計測自動制御学会 第 19 回自律分散システムシンポジウム（2007 年 1 月）pp.67–72

適正度の履歴を用いた自然勾配 Actor-Critic 法

Natural Gradient Actor-Critic using Eligibility Traces

九州大学 大学院工学研究院 海洋システム工学部門　○木村 元

○ Hajime Kimura, Graduate School of Engineering, Kyushu University

Abstract: By policy gradient theory, policy improvement RL methods making use of action value should adopt eligibilities of policy parameters for a basis function to approximate action values. The action values are represented by a linear parameterization of the eligibilities of the policy parameters, and then, the parameters of the value approximator are also corrsponding to a natural-gradient with respect to the policy parameters. This paper introduces eligibility traces of policy parameters to the natural-gradient policy improvement method, and demonstrate the performance in some non-Markov environments.

1　はじめに

環境との試行錯誤を通し、行動を改善したり最適な行動を見出していく強化学習法には、Q-learning のように政策に依存せずに最適な状態-行動評価値を推定する off-policy と呼ばれる方法と、SARSA や Actor-Critic のように政策に依存した状態評価値や行動評価値を推定して政策を改善する on-policy と呼ばれる方法がある。多数のセンサやモーターで構成されるロボットの制御規則を強化学習で獲得しようとする場合、高次元で膨大な状態空間や行動空間を扱う必要が生じることから、あらゆる状態-行動における経験を十分に行う必要のある off-policy 強化学習法よりも、保持している政策を改善していく on-policy 法のほうが学習速度の点で有望であると考えられる。その on-policy 強化学習法の一つである Actor-Critic 法は、状態や行動を評価する critic 部分と、確率的政策を保持する actor 部分より構成され、critic の評価値に基づいて actor の政策を改善していくものであり、実ロボットの強化学習において多くの実績がある。著者らは先行研究において actor の適正度の履歴（eligibility trace）を利用した actor-critic 法を提案し、critic の機能が損なわれると確率的傾斜法 [2] に等価となり、また適正度の履歴を使うことで、環境にある程度の非マルコフ性が存在しても学習可能であることを示した [3][4]。一方で、近年 actor の政策改善方法として「自然勾配」と呼ばれる勾配方向へ改善する方法の有効性が報告されている [1]。これは、やや乱暴な説明かもしれないが目的関数に対して 2 次微分の方向まで考慮した勾配方向へ政策パラメータを更新していく方法、すなわちニュートン法に近いと考えられる。ニュートン法は、通常 Hesse 行列の逆行列を必要とするなど計算コストがかかるが、強化学習の場合、特に価値関数近似のための基底関数として actor の適正度を用いた場合において、そのような重い計算が一切不要であるという驚くべき解析が示されている [7]。本研究では、著者らが先行研究において提案した actor の適正度の履歴を利用した actor-critic 法について、自然勾配の方向へ政策を更新するよう修正を加え、いくつかの実験を通じてその特徴を観察する。

2　問題の定式化

状態空間を $\mathcal{S}$，行動空間を $\mathcal{A}$，上下界を持つ実数の集合を $\mathcal{R}$ と表す．各時刻 t でエージェントは状態観測 $s_t \in \mathcal{S}$ に基づいて行動 $a_t \in \mathcal{A}$ を実行し，状態遷移に伴う報酬 $r_t \in \mathcal{R}$ を得る．本論文が環境のモデルとして仮定するマルコフ決定過程 (MDP) では，一般に次の状態や報酬は確率的で，その分布は s_t と a_t にのみ依存する．MDP では次の状態 s_{t+1} は遷移確率 $T(s_t, a, s_{t+1})$ に従って決まり，報酬 r_t も期待値 $r(s_t, a)$ によって与えられる．エージェントは予め $T(s_t, a, s_{t+1})$ や $r(s_t, a)$ についての知識を持っていない．強化学習の目的はエージェントのパフォーマンスを最適化する政策を得ることである．無限期間のタスクにおける自然な評価規範として，割引報酬の合計がある．$E\{\cdot\}$ は期待値，γ は割引率を表すものとすると，MDP での評価関数は以下のように定義される：

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, \pi\right], \qquad (1)$$

MDP における学習の目的は，各状態 s において式 (1) で定義される評価値を最大化するような最適政策を見つけることである．

3　政策勾配定理と自然勾配

Actor-Critic 強化学習法は、状態や行動を評価する critic 部分と、確率的政策を保持する actor 部分より構成され、critic の評価値に基づいて actor の政策を改善していくものであることから、critic における関数近似能力が不十分だと actor の政策改善に支障をきたす問題がある。筆者らの先行研究 [3][4] では、actor に適正度の履歴（eligibility trace）を導入することでこの問題を解決した。一方で、actor の政策を改善するのに必要十分な価値関数近似のための基底関数は何なのかについての研究が進められ、政策勾配定理（Policy Gradient Theorem）[9] により、actor の政策を改善するのに必要な価値関数近似のための基底関数は、適正度 (eligibility) すなわち行動選択確率関数（政策）の対数を政策パラメータで偏微分した関数であることが示されている。

確率的政策 π は，状態 s において行動選択確率 $\pi(s, a)$ に従って行動 a を実行する．この行動選択確率関数 $\pi(s, a)$ は政策パラメータ群 $\boldsymbol{\theta}$ によって値を調節でき，パラメータ群の各要素 θ によって偏微分可能な関数であるものとする．エージェントの学習目標である目的関数 $\rho(\pi)$ を次の

ように定式化する：

$$\rho(\pi) = V^{\pi}(s_0)\ , \qquad Q^{\pi}(s,a) = E\{V_t | s_t = s, a_t = a, \pi\}\ . \tag{2}$$

ここで $d^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi)$ とすると，政策勾配定理 [9] より以下の式が与えられる：

$$\frac{\partial \rho}{\partial \theta} = \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \frac{\partial \pi(s,a)}{\partial \theta} Q^{\pi}(s,a)\ . \tag{3}$$

次に，真の Q 関数 $Q^{\pi}(s,a)$ を関数近似によって $\hat{Q}^{\pi}(s,a)$ と表すことを考える。ただしこの近似 Q 関数 $\hat{Q}^{\pi}(s,a)$ は適正度 $\frac{\partial \ln \pi(s,a)}{\partial \theta}$ を各要素としたベクトル $\nabla_\theta \ln \pi(s,a)$ を基底関数とし、この各適正度に対応した要素を持つパラメータベクトル $\boldsymbol{w}$ との線形結合および状態価値関数 $V^{\pi}(s)$ を用いて以下のように表す：

$$\hat{Q}^{\pi}(s,a) = \left(\nabla_\theta \ln \pi(s,a)\right)^T \boldsymbol{w} + V^{\pi}(s)\ . \tag{4}$$

真の Q 関数 $Q^{\pi}(s,a)$ と近似 Q 関数を $\hat{Q}^{\pi}(s,a)$ との 2 乗誤差が最小になるようパラメータベクトル $\boldsymbol{w}$ を調節すると，政策勾配定理 [9] より以下が成り立つ：

$$\frac{\partial \rho}{\partial \theta} = \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \frac{\partial \pi(s,a)}{\partial \theta} \hat{Q}^{\pi}(s,a)\ . \tag{5}$$

つまり、critic において Q 関数を近似する場合に actor の適正度を基底関数として利用すると，Q 関数が近似であるにもかかわらず真の勾配方向へ政策を更新できることを意味する。一般に Q 関数の表現では，以下のように状態評価値 $V^{\pi}(s)$ と，advantage と呼ばれる評価値 $A^{\pi}(s,a)$ を用いて $Q^{\pi}(s,a) = V^{\pi}(s) + A^{\pi}(s,a)$ のように表すことができる。(4) 式の右辺左側の項 $\left(\nabla_\theta \ln \pi(s,a)\right)^T \boldsymbol{w}$ がこの advantage 関数 $A^{\pi}(s,a)$ を近似している。

この (5) 式の 1 次偏導関数で示される政策勾配を使った様々な政策改善アルゴリズム [6] が提案される一方で、2 次偏導関数も考慮に入れた「自然勾配 (natural gradient)」と呼ばれる勾配を用いる政策改善法が提案されている [1]。この種の勾配法は、1 次偏導関数しか使用しない勾配法に比べて格段に少ない反復回数で極値を得ることができ、一般的な最適化手法としてニュートン法が知られているが、一般に 2 次偏導関数行列（ヘッセ行列）の逆行列を求めるという大きな計算コストが要求される。ところが強化学習における政策パラメータに関する自然勾配は、情報理論的に興味深い性質を持ち、上記の逆行列計算が不要になる。Advantage 関数を構成する (4) 式のパラメータベクトル $\boldsymbol{w}$ において、ある政策パラメータ θ の適正度に対応する要素を w_θ と表す。これらが (5) 式を満たすとき、目的関数 $\rho(\pi)$ に対するパラメータ θ の自然勾配 $\frac{\partial \tilde{\rho}}{\partial \theta}$ は、以下のように驚くべき単純な式で表される [7]：

$$\frac{\partial \tilde{\rho}}{\partial \theta} = w_\theta\ . \tag{6}$$

すなわち、基底関数として actor の適正度を用いた advantage 関数を構成するパラメータベクトル $\boldsymbol{w}$ は、対応する政策パラメータ $\boldsymbol{\theta}$ の自然勾配になっている。以上の先行研究の知見より導かれる actor-critic アルゴリズムは Fig.1 のように構成される。このアルゴリズムは、政策 π のもとでの Q 値を推定することになるため、SARSA アルゴリズムの拡張と考えることができ、マルコフ性を満たす環境において、極めて効率良く政策改善を行うことが期待されるが、実ロボットなどのように非マルコフ性が存在する場合において学習できないという問題がある。

1. 状態 s_t を観測し，行動選択確率（または確率密度）$\pi(s_t, a_t)$ に従って行動 a_t を選択して実行する．その結果，報酬 r_t と遷移先の状態 s_{t+1} を観測する．
2. 政策 π のパラメータ $\boldsymbol{\theta}$ についての適正度 $\frac{\partial \ln \pi(s_t,a_t)}{\partial \theta}$ を計算し、これを各要素としたベクトル $\nabla_\theta \ln \pi(s_t, a_t)$ とパラメータベクトル $\boldsymbol{w}$ を用いて advantage の推定値 $\hat{A}^{\pi}(s_t, a_t)$ を計算：
$$\hat{A}^{\pi}(s_t, a_t) = \left(\nabla_\theta \ln \pi(s_t, a_t)\right)^T \boldsymbol{w}$$
3. 以下のように TD_error を計算して critic における状態評価値を更新する：
$$\begin{aligned} \mathrm{TD\_err}_t &= \Big(r_t + \gamma \hat{V}^{\pi}(s_{t+1})\Big) - \hat{V}^{\pi}(s_t)\ , \\ \hat{V}^{\pi}(s_t) &\leftarrow \hat{V}^{\pi}(s_t) + \alpha \mathrm{TD\_err}_t\ , \end{aligned}$$
ただし α は学習率，γ は報酬の割引率である．
4. 上記の TD_error と適正度を用いて advantage 関数のパラメータベクトル $\boldsymbol{w}$ を更新する：
$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha \Big(\mathrm{TD\_err}_t - \hat{A}^{\pi}(s_t, a_t)\Big) \nabla_\theta \ln \pi(s_t, a_t)$$
ただし α は学習率である。
5. Advantage 関数のパラメータベクトル $\boldsymbol{w}$ を用いて、政策パラメータ θ を更新：
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_p \boldsymbol{w}\ ,$$
ただし α_p は政策パラメータの学習率である。
6. 時刻を $t \leftarrow t+1$ に進めて step 1 より繰返す.

Figure 1: 自然勾配 ActorCritic 法

4 適正度の履歴を用いる自然勾配 ActorCritic 法の提案

Fig.2 に適正度の履歴を用いた自然勾配 actor-critic 法を示す。Q 値の表現と更新が、状態価値関数 $\hat{V}^{\pi}(s)$ と advantage 関数 $\hat{A}^{\pi}(s,a)$ の 2 つに分かれているが、SARSA(λ) アルゴリズムと同じになる。適正度の履歴を減衰させるパラメータ λ は SARSA(λ) の λ と同じである。advantage 関数のパラメータベクトル $\boldsymbol{w}$ を更新する処理において、新しい学習率 β を導入しているが、これは適正度の履歴 $\bar{\boldsymbol{e}}_t$ の絶対和ノルムを時間的に平均した値を $<\bar{\boldsymbol{e}}>$ と表すと、$\beta = \alpha / <\bar{\boldsymbol{e}}>$ とすると良い。実験では、時刻 t よりおよそ 100 ステップ分の過去の適正度の履歴の絶対和ノルムを

平均して用いている。

5 実験

Fig.3 に示すロボットに本手法を適用する．学習目標は，ロボットを前進させるために，アームを足のように作用させる動作の獲得である．関節は位置制御のサーボモーターによって角度を制御される．各時間ステップにおいて，エージェントは 8 つの関節モータの角度および 4 つの足先のタッチセンサの状態という 12 個の状態量を要素とした 12 次元ベクトルを観測する：関節角度 $\phi_1, \ldots \phi_8$ は 8 次元連続空間で定義され，タッチセンサ $\phi_9, \ldots \phi_{12}$ は 0 または 1 の 2 値である．行動は，関節角度の目標値を指示し，8 次元ベクトル $(a^1, \ldots a^8)$ の各要素がそれぞれ関節角度を表す．行動ベクトルの各要素は $[0, 1]$ の範囲に限定される．行動が選ばれると，モータは指示された目標位置へ動きはじめる．関節角度が指示された位置まで動くか，あるいはタッチセンサの値が変化すると，状態遷移の結果として報酬が与えられ，次の時刻へ進む．関節のモータが目標位置まで動く途中でセンサの値が変化すると，そこで意思決定イベントが発生して動きが打ち切られるため，次のステップでの関節角度は行動として出力された目標角度には一致しない．よって状態遷移には不確実性が存在する．報酬はボディが前進した距離と方向で与えられる．ロボットが後退した場合，報酬は負値になる．

状態表現のための特徴ベクトルは，Critic において状態評価値 $V^\pi(s)$ を表すためのものと，Actor において行動選択確率（政策）を表すためのものの 2 種類を用意した．前者は均一タイルコーディングで，状態空間を等分割のタイルで 3^{12} に分割する．後者は線形コーディングで，$X = (\phi_1, \phi_2, \cdots \phi_{12}, 1-\phi_1, 1-\phi_2, \cdots 1-\phi_{12})$ で与えられる 12 次元の連続ベクトルである．どんな状態においても特徴ベクトルの絶対和ノルムは一定に保たれる．

政策関数はコーシー分布に基づく連続分布関数を修正して用いた．それは行動 $(a_{(1)}, a_{(2)})$ の各要素が $[-1, 1]$ の範囲に限定されていることによる．エージェントは各モータ i 毎に式 8 のコーシー分布に従って，$[-1, 1]$ の範囲に収まっているサンプルを得るまでランダムにサンプリングを行い，得たサンプルを行動 $a_{(i)}$ として出力する．

$$P(a) = \frac{1}{\pi\sigma_{(i)}} \frac{1}{1 + \frac{(a-\mu_{(i)})^2}{\sigma_{(i)}^2}}. \tag{8}$$

ここでパラメータ $\mu_{(i)}$ および $\sigma_{(i)}$ は以下で与えられる：

$$\mu_{(i)} = \frac{2}{1 + \exp\left(-\sum_j x_j\, \theta_{j,(i)}\right)} - 1, \tag{9}$$

$$\sigma_{(i)} = \frac{1}{1 + \exp\left(-\theta_{\sigma,(i)}\right)} + 0.1, \tag{10}$$

ただし $\theta_{j,(i)}$ および $\theta_{\sigma,(i)}$ は政策パラメータである．$\theta_{j,(i)}$ 中の j は，特徴ベクトル X の要素 x_j に対応付けられる．このとき政策関数 $\pi(s, a)$ は以下で表される：

$$\pi(s, a_{(i)}) = \frac{1}{S_i \sigma_{(i)}} \frac{1}{1 + \frac{(a-\mu_{(i)})^2}{\sigma_{(i)}^2}}, \tag{11}$$

1. 状態 s_t を観測し，行動選択確率（または確率密度）$\pi(s_t, a_t)$ に従って行動 a_t を選択して実行する．その結果，報酬 r_t と遷移先の状態 s_{t+1} を観測し、再び確率 $\pi(s_{t+1}, a_{t+1})$ に従って行動 a_{t+1} を実行する．

2. 政策 π のパラメータ $\boldsymbol{\theta}$ についての適正度 $\frac{\partial \ln \pi(s_t, a_t)}{\partial \theta}$ を計算し、これを各要素としたベクトル $\nabla_\theta \ln \pi(s_t, a_t)$ とパラメータベクトル $\boldsymbol{w}$ を用いて advantage の推定値 $\hat{A}^\pi(s_t, a_t)$ および $\hat{A}^\pi(s_{t+1}, a_{t+1})$ を計算：

$$\begin{aligned} \hat{A}^\pi(s_t, a_t) &= \left(\nabla_\theta \ln \pi(s_t, a_t)\right)^T \boldsymbol{w}\ , \\ \hat{A}^\pi(s_{t+1}, a_{t+1}) &= \left(\nabla_\theta \ln \pi(s_{t+1}, a_{t+1})\right)^T \boldsymbol{w} \end{aligned}$$

3. 状態評価値 $V^\pi(s)$ と advantage より Q 値を計算:

$$\begin{aligned} \hat{Q}^\pi(s_t, a_t) &= \hat{V}^\pi(s_t) + \hat{A}^\pi(s_t, a_t) \\ \hat{Q}^\pi(s_{t+1}, a_{t+1}) &= \hat{V}^\pi(s_{t+1}) + \hat{A}^\pi(s_{t+1}, a_{t+1}) \end{aligned}$$

4. 以下のように適正度の履歴を更新する：

$$\bar{\boldsymbol{e}}_t = \nabla_\theta \ln \pi(s_t, a_t) + \lambda\gamma \bar{\boldsymbol{e}}_{t-1}$$

ただし λ は適正度の履歴の減衰率 $(0 \le \lambda \le 1)$

5. 状態評価値 $V^\pi(s)$ と適正度の履歴を用いて advantage 関数のパラメータベクトル $\boldsymbol{w}$ を更新：

$$\begin{aligned} \delta_t &= r_t + \gamma \hat{Q}^\pi(s_{t+1}, a_{t+1}) - \hat{Q}^\pi(s_t, a_t) \\ \boldsymbol{w} &\leftarrow \boldsymbol{w} + \beta\, \delta_t \bar{\boldsymbol{e}}_t \end{aligned} \tag{7}$$

ただし γ は割引率，β は学習率である。

6. Advantage 関数のパラメータベクトル $\boldsymbol{w}$ を用いて、政策パラメータ θ を更新：

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_p \boldsymbol{w}\ ,$$

ただし α_p は政策パラメータの学習率である。

7. 以下のように TD_error を計算して critic における状態評価値を更新する：

$$\begin{aligned} \text{TD\_err}_t &= \left(r_t + \gamma \hat{V}^\pi(s_{t+1})\right) - \hat{V}^\pi(s_t)\ , \\ \hat{V}^\pi(s_t) &\leftarrow \hat{V}^\pi(s_t) + \alpha \text{TD\_err}_t\ \ , \end{aligned}$$

ただし α は学習率である。

8. 時刻を $t \leftarrow t+1$ に進めて step 1 より繰返す.

Figure 2: 適正度の履歴を用いる自然勾配 ActorCritic 法

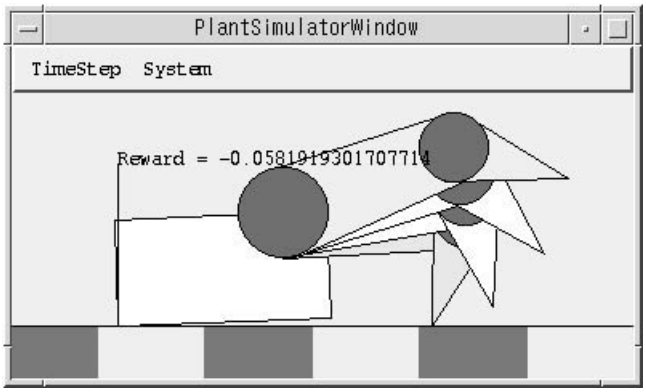

Figure 3: 多自由度ほふくロボットシミュレータ．右側が前方．足は 4 本で，各足にはモータ 2 個ずつ計 8 個取り付けられている．各足の先端にはタッチセンサが付いている．状態は 8 関節の角度と 4 個のタッチセンサの値の計 12 次元で，行動は 8 個の関節の目標値で計 8 次元である．

ただし $\pi(s,a)=\prod_i \pi(s,a_{(i)})$ で，S は以下で与えられる：

$$\begin{aligned} S_i &= \int_{-1}^{1} \frac{1}{\sigma_{(i)}} \frac{1}{1+\frac{(a-\mu_{(i)})^2}{\sigma_{(i)}^2}} da \\ &= \tan^{-1}\left(\frac{1-\mu_{(i)}}{\sigma_{(i)}}\right) - \tan^{-1}\left(\frac{-1-\mu_{(i)}}{\sigma_{(i)}}\right). \end{aligned} \quad (12)$$

このとき $\mu_{(i)}$ と $\sigma_{(i)}$ の適正度は以下で計算される：

$$\begin{aligned} &\frac{\partial \ln \pi(s,a_{(i)})}{\partial \mu_{(i)}} \\ &= \frac{1}{\sigma_{(i)}}\left(\frac{1}{S_i}\left(\frac{1}{1+\left(\frac{1-\mu_{(i)}}{\sigma_{(i)}}\right)^2} - \frac{1}{1+\left(\frac{-1-\mu_{(i)}}{\sigma_{(i)}}\right)^2}\right)\right. \\ &\quad \left. +2(a_{(i)}-\mu_{(i)})S_i\pi(s,a_{(i)})\right), \end{aligned} \quad (13)$$

$$\begin{aligned} &\frac{\partial \ln \pi(s,a_{(i)})}{\partial \sigma_{(i)}} \\ &= \frac{1}{\sigma_{(i)}^2 S_i}\left(\frac{1-\mu_{(i)}}{1+\left(\frac{1-\mu_{(i)}}{\sigma_{(i)}}\right)^2} - \frac{-1-\mu_{(i)}}{1+\left(\frac{-1-\mu_{(i)}}{\sigma_{(i)}}\right)^2}\right) \\ &\quad -\left(1-\frac{(a_{(i)}-\mu_{(i)})^2}{\sigma_{(i)}^2}\right) S_i\pi(s,a_{(i)}) \end{aligned} \quad (14)$$

式 9, 9, 13, 14 より，政策パラメータの適正度は：

$$\begin{aligned} \frac{\partial \ln \pi(s,a_i)}{\partial \theta_{j,(i)}} &= \frac{\partial \mu_{(i)}}{\partial \theta_{j,(i)}} \frac{\partial \ln \pi(s,a_{(i)})}{\partial \mu_{(i)}} \\ &= \frac{x_j}{2}(1+\mu_{(i)})(1-\mu_{(i)}) \frac{\partial \ln \pi(s,a_{(i)})}{\partial \mu_{(i)}} \end{aligned} \quad (15)$$

$$\begin{aligned} \frac{\partial \ln \pi(s,a_i)}{\partial \theta_{\sigma,(i)}} &= \frac{\partial \sigma_{(i)}}{\partial \theta_{\sigma,(i)}} \frac{\partial \ln \pi(s,a_{(i)})}{\partial \sigma_{(i)}} \\ &= \sigma_{(i)}(1-\sigma_{(i)}) \frac{\partial \ln \pi(s,a_{(i)})}{\partial \sigma_{(i)}} \end{aligned} \quad (16)$$

本実験では，Fig.1 や Fig.2 中の適正度 $\frac{\partial \ln \pi(s_t,a_t)}{\partial \theta}$ は式 15,16 で計算される．全ての実験において割引率 $\gamma=0.9$, critic の価値関数 $\hat{V}^\pi(s)$ の学習率 $\alpha=0.1$，advantage 関数 $\hat{A}^\pi(s,a)$ の学習率 $\beta=\alpha/<\bar{\boldsymbol{e}}>$ ただし $<\bar{\boldsymbol{e}}>$ は時刻 t よりおよそ 100 ステップ分の過去の適正度の履歴 $\bar{\boldsymbol{e}}_t$ の絶対和ノルムを平均した値を表す。Actor の政策更新の学習率として $\alpha_p=0.02$ を用いたが、学習が不安定になるのを回避するために更新時に自然勾配 $\boldsymbol{w}$ の絶対和ノルムを計算し、これが 1 より大きい場合は自然勾配 $\boldsymbol{w}$ をその絶対和ノルム $|\boldsymbol{w}|$ で除して正規化してから更新した。自然勾配 $\boldsymbol{w}$ の絶対和ノルムが 1 より小さい場合はそのまま更新した。

比較対象として、Actor の適正度の履歴を用いる Actor-Critic アルゴリズム [4][5] を少し修正し、パラメータ設定が同じになるよう配慮した。Fig.2 で示されるアルゴリズムでは、政策パラメータ θ が自然勾配 $\boldsymbol{w}$ で更新され、絶対和ノルムが 1 より大きい場合はその値で除した勾配方向へ更新しているので、従来の Actor-Critic 法の更新部分を変更し、$\Delta\theta$ を平均化することで政策勾配を推定し、この値を用いて同じように政策パラメータ θ を更新するように変更した。よって、Actor の政策更新の学習率 $\alpha_p=0.02$ を同じに設定すれば、性能の違いは政策勾配の推定性能の差を意味することになる。

Fig.4 は、従来の ActorCritic アルゴリズムと Fig.2 で提案したアルゴリズムを全く同一の関数近似・行動選択確率関数のもとで多自由度ほふくロボットシミュレータにて学習性能を比較した結果を示す。ただし、双方の手法ともに適正度の履歴のパラメータ設定は $\lambda=1$、すなわち Actor の適正度の履歴を用いて学習している。横軸はエージェントの学習ステップ数、縦軸は対応する学習ステップ数における政策を、学習を停止して 500 ステップ分動作させた場合の 1 ステップあたりの報酬である。政策は 1000 ステップ毎に学習を停止させて評価を行った。グラフは 10 試行の平均と標準偏差を示す。学習によって得られた政策には 6 倍近い性能の差が見られ、学習の早さも自然勾配を用いた方法のほうが 10 倍以上高速である。

Fig.5 は、同様の比較の結果を示すが、双方の手法ともに適正度の履歴のパラメータ設定は $\lambda=0$，すなわち適正度の履歴を用いない設定である。このとき、双方ともに得られる政策の質が半分程度に下落し、また学習が進むと政策の質が悪化する傾向が見られる。これは適正度の履歴を用いていないことから、非マルコフ性に対処できないアルゴリズムになっていると考えられ、環境の有する非マルコフ性が影響しているものと推測される。とはいえ、自然勾配を用いた手法は、従来手法よりはるかに良い政策を獲得している。

Fig.6 も同様に比較の結果を示すが、双方の手法ともに適正度の履歴のパラメータ設定は $\lambda=1$ だが，Critic の値（状態評価値 $V^\pi(s)$）をゼロに固定し、Critic を用いない設定になっている．この場合、双方の手法ともにあまり影響を受けていないように見える。Actor の適正度の履歴を用いる ActorCritic 法 [4][5] では、Critic の性能が不完全でも政策改善が可能であるという特徴を有していたが、自然勾配を用いる ActorCritic 法でも適正度の履歴を用いることにより全く同様の効果が得られることが分かる。

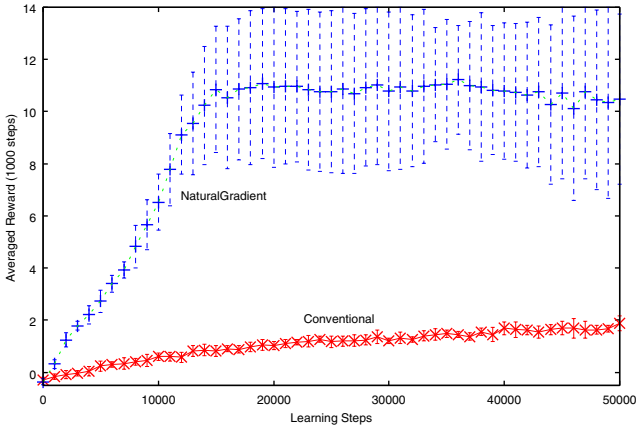

Figure 4: 多自由度ほふくロボットシミュレータにおける従来手法と自然勾配を利用した方法による学習を 10 試行平均して比較した様子．双方の手法ともに適正度の履歴のパラメータ設定は $\lambda = 1$.

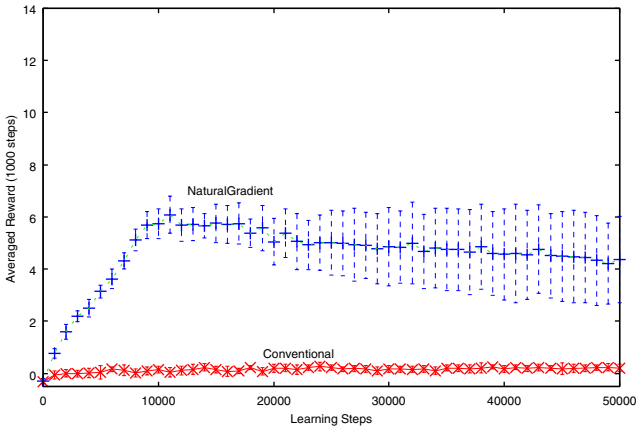

Figure 5: 多自由度ほふくロボットシミュレータにおける従来手法と自然勾配を利用した方法による学習を 10 試行平均して比較した様子．双方の手法ともに適正度の履歴のパラメータ設定は $\lambda = 0$，すなわち適正度の履歴を用いない設定.

6 考察

【性能の差について】
本実験より以前に、離散的な 7 状態-2 行動の環境で予備実験を行ったのだが、そのとき従来法との性能差はあまり観察されなかった。本実験の設定では政策パラメータの次元数が極めて高いため、単なる勾配法と 2 次微分を考慮した勾配法との性能差が顕著に現れたのではないかと考えられる。単なる勾配では、勾配が小さくほぼ平坦と思われる政策パラメータ領域においても、自然勾配で 2 次微分の方向まで考慮すると政策改善が滞ることなく行える点や、2 次微分まで考慮することにより（局所的な）最適点まで近道で到達できることが、特に高次元の空間において顕著な効果を示すのではないかと考えられる。従来手法 [5] では、実ロボットの歩行動作学習におよそ 100 分かかっていたのだが、本実験結果から考えると 10 分程度で学習できると予想され、強化学習の実用性がかなり高まるものと期待される。しかしながら、得られた政策の質には試行毎にばらつきが生じた。これは勾配法であるために局所解に陥りやすい傾向を有しているためと考えられる。

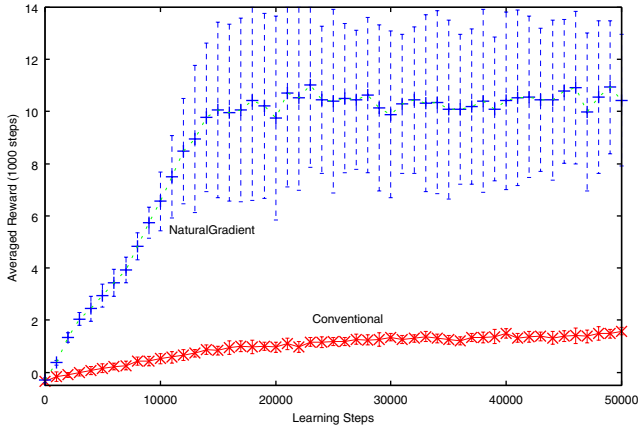

Figure 6: 多自由度ほふくロボットシミュレータにおける従来手法と自然勾配を利用した方法による学習を 10 試行平均して比較した様子．双方の手法ともに適正度の履歴のパラメータ設定は $\lambda = 1$ だが，Critic の値（状態評価値 $V^{\pi}(s)$）をゼロに固定し、Critic を用いない設定.

【適正度の履歴の影響】
Fig.5 において、双方の手法ともに適正度の履歴を用いない設定では、双方ともに得られる政策の質が半分程度に下落し、また学習が進むと政策の質が悪化する傾向が見られたが、これは適正度の履歴を用いていないことから、非マルコフ性に対処できないアルゴリズムになっていると考えられ、環境の有する非マルコフ性が影響しているものと推測される。ある程度政策を学習した後に、性能が悪化していく理由については、学習が進むと政策のランダムさが減少していくと考えられ、これが非マルコフの環境において誤った方向へと導かれるために性能が悪化していくのではないかと思われる。

【状態評価値の影響】
Fig.6 において、双方の手法ともに適正度の履歴は用いるが Critic を用いない設定では、双方の手法ともにあまり影響を受けないという類似した傾向が見られた。これは、適正度の履歴が同じような機能・役割を持っているためと考えられる。Fig.2 で示したアルゴリズムは、SARSA(λ) アルゴリズムの一種と考えられるが、Q 関数の表現が状態価値関数と Advantage 関数に分けられているため、状態価値関数の推定が不可能なのに Advantage 関数の推定ができている点は大変不思議である。

【実装上の工夫について】
Fig.6 のアルゴリズムは、このまま実装しても学習は大変不安定で、すぐに重み変数 $\boldsymbol{w}$ の数値が発散してしまう傾向が見られた。これは、Advantage 関数を表すための基底関数である適正度ベクトルや適正度の履歴ベクトルのノルムの上界や平均が予め計算しておくことが困難であることが主な原因であった。そのため、適正度の履歴ベクトルのノルムの平均値をオンラインで計算して Advantage 関数の学習率 α を調節したところ、学習はかなり安定した。しかしながら、本実験ではコーシー分布の分散をコントロールするパラメータがゼロに近づくと、適正度が極端に大きな値になることがあり、学習が発散するなど不安定な現象がまれに見られた。よって、(10) 式の右辺第 2 項におい

て 0.1 を足すことにより、このパラメータが 0.1 より小さくなることを回避し、適正度の値が極端な値にならないよう配慮した。

7 おわりに

本研究では自然勾配 Actor-Critic 法に適正度の履歴を導入したアルゴリズムを提案した。これは目的関数に関して 2 次微分まで考慮した勾配法と考えられ、単なる勾配法だった従来手法と比較すると、高次元の政策パラメータを持つ強化学習問題において劇的な性能向上が認められた。またこの新しい勾配法は、ある程度の非マルコフ環境でも対処可能で、また状態評価値を推定する Critic 部分が機能しなくても学習可能であるなど従来手法の有する特徴をそのまま引き継いでいることを実験により確認した。本手法を実ロボットの学習へ適用することが今後の課題である。

References

[1] Kakade, S.: A Natural Policy Gradient, *Advances in Neural Information Processing Systems 14*, pp.1531–1538 (2002).

[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] 木村 元, 小林 重信, Actor に適正度の履歴を用いた Actor-Critic アルゴリズム– 不完全な Value-Function のもとでの強化学習, 人工知能学会誌, Vol.15, No.2, pp.267–275 (2000).

[5] 木村 元, 山下 透, 小林 重信：強化学習による 4 足ロボットの歩行動作獲得, 電気学会 電子情報システム部門誌, Vol.122-C, No.3, pp.330–337 (2002).

[6] Konda, V.R. & Tsitsiklis, J.N.: Actor-Critic Algorithms, *Advances in Neural Information Processing Systems 12*, pp. 1008–1014 (2000).

[7] Peters, J., Vijayakumar, S. & Schaal, S.: Reinforcement Learning for Humanoid Robots - policy gradients and beyond, 3rd IEEE-RAS International Conference on Humanoid Robotics (2003).

[8] Sutton, R.S. & Barto, A.: Reinforcement learning: An introduction, *A Bradford Book*, The MIT Press (1998).

[9] Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, *Advances in Neural Information Processing Systems 12 (NIPS12)*, pp. 1057–1063 (2000).