3D9-4

複素強化学習を用いたAcrobotの振り上げ制御

Swinging up Acrobot using Complex-Valued Reinforcement Learning

澁谷 長史 Takeshi Shibuya　島田 慎吾 Shingo Shimada　濱上 知樹 Tomoki Hamagami

横浜国立大学大学院工学府
Graduate School of Engineering, Yokohama National Universty

Swinging up acrobot with incomplete sensors is conducted to declare that complex-valued reinforcement learning enables an agent to achieve difficult task. The agent controlling the acrobot observes only angular of each joint. The experimental result shows complex-valued reinforcement learning is effective for the acrobot task.

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．強化学習は，教師あり学習・教師なし学習とならぶ機械学習のひとつの枠組みである．強化学習の特徴は，試行錯誤と報酬である．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，強化学習を適用する際のボトルネックとなっている．

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.[4] この枠組みでは価値関数の複素数値化によって時系列の取り扱いが可能である．複素強化学習を用いるメリットは２つある．ひとつめは Q-learning や Profit Sharing のような従来の強化学習アルゴリズムに容易に適用できることでである．ふたつめは直接過去の履歴を参照しないことで，メモリベースト法とくらべて計算資源の乏しいロボットの制御に適用できることである．

先の研究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．[4]

本稿では，より複雑なタスクとして角速度センサを利用できない Acrobot の振り上げ制御タスクを導入する．通常の Acrobot においては，エージェントは２つの関節それぞれについての角度と各速度を観測する．これに対して本稿で導入する Acrobot では，それぞれの関節の角度しか観測できないという制限を設ける．

2. 複素強化学習

筆者らが提案している強化学習の枠組みである複素強化学習について述べる．複素強化学習では，従来の価値関数を複素数値化した複素価値関数と，内部参照値という変数を導入することで時系列の取り扱いができる．これにより，Q-learning や Profit Sharing のような単純なアルゴリズムを用いて行動の文脈を効率的に獲得することが可能である．

連絡先: 澁谷長史 tks-shibuya@mail.dislab.ynu.ac.jp

2.1 複素強化学習の枠組み

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．複素価値は，絶対値で従来の価値の大きさを，位相で時系列情報を それぞれ表すことができるため，従来の価値よりも豊かな表現力をもつ．

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．

エージェントは，複素行動価値と内部参照値との相互作用によってとるべき行動を決定し，その決定に応じて内部参照値を変化させることを繰り返す．

この操作はメモリベースト法のように過去の状態遷移履歴の代わりに 内部参照値によって 同じ観測に対する異なる行動選択を可能にする．

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．

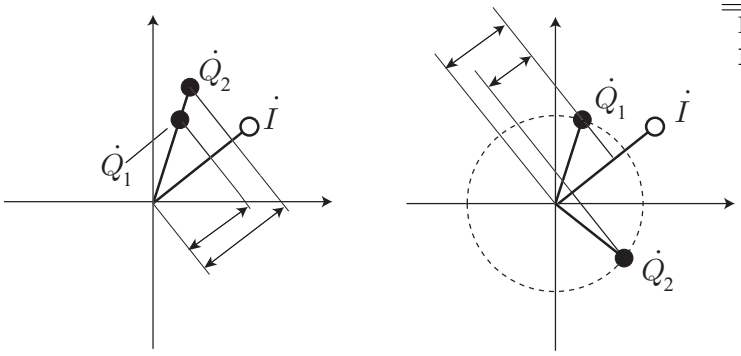
仮定 1 複素行動価値の絶対値が大きいほど，その行動は選ばれやすい

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．

図 1(a)(b) は，それぞれ 仮定 1，仮定 2 を視覚的に表している．黒丸の点 $\dot{Q}_1$ と $\dot{Q}_2$ は複素平面上の一点であり，複素行動価値を表す．白丸の点 $\dot{I}$ は，同じく複素平面上の一点であり，内部参照値を表す．以降，特に断りのない限り，ドットのつく変数・関数は複素数を表すことにする．特別な例として，内部参照値の絶対値が 1 である場合を考えると，内積の実部は内部参照値から原点へ引いた直線にそれぞれの複素価値からの垂線の足と原点との距離に等しい．

たとえば，図 1(a) においては，$\dot{Q}_1$ と $\dot{Q}_2$ を比べると，$\dot{Q}_2$ のほうが大きな収益を期待でき，ふたつの行動価値の位相は等しい．この場合，仮定 1 によって 絶対値の大きい $\dot{Q}_2$ が選択される．また，図 1(a) においては，$\dot{Q}_1$ と $\dot{Q}_2$ を比べると，ふたつの行動価値の期待収益の大きさは等しいが，$\dot{Q}_1$ のほうが $\dot{Q}_2$ よりも 内部参照値 $\dot{I}$ に位相が近い．この場合，仮定 2 によって 絶対値の大きい $\dot{Q}_1$ が選択される．



(a) assumption 1 (b) assumption 2

図 1: 複素強化学習の原理図

このように，複素強化学習では複素行動価値と内部参照値との複素平面上での位置関係によって “もっともよい行動” を変化させることができる．

2.2 $\dot{Q}$-learning

Q-learning に複素価値関数を導入したアルゴリズムが $\dot{Q}$-learning である．ある時刻 t において状態 s_t にいたエージェントが行動 a_t をとったときの複素価値 $Q(s_t, a_t)$ を次式で更新する．

$$\begin{aligned}\dot{Q}(s_t,a_t) &\leftarrow (1-\alpha)\dot{Q}(s_t,a_t) \\ &\quad +\alpha\left(r_{t+1}+\gamma\dot{Q}^{(t)}_{\max}\right)\dot{\beta} \qquad (1)\\ \dot{Q}^{(t)}_{\max} &= \dot{Q}(s_{t+1},a) \qquad (2)\\ a &= \max_b\left(Re[\dot{Q}(s_{t+1},b)\bar{\dot{I}_t}]\right) \qquad (3)\end{aligned}$$

I_t は，時刻 t における内部参照値である．

I_t の定め方はさまざま考えられるが，本稿では，

$$\dot{I}_t = \begin{cases} \dot{Q}(s_t,a_t)/\dot{\beta} & t \geq 0 \\ \dot{Q}(s_{t+1},a) \quad a=\arg\max_b|\dot{Q}(s_{t+1},b)| & otherwise \end{cases} \qquad (4)$$

とする．

Initialize $\dot{Q}(s,a)=0$
Repeat(each episode):
  Initialize s
  Repeat(for each step of episode)
    Choose a from s using policy delived from $\dot{Q}, \dot{I}_t$
    Take action a, observe r, s'
    $\dot{Q}(s_{t-k},a_{t-k}) \leftarrow (1-\alpha)\dot{Q}(s_{t-k},a_{t-k})$
      $+\alpha(r_{t+1}+\gamma\dot{Q}^{(t)}_{\max}\dot{u}(k))$
    $s \leftarrow s'$
    $\dot{I}_t = \dot{Q}(s_t,a_t)/\dot{\beta}$
  Until(s is terminal)

図 2: $\dot{Q}$-learning の手続き

さらに，学習速度の向上のために，適格度トレースを用いる．

$$\begin{aligned}\dot{Q}(s_{t-k},a_{t-k}) &\leftarrow (1-\alpha)\dot{Q}(s_{t-k},a_{t-k}) \\ &\quad +\alpha(r_{t+1}+\gamma\dot{Q}^{(t)}_{\max})\dot{u}(k) \qquad (5)\\ \dot{u}(k) &= \beta^{k+1} \qquad (6)\end{aligned}$$

ただし，$k=0,1,\cdots,N_e$ で，N_e をトレース数とよぶ．トレース数は，価値まで遡って行動価値の更新をステップ数を定めるパラメータである．Q-learning と同じように，学習は報酬が与えられる状態の行動価値から，この状態へと遷移する状態の行動価値の方へと進んでいく．このとき，$\dot{\beta}$ だけ位相回転を加えながら価値を伝播させていく．図 2 に $\dot{Q}$-learning の手続きを，図 3 に価値の更新のブロック図を示す．

2.3 複素価値関数のための方策

複素強化学習での方策は，複素価値関数と内部参照値との相互作用によって各々の行動の選択確率を定める．本稿では，この相互作用を内積の実部としているので，形式的には従来の方策における価値を $Re[\dot{Q}(s,a)\bar{\dot{I}}]$ で置き換えた表現となる．すなわち，$Re[\dot{Q}(s,a)\bar{\dot{I}}]$ をその文脈にあった実効的な価値とみなして方策を計算する．従来の Q-learning とあわせて Boltzmann 方策がよく用いられているのに対し，本稿では $\dot{Q}$-learning とあわせて 複素価値関数のための Boltzmann 方策を用いる．

複素価値関数のための Boltzmann 方策は，文脈にあった実効的な価値の大きさに対して行動の選択確率が Gibbs 分布と

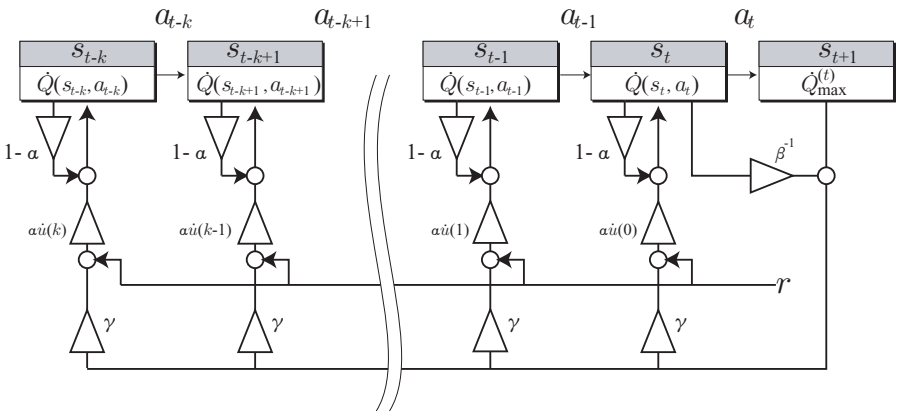


図 3: $\dot{Q}$-learning のブロック図

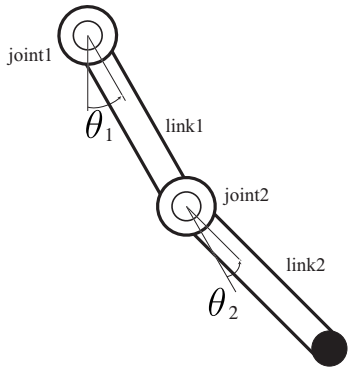


図 4: Acrobot

なるような方策である．この方策は，次式で定式化される．

$$\pi(s,a) = \frac{e^{Re[\dot{Q}(s,a)\bar{I}]/T}}{\sum_{a' \in \mathcal{A}} e^{Re[\dot{Q}(s,a')\bar{I}]/T}} \tag{7}$$

3. Acrobot の振り上げタスク

複素強化学習が不完全知覚を含む簡単な迷路タスクを達成することができた．[4] 本稿では，より複雑なタスクとして 角速度センサを利用できない Acrobot の振り上げタスクを導入する．このタスクでは，エージェントが観測するすべての状態は速度について不完全知覚があり，すでに明らかにしている迷路タスクとは異なるクラスのタスクである。

Acrobot は 図 4 に示す 2 リンク 2 関節のロボットである．第 1 関節は受動関節であり，自由に回転する．第 2 関節は駆動関節であり，取り付けられたアクチュエータによってトルクを加えることができる．

振り上げタスクにおけるエージェントの目標は，予め設定された目標の高さ h_0 までロボットの先端を振り上げることである．ただし，既に述べたように第 1 関節には直接トルクを加えることができないので，第 2 関節のアクチュエータをつかって振れ幅を励振させなければならない．

Acrobot のダイナミクスは，それぞれのリンクの長さ $l_1 = l_2 = 1$[m]，それぞれのリンクの重心までの距離 $l_{c1} = l_{c2} = 0.5$[m]，それぞれのリンクの質量 $m_1 = m_2 = 1$[kg]，リンクの重心まわりの慣性モーメント $I_1 = I_2 = 1$[kg · m^2]，重力加速度 $g = 9.8$[m/s^2] を考慮して次のような運動方程式で定式化できる．

$$d_{11}\ddot{\theta_1} + d_{12}\ddot{\theta_2} + h_1 + \phi_1 = 0 \tag{8}$$
$$d_{21}\ddot{\theta_1} + d_{22}\ddot{\theta_2} + h_2 + \phi_2 = \tau \tag{9}$$

$$\begin{aligned} d_{11} &= m_1 r_1^2 + m_2 l_1^2 + m_2 r_2^2 \\ &\quad + 2m_2 l_1 r_2 \cos\theta_2 + I_1 + I_2 && (10) \\ d_{12} &= d_{21} = m_2 r_2^2 + m_2 l_2 r_2 \cos\theta_2 + I_2 && (11) \\ d_{22} &= m_2 r_2^2 + I_2 && (12) \\ h_1 &= -m_2 l_1 r_2 (2\dot{\theta}_1 + \dot{\theta}_2)\dot{\theta}_2 \sin\theta_2 && (13) \\ h_2 &= m_2 l_1 r_2 \dot{\theta}_1^2 \sin\theta_2 && (14) \\ \phi_1 &= (m_1 r_1 + m_2 l_1) g \sin\theta_1 \\ &\quad + m_2 r_2 g \sin(\theta_1 + \theta_2) && (15) \\ \phi_2 &= m_2 r_r g \sin(\theta_1 + \theta_2) && (16) \end{aligned}$$

ただし，$\theta_1, \theta_2 \in [-\pi, \pi]$，$\dot{\theta}_1 \in [-4\pi, 4\pi]$，$\dot{\theta}_2 \in [-9\pi, 9\pi]$ を状態空間に対する機械的な制約条件とする．表記を簡単にするため，式 8 〜 16 に限りドットで時間微分を表した．

このタスクは，システムが強い非線形性をもっていること，状態空間が連続空間であることから，強化学習のなかでも難しいタスクとして知られている.[5]

本稿では，角速度センサがはじめから壊れていた場合を想定している．完全知覚な状態空間を構成するためには各関節の角度と角速度が必要であるから，このタスクでは各角速度に関して常に不完全知覚が生じる．

このタスクでは，リンクを左右にふって励振をしなければならないため，両方のリンクが下を向いている状態が 1 周期に 2 回現れる．角速度センサから情報を得ることができれば，これら二つの状態は区別され適切なトルクを付加することができる．しかし，角速度センサからの情報が得られず不完全知覚が生じているとこれらの二つの状態は区別されず，同じ状態として扱われる．この場合，既に得られたエネルギーを逆に消失させる方向にトルクを付加する状況が起こってしまい，振幅を励振させることができない．

表 1: Acrobot 振り上げタスク 実験パラメータ

Q-learning	
学習係数 α	0.25
割引率 γ	0.9
ボルツマン温度 T	0.1
$\dot{\text{Q}}$-learning	
学習係数 α	0.25
位相回転 $\dot{\beta}$	exp $(j$[deg])
割引率 γ	0.9
ボルツマン温度 T	0.4
トレース数 N_e	6

表 2: Acrobot センサの離散化レベル

sensor type	quantization threshold[%]
angular	1,5,10,30,50,70,90,95,99

4. シミュレーション実験

不完全なセンサの Acrobot の振り上げタスクに Q-learning と $\dot{\text{Q}}$-learning を適用し，比較を行った．エージェントの 1 回の行動選択を 1 ステップ，初期状態から目標高さに振り上げるまでを 1 エピソード，500 エピソードを 1 学習とした．

4.1 実験条件

すべてのエピソードは，両方のリンクが下を向いて静止している状態から始めた．そして，エージェントが行動を繰り返すうち，ロボットの先端の高さが

$$h = -l_1 \cos\theta_1 - l_2 \cos(\theta_1 + \theta_2) \tag{17}$$

目標の高さ h_0 を超えた場合にはエピソードを終了しエージェントには 報酬 $r = 100$ を与えた．実験では，0.05[sec] ごとに

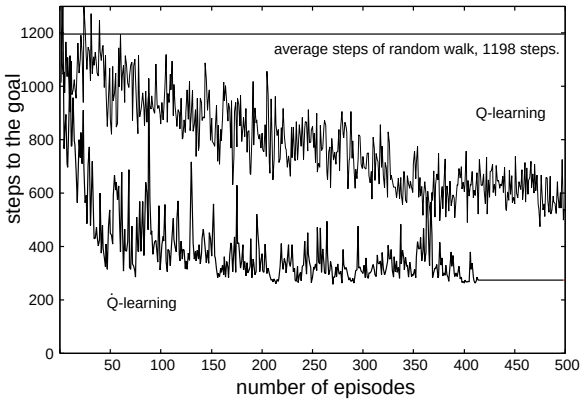


図 5: 学習曲線

運動方程式を逐次計算し，エージェントには 0.2[sec] に 1 度行動としてトルク $\tau \in [-1, 0, 1]$ を選択させた. また $h_0 = 0.5$ とした.

簡単のため，角度センサと角速度センサは表 2 のように量子化を行った.

なお，運動方程式の計算には Runge - Kutta 法 (4 次) を用いた. 表 1 に予備実験によって定めた実験パラメータを示す.

4.2 実験結果

Q-learning と Q̇-learning のそれぞれについて経過エピソードと，成功した 50 学習の平均ステップ数の関係を図 5 に示した. エージェントがランダムに行動をした場合（ランダムウォーク）の平均も同一平面に示した.

Q-learning はエピソードを経過してもステップ数が低下せず，適切な振る舞いを獲得できていないことがわかる. 一方，Q̇-learning はエピソードが経過するごとに振り上げるのに必要なステップ数が低下し，収束している.

図 6 は学習によって獲得された典型的な振る舞いを示している. この図ではそれぞれの関節の角度を時系列順に示している. Q-learning は 250 ステップ前後で 振幅が大きくなっているが，不完全知覚問題が生じているために振幅を消失させる方向にトルクを加えてしまい，その後減衰している. これに対して Q̇-learning では，振幅を大きく励振させタスクを達成していることを確認できる. (b) の (1) と (2) は グラフが表す実際の様子をストロボ的に表している. (1) ではエピソード開始から 250 ステップまで，(2) では 251 ステップからエピソード終了までの動きを表している. この図からもエネルギーを蓄えたのち，第 2 リンクを振り上げる様子を確認できる.

5. 考察

400 ステップ以下の学習回数は，Q-learning で 30 学習，Q̇-learning で 47 学習である. この実験では，Q̇-learning は Q-learning と比べて良質な学習を行うことができた.

Q̇-learning は Q-learning よりも学習速度が速いのは，とった行動の文脈が強化され，重点的に学習が行われたためである.

6. おわりに

複素強化学習の枠組みと Q̇-learning のアルゴリズムについて述べ，角速度センサを利用できない Acrobot の振り上げ制御タスクを導入した. シミュレーション実験の結果，常にエージェントの観測に不完全知覚が生じていても，複素強化学習は有効であることが確認された.

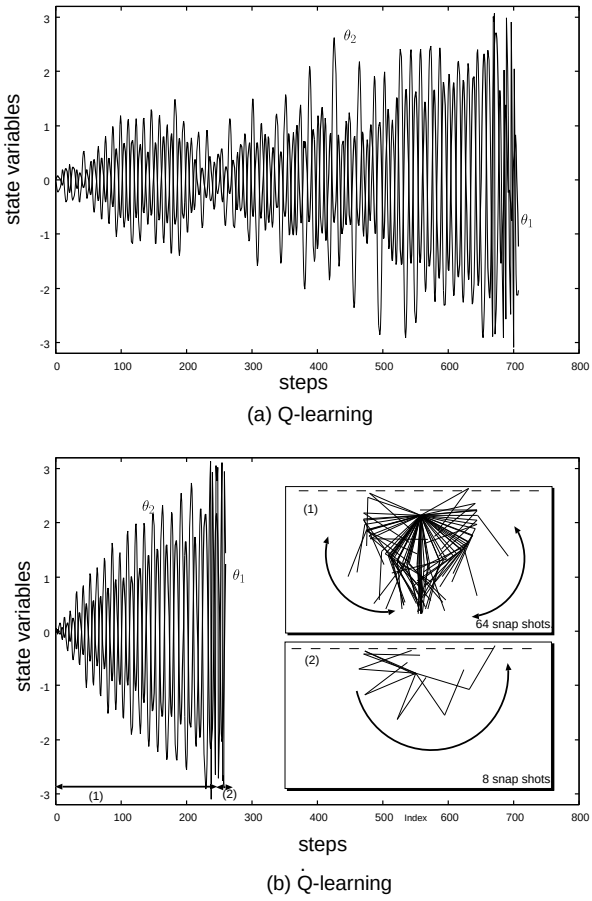


図 6: 典型的な振る舞い

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] T.Hamagami, T.Shibuya, S.Shimada, “Complex-Valued Reinforcement Learning,” Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp.3235-3530, 2006.

[5] 吉本潤一郎, 石井信, 佐藤雅昭, “オンライン EM アルゴリズムによる強化学習法の acrobot 制御への応用,” 電子情報通信学会論文誌, J83-D-II(3), pp.1024-1033, 2003.