適正度の履歴を用いた Actor-Critic 強化学習アルゴリズムの解析

○木村 元,　　小林 重信
東京工業大学 大学院総合理工学研究科

1　はじめに

Actor-critic 学習システムは，報酬の遅れを取り扱い可能な強化学習法の一つであり，Q-learning[7] と比較すると以下の実用的利点を持つ．

1. 連続値を含む行動出力への拡張が容易
2. 確率的政策により，隠れ状態を含む環境やマルチエージェント，ゲームなどへの適用も可能
3. 従来の教師付学習を Actor へ適用することにより，エキスパートの知識との統合が容易

従来，critic の Value function 表現と actor の政策表現の両方に十分な関数近似能力がなければ学習できなかった．本研究では actor の政策修正アルゴリズムとして確率的傾斜法 [2] を用いることにより，critic の能力が不十分でも学習できることを示す．

一般的 actor-critic アルゴリズムの構成を図 1 に示す [1]．以下の手順により学習する．

1. エージェントは環境において状態 x_t を観測する．Actor は，確率的政策 π に従って行動 a_t を実行する．
2. Critic は報酬 r_t を受け取り，次の状態 x_{t+1} を観測し，actor への強化信号として図 1 の TD-error を計算する．γ $(0 \leq \gamma \leq 1)$ は割引率，$\hat{V}(x)$ は Critic が出力した利得 (return) の期待値，つまり割引報酬の期待値を表す．
3. TD error を用いて actor の行動選択確率を更新する．(TD-error) > 0 ならば，実行した行動 a は比較的好ましいもの考えられるので，この選択確率を増やす．逆に (TD-error) < 0 ならば，a の選択確率を減らす．
4. TD 法 [6] を用いて critic の value の推定値を更新する．例えば TD(0) ならば, $\hat{V}(x) \leftarrow \hat{V}(x) + \alpha$ (TD-error), ただし α は学習率である．
5. $t \leftarrow t+1$ としてステップ 1 から繰り返す．

2　Actor への確率的傾斜法の適用

観測 X においてエージェントが行動 a を選択する確率を政策 π と呼び，関数 $\pi(a, W, x)$ で表す．$\pi(a, W, X)$

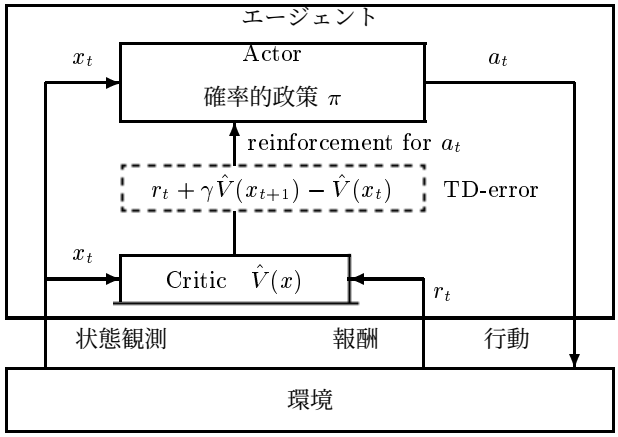

図 1: 一般的 actor-critic アルゴリズムの枠組

は行動 a の集合が連続値の場合は確率密度関数である．$W = (w_1, w_2, \cdots, w_i \cdots)$ は政策パラメータを表す．エージェントは W を調節することにより π を変える．確率的傾斜法では，実行した行動 a_t と w_i との相関である適正度 $e_i(t)$ を計算する．過去の適正度ほど割引率 γ で減衰して合計した適正度の履歴 $D_i(t)$ を用いることで，行動 a_t だけでなく全ての過去の行動系列を強化するのが特徴である．Actor の行動選択確率の処理ステップ 3 を以下のようにする．

$$
\begin{aligned}
\text{Eligibility:} \quad e_i(t) &= \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, x_t)\Big), \\
\text{Eligibility Trace:} \quad D_i(t) &= e_i(t) + \beta D_i(t-1), \\
\Delta w_i(t) &= (\text{TD-error})\, D_i(t) \\
W &\leftarrow W + \alpha_p \, \Delta W(t),
\end{aligned}
$$

ただし，β $(0 \leq \beta < 1)$ は適正度の履歴の割引率，α_p は学習定数を表す．

3　提案手法の解析

提案手法は $\beta = 0$ の場合のみ，図 1 に示した従来手法の完全なサブクラスとなる．$\beta = \gamma$ の場合，提

案手法の $\Delta w_i(t)$ の合計は

$$\begin{aligned}\sum_{t=0}^{\infty}\Delta w_i(t) &= \sum_{t=0}^{\infty}\Big(r_t+\gamma\hat{V}(x_{t+1})-\hat{V}(x_t)\Big)\ D_i(t)\\ &= \sum_{t=0}^{\infty}\Big(r_t+\gamma\hat{V}(x_{t+1})-\hat{V}(x_t)\Big)\left(\sum_{\tau=0}^{t}\gamma^{t-\tau}e_i(t)\right)\\ &= \sum_{t=0}^{\infty}e_i(t)\left(\sum_{\tau=t}^{\infty}\gamma^{\tau-t}\Big(r_\tau+\gamma\hat{V}(x_{\tau+1})-\hat{V}(x_\tau)\Big)\right)\\ &= \sum_{t=0}^{\infty}e_i(t)\left(\left(\sum_{\tau=t}^{\infty}\gamma^{\tau-t}r_\tau\right)-\hat{V}(x_t)\right) \qquad (1)\\ &= \sum_{t=0}^{\infty}e_i(t)\Big(V_t-\hat{V}(x_t)\Big) \qquad (2)\end{aligned}$$

ただし $t<0$ のとき $D_i(t)=0$ とする．Critic の推定した $\hat{V}(x_t)$ と a_t は独立である．REINFORCE algorithm の定理 [8] より，提案手法はある条件下では critic の推定した $\hat{V}$ の勾配ではなく，実際の時系列より計算される利得 (actual return) V_t の勾配の方向へ政策を改善する．このとき critic の $\hat{V}$ は報酬基底として働き，学習の効率には関係するが，政策の改善方向の期待値には関係しない．すなわち $\beta=\gamma$ の場合，提案手法は critic が不完全でも学習できる．

4 実験および考察

以下の線形 2 次形式制御問題を設定した．

- 環境の状態遷移規則：$x_{t+1}=x_t+a_t+noise$，ただし $noise=N(0,0.5)$，$x_i,a_i\in R$
- 直接報酬：$r_t=-x_t^2-a_t^2$
- 学習目標：割引報酬の合計 $\sum_{t=0}^{\infty}\gamma^t r_t$ の最大化
- Actor の政策表現：政策パラメータ $W=(w_1,w_2)$

$$\pi(a,W,x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(\frac{-(a-\mu)^2}{2\sigma^2}\Big),\ \text{where}$$
$$\mu=w_1\,x_t \quad,\quad \sigma=\frac{1}{1+\exp(-w_2)}.$$

- Critic の状態離散化：$-4\le x\le 4$ を 3 等分
- Critic のアルゴリズム：TD(0) 法

実験結果を図 2 に示す．$\beta=0$ とした従来の枠組では Critic の関数近似能力が不十分なため学習できないが，$\beta=\gamma$ とすることにより学習できた．

本研究では DP に基づく手法と勾配法による政策改善法との融合を示した．

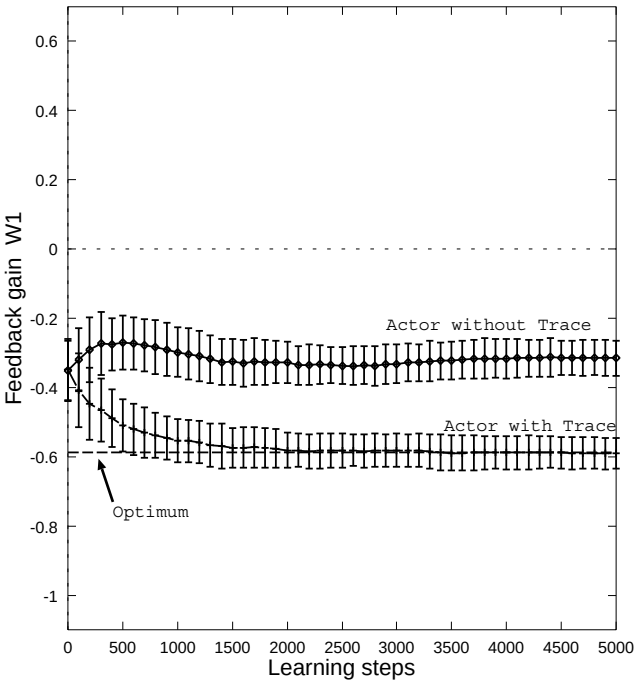

図 2: 従来手法と提案手法の比較．割引率 γ=0.9，横軸は学習ステップ，縦軸は獲得したフィードバックゲイン w_1，100 試行の平均と標準偏差をあらわす．

参考文献

[1] Crites, R. H. and Barto, A. G.: An Actor/Critic Algorithm that is Equivalent to Q-Learning, *Advances in Neural Information Processing Systems 7*, pp. 401-408 (1994).

[2] 木村 元, 山村 雅幸, 小林 重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近, 人工知能学会誌, Vol.11, No.5, pp.761-768 (1996).

[3] Kimura, H. & Kobayashi, S.: An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Function, *15th International Conference on Machine Learning*, pp.278–286 (1998).

[4] Lin, C. J. and Lin, C. T.: Reinforcement Learning for An ART-Based Fuzzy Adaptive Learning Control Network, *IEEE Transactions on Neural Networks*, Vol.7, No. 3, pp. 709-731 (1996).

[5] Singh, S. P., & Sutton, R.S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning 22*, pp. 123-158 (1996).

[6] Sutton, R. S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning 3*, pp. 9-44 (1988).

[7] Watkins, C. J. C. H., & Dayan, P.: Technical Note: *Q*-Learning, *Machine Learning 8*, pp. 55-68 (1992).

[8] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp. 229-256 (1992).