F-041

複素強化学習における負の報酬の及ぼす影響について

An effect of negative reward in complex-valued reinforcement learning

澁谷 長史 †　　濱上 知樹 †
Takeshi SHIBUYA　Tomoki HAMAGAMI

1　はじめに

ロボットの振る舞いをあらかじめ設計しておく代わりに，ロボットが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][2] が知られている．この枠組みのもとでは，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，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を得られるような振る舞いの獲得を目指す．エージェントは行動選択の基準として，行動価値と呼ばれる指標を用いる．強化学習においては行動価値を適正に調整することが必要となる．

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．すなわち，複数の異なる状態を同じ状態としてみなしてしまうという問題が発生する．この問題は不完全知覚問題 [3] とよばれ，強化学習の分野において重要な課題となっている．

この不完全知覚問題に対して，複素強化学習とよばれる枠組みを提案している [4]．提案する枠組みにおいて，複素化された行動価値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．

しかし，これまでの複素強化学習の手法では負の報酬（罰）を扱うことを想定しておらず，負の報酬を用いた場合の影響は明らかではなかった．負の報酬とは，設計者が抑止したい行動をエージェントが選択しないよう学習するために用いられる手法であり，実応用の観点から重要な概念である．

本稿では複素強化学習において負の報酬が与える影響を明らかにし，負の報酬を用いるための手法について検討する．

2　複素強化学習

文脈依存な行動価値を実現するために複素数で表現された価値 (複素価値) を導入する．複素価値の絶対値で従来の強化学習における価値の大きさを，複素価値の位相で時系列情報をそれぞれ表すことにする．

複素強化学習では，もうひとつの複素数である内部参照値を導入する．内部参照値はエージェントの文脈を保持する変数である．不完全知覚によって本来異なる状態を同じ状態としてみなしてしまっても，内部参照値が異なれば区別することができ，とる行動を変えることができる．

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．

筆者らは，複素価値関数を Q-learning に適用した Q̇-learning を提案している [4]．

3　複素強化学習における負の報酬の及ぼす影響に関する考察

設計者にとって望ましくない行動を抑止するために，強化学習において負の報酬を用いることがある．正の報酬が褒美であるのに対して，負の報酬は罰に相当する．

図 1(a) に示すように，実数の強化学習において，負の報酬は行動価値を減じる働きがある．行動価値の大きい行動ほど選択されるので，負の報酬は行動を抑止する手段として機能する．

一方，複素強化学習においては負の報酬を単純に用いただけでは行動を抑止する手段として機能しない．図 1(b) に示すように，負の報酬を与えることは単に左側に平行移動させるという意味になる．次に述べるように，絶対値と位相の両面から，負の報酬は行動選択を抑止する手段としては用いることができない．

複素強化学習では，行動の選択確率は複素行動価値の絶対値に依存する．前述の操作によって，左半平面にある複素行動価値の絶対値が大きくなるため，かえってこれらの行動の選択確率があがってしまう．

さらに，複素強化学習では，時系列の取扱いのために，

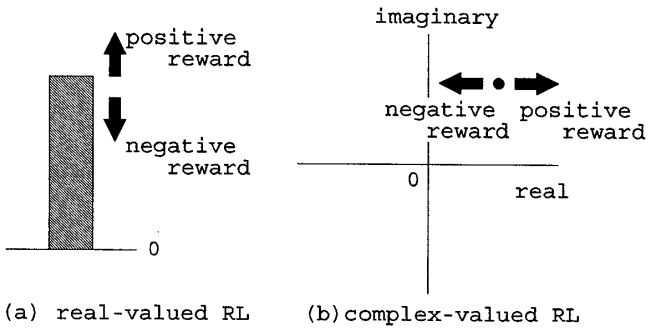

図 1　実数強化学習 (real-valued RL) と複素強化学習 (complex-valued RL) における負の報酬がもたらす価値の更新のされ方の違い

† 横浜国立大学大学院工学府

複素行動価値の位相を重視する．前述の方法で負の報酬を与えるということは複素行動価値の位相の関係性を π に近づける操作となり，その極限では負の報酬が与えられる複素行動価値の位相の関係性を破壊してしまう．

4 負の報酬を取り扱うための一手法

前節の考察を受け，本節では，エージェントが負の報酬を得た場合に行動価値の絶対値を小さくする手法を提案する．

- 行動価値の初期値を，複素平面上の適当な範囲の乱数とする．
- 報酬が正である場合には，これまで通常の Q-learning を行う．
- 報酬が負である場合には，更新対象の複素行動価値に絶対値が 1 未満の複素数を乗ずる．
- 観測や行動の系列がループしたら，内部参照値に絶対値が 1 の複素数を乗ずる．

本稿では，簡単のために，エージェントが負の報酬を得た場合には，その値を問題にせずに一定の処理を行う．

5 実験

提案手法の有効性を確認するために，図 2 に示す環境においてシミュレーション実験を行った．エージェントは，太い枠のマスの観測に対して，一度以上異なる行動を選択する必要がある．この環境では，黒いマスにおいてエージェントに負の報酬が与えられる．すなわち，白いマスのみを進んでゴールすることが望まれている．

学習パラメータ [4] は，学習率 $\alpha = 0.25$，位相回転 $\dot{\beta} = \exp(j\pi/6)$，割引率 $\gamma = 0.9$，ボルツマン温度 $T = 20$，トレース数 $N_e = 2$ 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，500 エピソードを 1 学習として，50 学習おこなった．

乱数の初期値は，$\{z|\mathrm{Re}[z] \in [-50, 50]$ and $\mathrm{Im}[z] \in [-50, 50]\}$ の領域で一様分布により決定した．負の報酬を得た行動の行動価値には，$0.99\exp(-\pi/8)$ を乗じた．観測と行動の系列がループした場合には，内部参照値に -1 を乗じた．これらのパラメータは予備実験によって決定した．

学習の平均ステップ数を図 3 に示す．学習開始直後から平均ステップ数が減少し，エージェントが行動を学習している様子が観察された．なお，５００エピソード終了時点では平均ステップ数は約１２まで減少していた．

負の報酬が与えられない最短経路を獲得した割合は，約 60% であった．すなわち，エージェントは，約 60% の確率でほぼ決定論的にこの最短経路を選択するような振る舞いを獲得する．ランダムウォークがこの経路を実現する確率 $1/2^8 \simeq 4\%$ と比べて，提案手法が望ましい経路を獲得する確率が十分に高いことがわかる．

パラメータの調整によって，さらなる向上が期待できる．

6 おわりに

複素強化学習において，負の報酬が行動価値に及ぼす影響について明らかにし，実数の強化学習における負の報酬と同じ役割を果たす手法を提案した．実験の結果，この手法の有効性が確認された．

今回の手法は，目的を達成する一手法にすぎず，ほかの手法を検討する余地が十分にある．今後はこの検討と合わせて，本手法パラメータの決定方法なども行う．

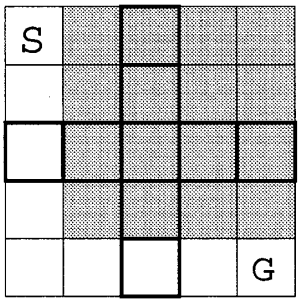

図 2 実験に用いた迷路．S はスタート，G はゴールを表す．太枠で囲まれたマスはすべて同一の観測を与える．網掛けで示されたマスでは負の報酬が与えられる．ゴールでは，報酬 100 が与えられる．

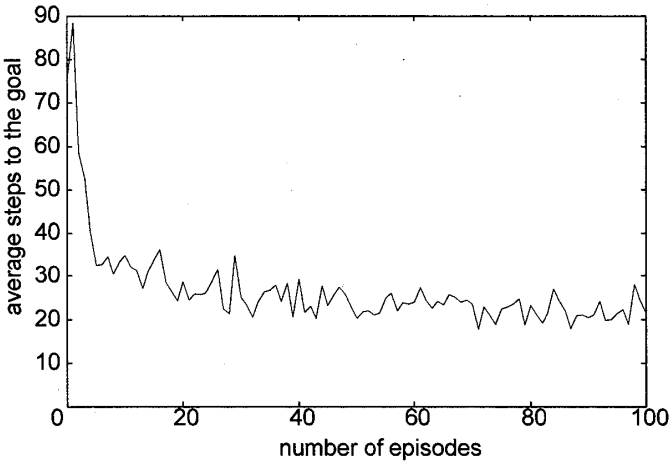

図 3 学習曲線

参考文献

[1] Leslie P. Kaelbling and Michael Littman and Andrew Moore, "Reinforcement Learning: A Survey," Journal of Artificial Intelligence Research, Volume 4, pp. 237–285, 1996

[2] Richard S. Sutton and Andrew G. Barto, "REINFORCEMENT LEARNING: An Introduction," MIT Press, 1998.

[3] 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

[4] 澁谷長史，濱上知樹，"複素数で表現された行動価値を用いる Q-learning"，電子情報通信学会論文誌 D Vol.J91-D,No.5,pp.1286-1295,2008