5C-3

周期的環境に対するフーリエ混合強化学習法

野田 五十樹 †
(独) 産業技術総合研究所 情報技術研究部門

1 まえがき

動的環境に対する適用は強化学習を実用化する上で重要な課題である。これに対し筆者らはこれまで、学習の過程でも時々刻々変化するような環境において強化学習のパラメータ (ステップサイズパラメータ) を適切に調節する方法 (Recursive Adaptation of Stepsize Parameter, RASP) を提案してきた [1]。一方、実世界の問題においては、周期的に変化する環境など、学習すべき対象が規則的に絶えず変化することがよくある。本稿ではそのなかで特に一定の周期で変化する環境を取り上げ、そのようななかで適応が可能な強化学習の方式を提案する。

2 混合周期複素強化学習

2.1 混合周期複素時系列

本稿では、周期的に変化する値を表すため、以下のような複素時系列を用いる。

$$z_t = \mathring{z}\mathrm{e}^{j\omega t}$$

ここで、j は虚数単位、ω は実角速度、t は時刻を表す整数、$\mathring{z}$ はこの時系列の初期値である。実際には複数の周期が重畳する場合も考慮し、以下のような混合フーリエ時系列に拡張して議論を進めていく。

$$z_t = \sum_k z_{kt} = \sum_k \mathring{z}_k \mathrm{e}^{j\omega_k t} \quad (1)$$

ただし ω_k は、$\forall k \in K : 0 \leq \omega_k \leq \pi$ および $\forall k, k' \in K, k \neq k' : \omega_k \neq \omega_{k'}$ を満たすものとする。

2.2 混合周期複素時系列の指数移動平均

通常の実数時系列 $\{r_t\}$ の EMA $\{x_t\}$ は $x_{t+1} = x_t + \alpha(r_t - x_t)$ で与えられる。

ここで、α は EMA のステップサイズパラメータ (実数) である。この EMA を混合フーリエ時系列 ((1) 式) に拡張した、以下のような混合フーリエ EMA $\{y_t\}$ (Fourier Mixture EMA, FM-EMA) を考える

$$\begin{aligned} \forall k : y_{kt+1} &= \mathrm{e}^{j\omega_k}\left(y_{kt} + \alpha(r - \Re(y_t))\right) \quad (2) \\ y_t &= \sum_{k'} y_{k't} \end{aligned}$$

ただし、$\Re(z)$ は複素数 z の実部を取出す関数である。

この EMA に対し、以下の定理が成立する。

定理 2.1
(2) 式 は *2* 乗誤差 $\frac{1}{2}(r_t - \Re(y_t))^2$ の平均を極小化する最急降下法となっている。*[2]* □

2.3 混合周期複素強化学習法

強化学習における期待報酬 Q を時間の関数とし、(2) 式を適用して以下のような更新式を考える。

$$\begin{aligned} Q_t(s_t, a_t) &= \sum_k Q_{kt}(s_t, a_t) \\ Q_{kt+1}(s_t, a_t) &= \mathrm{e}^{j\omega_k} \cdot (Q_{kt}(s_t, a_t) \\ &\quad + \alpha(r_t + \gamma \max_{a'} Q'_t(s_{t+1}, a') \\ &\quad - Q_t(s_t, a_t))) \quad (3) \\ Q'_t(s_{t+1}, a')) &= \sum_{k'} \mathrm{e}^{j\omega_{k'}} Q_{kt}(s_{t+1}, a') \end{aligned}$$

また、時刻 t に於て選択されなかった状態-行動対 $\langle s, a \rangle$ については、(3) 式による修正がなかったとして、以下のように $t+1$ 用に期待報酬関数を求めるものとする。

$$\begin{aligned} &\forall \langle s, a \rangle \neq \langle s_t, a_t \rangle, \forall k \\ &\quad Q_{kt+1}(s, a) = \mathrm{e}^{j\omega_k} Q_{kt}(s, a) \quad (4) \end{aligned}$$

Fourier Mixture Reinforcement Learning for Periodic Environment

† Itsuki Noda, ITRI, AIST <i.noda@aist.go.jp>

定理 2.1 により、この (3) 式で与えられる Q 値の修正方法は、通常の強化学習と同じく、環境から得られる報酬 r と状態遷移後からのバックアップの和に Q 値を近づけるように働く。一方、ここでの各 Q は混合フーリエ時系列として表されており、周期的に変化する r に適応する形で学習が進むと期待できる。そこで、この (3) 式および (4) 式による期待報酬関数の修正を混合フーリエ強化学習法 (Fourier Mixture Reinforcement Learning, FM-RL) と呼ぶことにする。

3 実験

前節でもとめた FM-RL の動作を確認するために、以下のような実験を行った。

異なる周期 (周期 5 と周期 3) ので変動する報酬が別々のセル (0,0) と (0,2) に与えられるようなグリッドワールドを考える (図 1)。この設定では、より大きい報酬を提供するセルへタイミング良く移動することが必要となる。

FM-RL による学習後の Q 値の変化と greedy 戦略によるエージェントの移動の様子を図 2 に示す。この図では、セル (0,0), (0,1) (0,2) における各行動の Q 値の時間変化が上部 3 つのグラフで、また、エージェントの移動結果が最下部のグラフで示されている。前の実験と同じく、セル (0,0) および セル (0,2) における stay の行動に対応する Q 値は、そのセルに与えられる報酬とほぼ同期した変化をとるように学習されている。同時に、セル (0,0) における down の Q 値は、セル (0,2) における報酬の周期に対応した成分が重畳された変化を示していることが読み取れる。つまり、セル (0,2) における周期的な報酬の変化が、セル (0,1) を介して学習されていることになる。これも (3) 式の Q' の効果が多段にわたって波及しているものであり、グリッドに与えられる複数の報酬の変化をうまく Q の関数の中に取り込んでいることが分かる。結果として、この例においても学習後のエージェントは獲得報酬を最大化する周期 15 の行動をとっていることが、図 2 のグラフより分かる。

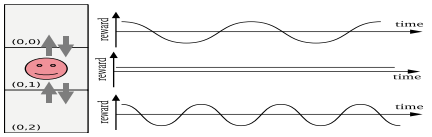

図 1: 周期的報酬を 1 つ含むグリッドワールド

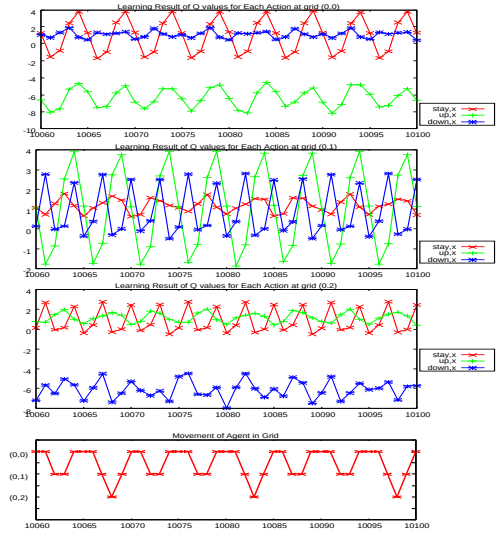

図 2: Exp.3-b: 周期的報酬を 2 つ含むグリッドワールドでの行動学習

4 おわりに

本稿では、混合周期的複素 EMA が、変動する値を近似するための最急勾配法になっていることを利用して、時間的に変動する状態-行動価値 Q を学習する混合周期的複素強化学習 (PCM-RL) を提案した。

謝辞 本研究は科研費 21500153 の助成を受けたものである。

参考文献

[1] 野田五十樹. 指数的移動平均 2 乗誤差の最小化によるステップサイズパラメータの調整法. In *JAWS* **予稿集**. JAWS, 10 月 2009.

[2] 野田五十樹. 複素数指数移動平均を用いた強化学習による周期的環境への適応. In *JAWS 2010* **予稿集**, 10 月 2010.