

連続的な状態空間における複素強化学習

Complex-valued Reinforcement Learning in Continuous State Space

有田 秀彰†

澁谷 長史†

濱上 知樹†

Hideaki ARITA Takeshi SHIBUYA Tomoki HAMAGAMI

1 はじめに

学習主体であるエージェントが環境に応じた動作を自律的に獲得する手法として、強化学習がある。強化学習は、エージェントにとるべき行動をあらかじめ設計しておく必要はなく、目的に対し報酬を与えるだけで行動を学習するといった特長がある。

しかし、センサから得られる状態の観測に必要な情報が不十分である場合、状態を一意に識別できず、異なる状態を同じ状態と認識してしまうことがある。このように状態を誤って認識してしまうことで適切な行動を学習することが困難になる問題を不完全知覚問題という。不完全知覚問題に対し、従来の強化学習の行動価値を複素数で定義した複素強化学習が有効であることが示されている [1]。

これまでの複素強化学習の研究では、センサから得られる観測値が連続である場合、観測値を離散化していた。しかし、離散化の度合いをあらかじめ適切に設定しておくことは困難である。従来の強化学習では連続な観測値を扱うために動径基底関数 (Radial Basis Function: RBF) ネットワークを用いた手法が提案されている [2]。

そこで、本稿では複素強化学習に複素数値化された RBF ネットワークを用いる手法を提案する。これにより、不完全知覚を含む連続的な状態空間における学習が期待できる。

2 複素強化学習

2.1 複素強化学習の原理

複素強化学習は従来の強化学習の行動価値を複素数で定義したもので、環境に対する事前知識や多くのメモリを必要としないといった特長がある。

複素強化学習では内部参照値という複素数で定義された変数を導入する。内部参照値は、エージェントが現在の状態に到達するまでの過去の履歴である文脈を表す。これにより、エージェントが異なる状態を同じ状態とみなしても、内部参照値が異なれば状態を区別することができる。選択する行動を変えることができる。

複素強化学習では、複素行動価値の絶対値が大きい行動ほど、また、複素行動価値と内部参照値の位相が近い行動ほど選択されやすいとしている。本稿ではこれらを満たすために複素行動価値と内部参照値の内積が大きい行動ほど選択されやすい方策を用いる。

2.2 Q-learning

Q-learning とは複素強化学習の一手法であり、強化学習の代表的な手法である Q-learning の行動価値を複素数で定義した手法である [1]。ある時刻 t においてエー

ジェントが $\mathbf{x}_t$ を観測し、行動 a_t を選択するときの複素行動価値 $\dot{Q}(\mathbf{x}_t, a_t)$ は次式により更新される。

$$\dot{Q}(\mathbf{x}_t, a_t) \leftarrow \dot{Q}(\mathbf{x}_t, a_t) + \alpha \delta_t \quad (1)$$

$$\delta_t = (r_{t+1} + \gamma \dot{Q}_{\max}^{(t)})\beta - \dot{Q}(\mathbf{x}_t, a_t) \quad (2)$$

$$\dot{Q}_{\max}^{(t)} = \dot{Q}(\mathbf{x}_{t+1}, a) \quad (3)$$

$$a = \operatorname{argmax}_{a' \in \mathcal{A}(\mathbf{x}_{t+1})} \left(\operatorname{Re} \left[\dot{Q}(\mathbf{x}_{t+1}, a') \bar{I}_t \right] \right) \quad (4)$$

時刻 t における内部参照値 I_t は次式で更新される。

$$\dot{I}_t = \dot{Q}(\mathbf{x}_t, a_t) / \beta \quad (t \geq 0) \quad (5)$$

$$\dot{I}_{-1} = \dot{Q}(\mathbf{x}_0, a) \quad (6)$$

$$a = \operatorname{argmax}_{a' \in \mathcal{A}(\mathbf{x}_0)} \left| \dot{Q}(\mathbf{x}_0, a') \right| \quad (7)$$

β は位相の回転量を表し、複素行動価値や内部参照値の位相を変化させる。

また、行動選択には複素強化学習における Boltzmann 方策を用いる。 T は Boltzmann 温度である。

$$\pi_{I_{t-1}}(\mathbf{x}_t, a) = \frac{\exp \left(\operatorname{Re} \left[\dot{Q}(\mathbf{x}_t, a) \bar{I}_{t-1} \right] / T \right)}{\sum_{a' \in \mathcal{A}(\mathbf{x}_t)} \exp \left(\operatorname{Re} \left[\dot{Q}(\mathbf{x}_t, a') \bar{I}_{t-1} \right] / T \right)} \quad (8)$$

3 RBF ネットワーク

RBF ネットワークとは RBF の線形和により任意の関数を表現する手法である。RBF には一般にガウス関数が用いられる。RBF ネットワークにより、連続な観測値に対して連続な行動価値、つまり行動価値関数を表すことが可能となる。

Q-learning の更新式より、RBF ネットワークを用いて行動価値関数を表現する手法が提案されている [2]。各行動ごとの行動価値関数を $Q(\mathbf{x}, a)$ 、観測値の数を m 個とし、時刻 t の観測値を $\mathbf{x}_t \in \mathbb{R}^m$ 、行動を a_t 、 k 番目の基底関数を $\phi_{a,k}$ 、基底関数の数を n_a とする。ただし基底関数の数の増減は行わないとする。

最急降下法を用い、RBF パラメータである重み $\omega_{a,k}$ 、分散値 $\sigma_{a,k}$ 、中心値 $\mu_{a,k} \in \mathbb{R}^m$ を次式により更新することで行動価値関数を表現する。

$$Q(\mathbf{x}, a) = \sum_{k=1}^{n_a} \omega_{a,k} \phi_{a,k}(\mathbf{x}) \quad (9)$$

$$\phi_{a,k}(\mathbf{x}) = \exp \left(-\frac{\|\mathbf{x} - \mu_{a,k}\|^2}{\sigma_{a,k}^2} \right) \quad (10)$$

$$\delta_t = r_{t+1} + \gamma \max_a Q(\mathbf{x}_{t+1}, a) - Q(\mathbf{x}_t, a_t) \quad (11)$$

$$\omega_{a,k} \leftarrow \omega_{a,k} + \alpha_\omega \delta_t \phi_{a,k}(\mathbf{x}_t) \quad (12)$$

$$\sigma_{a,k} \leftarrow \sigma_{a,k} + \alpha_\sigma \delta_t \omega_{a,k} \frac{\|\mathbf{x}_t - \mu_{a,k}\|^2}{\sigma_{a,k}^3} \phi_{a,k}(\mathbf{x}_t) \quad (13)$$

$$\mu_{a,k} \leftarrow \mu_{a,k} + \alpha_\mu \delta_t \omega_{a,k} \frac{\mathbf{x}_t - \mu_{a,k}}{\sigma_{a,k}} \phi_{a,k}(\mathbf{x}_t) \quad (14)$$

† 横浜国立大学大学院工学府

ここで δ_t は Q-learning の TD 誤差, α_ω , α_σ , α_μ はそれぞれ $\omega_{a,k}$, $\sigma_{a,k}$, $\mu_{a,k}$ の学習率である.

4 提案手法

本稿では, 連続な観測値に対して連続な複素行動価値, つまり複素行動価値関数を表す複素 RBF ネットワークを提案する. 提案手法では, Q-learning の複素行動価値関数 $\dot{Q}(x, a)$ を実数部分 $Q_1(x, a)$ と虚数部分 $Q_2(x, a)$ に分け, それぞれを RBF ネットワークで表す. $i = 1$ は複素 RBF ネットワークの実数部分, $i = 2$ は虚数部分を表し, 基底関数を $\phi_{i,a,k}$, 基底関数の数を $n_{i,a}$ とする. 重み $\omega_{i,a,k}$, 分散値 $\sigma_{i,a,k}$, 中心値 $\mu_{i,a,k} \in \mathbb{R}^m$ をそれぞれの RBF ネットワークにおいて更新し, 複素行動価値関数を表現する.

実数部分と虚数部分それぞれの RBF パラメータを次式により更新することで, 不完全知覚を含む連続的な状態空間に対応した複素行動価値関数を表現することができる.

$$\dot{Q}(x, a) = Q_1(x, a) + jQ_2(x, a) \quad (15)$$

$$Q_i(x, a) = \sum_{k=1}^{n_{i,a}} \omega_{i,a,k} \phi_{i,a,k}(x) \quad (i = \{1, 2\}) \quad (16)$$

$$\phi_{i,a,k}(x) = \exp\left(-\frac{\|x - \mu_{i,a,k}\|^2}{\sigma_{i,a,k}^2}\right) \quad (17)$$

$$\dot{\delta}_t = (r_{t+1} + \gamma \dot{Q}_{\max}^{(t)})\beta - \dot{Q}(x_t, a_t) \quad (18)$$

$$= \delta_{1,t} + j\delta_{2,t} \quad (19)$$

$$\omega_{i,a,k} \leftarrow \omega_{i,a,k} + \alpha_\omega \delta_{i,t} \phi_{i,a,k}(x_t) \quad (20)$$

$$\sigma_{i,a,k} \leftarrow \sigma_{i,a,k} + \alpha_\sigma \delta_{i,t} \omega_{i,a,k} \frac{\|x_t - \mu_{i,a,k}\|^2}{\sigma_{i,a,k}^3} \phi_{i,a,k}(x_t) \quad (21)$$

$$\mu_{i,a,k} \leftarrow \mu_{i,a,k} + \alpha_\mu \delta_{i,t} \omega_{i,a,k} \frac{x_t - \mu_{i,a,k}}{\sigma_{i,a,k}} \phi_{i,a,k}(x_t) \quad (22)$$

以上のように, 実数部分と虚数部分それぞれの RBF パラメータは従来の RBF パラメータと同様の式により更新する. 提案手法はこのような簡潔な処理により複素行動価値関数を表現することが可能となる.

5 シミュレーション実験・結果

提案手法の有効性を確認するため, 連続的な状態空間における強化学習の標準的なタスクである図1の Mountain Car タスクで実験を行った. 車は前方へ加速, 加速しない, 後方へ加速の3種類の行動を選択することができる. 車は前方の山頂へたどり着くと報酬が与えられる. しかし, エンジンの出力が小さいため後方の山を登り, 勢いをつけてから前方の山を登る必要がある. 本実験では車の位置 $x \in [-1.2, 0.5]$ を観測可能とし, 速度を観測不可能とする. よって車の状態観測は常に不完全知覚を含む観測となる. 実験で用いたパラメータを表1, 2に示す. 車の1回の行動を1ステップ, スタートからゴールへたどり着くまでを1エピソードとする. 300 エピソードを1学習とし, 20 学習行った.

表1 強化学習のパラメータ

	Q-learning	Q-learning
Boltzmann 温度 T	150/(1+episode)	0.5
位相回転 β	-	$\exp(j\pi/180)$
学習率 α	0.1	0.1
割引率 γ	0.7	0.7
報酬 r	100	100

表2 RBF ネットワークの学習パラメータ

	RBF	複素 RBF
ω の学習率 α_ω	0.001	0.01
σ の学習率 α_σ	0.001	0.001
μ の学習率 α_μ	0.001	0.001
各基底関数の数 $n_{i,a}$	30	30

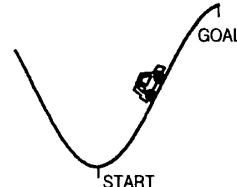


図1 Mountain Car

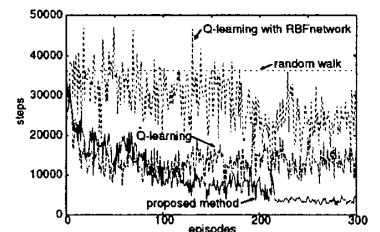


図2 平均学習曲線

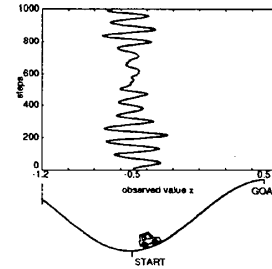


図3 従来手法による軌跡

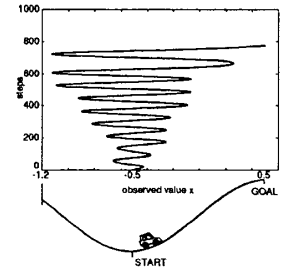


図4 提案手法による軌跡

従来手法である Q-learning に RBF ネットワークを用いた学習, 観測値を 10 等分に離散化した Q-learning による学習と比較した結果を図2に示す. 図2より提案手法は従来手法に比べ, 効率よく目的を達成できることがわかる. 本実験では車の速度が観測不可能であるため, どちらの方向へ進んでいるかを区別できない. Q-learning に RBF ネットワークを用いた学習は図3のように振幅を小さくする行動を選択し, 車の勢いを弱めてしまうことで目的の達成が困難になっている. 提案手法では図4のように振幅を大きくする行動をとることで目的を達成している.

6 おわりに

複素 RBF ネットワークを用いて, 連続的な状態空間における複素強化学習を提案した. 実験により, 提案手法が不完全知覚を含む連続的な状態空間において有効であることを示した.

参考文献

- [1] 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.
- [2] 淵田孝康, 前原正和, 森邦彦, 村島定行: “強化学習の手法を用いた RBF ネットワークの学習”, 電子情報通信学会技術研究報告 NC99-113, pp.157-163, 2000.