単なタスクにおいて有効性が確認されている. しかし，ある観測に対して，ある行動をとることが複数回求められるようなタスクでは，十分な学習が行えない場合がある. これは，ある行動が選ばれるためには行動価値と内部参照値の位相が近くなければならないが，複数の内部参照値すべてに対して行動価値の位相を近づける学習ができないためである.

このことを図 1 の迷路タスクを用いて具体的に述べる. エージェントの目標は start から goal までたどり着くことである. エージェントの観測可能な情報は東西南北周囲 4 マスの壁の有無であるとする. すなわち，観測の種類は o_0 から o_7 までの 8 種類であるとする. o_3 を観測するエージェントはいつも北を選択しなければならない. しかし，o_6 を観測して北を選択したエージェントの内部参照値は，そのあと o_5 を観測して北を選択したあとの内部参照値と，間の状態遷移の分だけ異なる. 内部参照値の位相の違いが大きいとき，位相がずれてしまうために o_5 では北を選択できなくなってしまう.

そこで，本稿ではひとつの行動に複数の行動価値を割り当てる，行動価値の多重化を提案する. 行動価値の多重化を行うと，異なる内部参照値に対しても常にある行動を選択できるようになる.

図 2 は従来の複素強化学習における行動価値を表している.(a) はある観測における複素平面上の行動価値，(b) は内部参照値の位相に対する実効的な行動価値の大きさの関係を表している. 横軸は内部参照値の位相 $\angle \dot{I}$，縦軸は内部参照値の相互作用 $Re[\dot{Q}(s,a)\bar{\dot{I}}]$ である. 縦軸が大きいほど，内部参照値を基準として文脈に沿った行動であること，すなわち実効的な行動価値が高いことを示している. (a) で示すような複素価値が獲得されているとき，ふたつの行動 a_0 と a_1 のうち，(b) の矢印の領域で

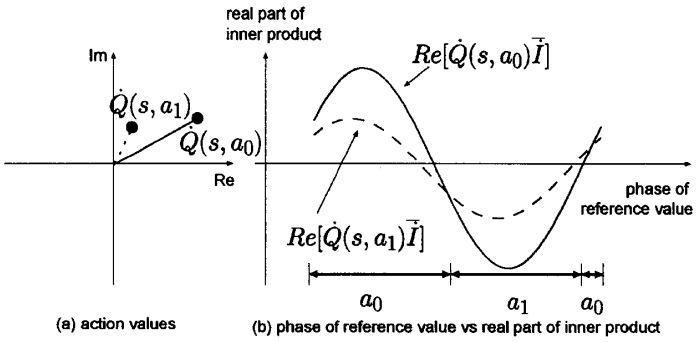

図 2　従来の複素強化学習における行動価値 ($|\dot{I}| = 1$ に規格化した場合)

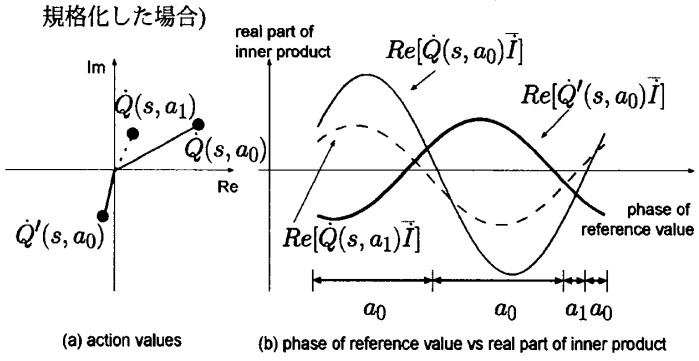

図 3　多重化された行動価値 ($|\dot{I}| = 1$ に規格化した場合)

示された行動が選択される. これに対して 図 3 では，具体例として a_0 の行動価値を 2 倍に多重化している. 行動価値を多重化することで，a_o と a_1 の選択される領域が変化している.$\dot{Q}(s, a_0)$ と $\dot{Q}'(s, a_0)$ を独立に学習することにより，単一の行動価値では選択されない内部参照値においても，a_0 が再び選択されることを可能にする.

行動価値の多重化は，複素価値の性質を利用した手法である. 従来の実数価値を用いた強化学習においては，文脈に依存せず常に価値の高い行動がよい行動とされるため，行動価値の多重化を行うことはできない.

4　シミュレーション実験

図 1 のグリッドワールドにおいて迷路タスクによるシミュレーション実験を行った. この実験の目的は，提案手法である行動の多重化を適用した Q̇-learning と 従来の Q̇-learning の比較である.

エージェントは，東西南北のうち壁のない方向へと進むことができる. ただし，提案手法ではそれぞれの行動価値を 2 倍に多重化した. エージェントが goal にたどり着いたときに報酬 $r = 100$ を与えた.

学習パラメータは，$\dot{\beta} = \exp(j\pi/6)$, $\gamma = 0.9$, $T = 100/(episode + 1)$, $N_e = 5$ とした. 一回の状態遷移を 1 ステップ，一回 goal にたどり着くまでを 1 エピソード, 100 エピソードを 1 学習として, 100 学習おこなった.

エピソード数に対する平均のステップ数を図 4 に示す. 行動の多重化を行っていない Q̇-learning は振る舞いを獲得できていないのに対し，提案手法は最短経路をとる振る舞いを獲得できたことがわかった.

5　考察

内部参照値の位相と観測 o_3 における行動価値の関係を図 5 に示す. このタスクでは観測 o_3 では常に北を選択

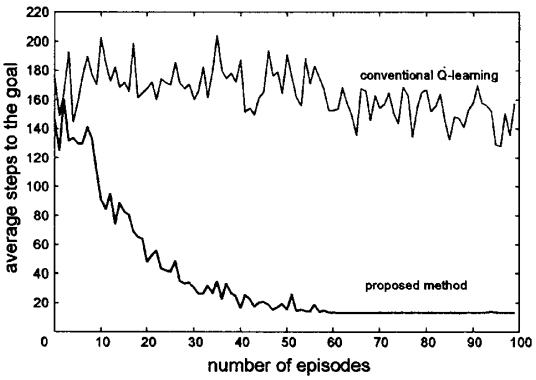

図 4　迷路タスクにおける実験結果

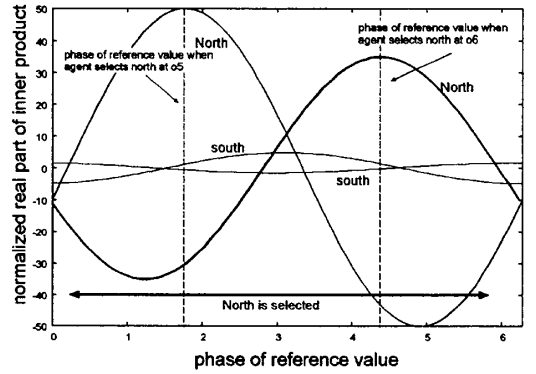

図 5　内部参照値の位相と観測 o_3 における行動価値の関係

する行動が望ましいため，o_5，o_6 のどちらを観測する状態から遷移したときにも北が選ばれるような行動価値を学習をしている.

6　おわりに

複素強化学習における行動価値の多重化手法を提案し，ある観測に対して，ある行動をとることが複数回求められるような簡単なタスクに適用した. その結果，行動価値の多重化は “もっともよい行動” となる領域を柔軟に表現できる効果があることが確認された. 今後は，より大規模なタスクに本手法を適用して有効性を確認していくとともに，本稿では固定であった多重化の倍数を動的に変化させる手法を検討していく予定である.

参考文献

[1] Richard S. Sutton and Andrew G. Barto, “REINFORCEMENT LEARNING: An Introduction,” MIT Press, 1998.

[2] Steven D. Whitehead and Dana H. Ballard, “Learning to Perceive and Act by Trial And Error,” Machine Learning, Volume 7, Number 1, pp. 45-83, 1991

[3] 澁谷長史，濱上知樹,“複素価値関数を用いた強化学習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, 2005.

[4] T.Hamagami, T.Shibuya, S.Shimada, “Complex-Valued Reinforcement Learning,” Proc. of IEEE international Conference Systems, Man and Cybernetics, pp.3235-3530, 2006.