1M2-5

逐次的な自然方策勾配推定法の解析と勾配推定分散の最小化による効率的な強化学習法の提案

Analysys of incremental natural policy gradient method and minimizing the variance of gradient estimation

岩城 諒 *1 Ryo Iwaki　横山 裕樹 *2 Hiroki Yokoyama　浅田 稔 *1 Minoru Asada

*1大阪大学大学院工学研究科 知能・機能創成工学専攻
Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University

*2玉川大学工学部 機械情報システム学科
Department of Intelligent Mechanical Systems, College of Engineering, Tamagawa University

We propose a natural actor critic method based on SARSA(λ). The proposed actor not only estimates unbiased natural policy gradient but also is independent of an approximated value function. Our critic explicitly minimizes the variance of actor's gradient estimation and does not need any auxiliary function, while conventional methods estimate an optimal baseline in order to reduce the variance of natural policy gradient. We apply our actor-critic algorithm to simple pendulum swing-up problem and show that our algorithm can learn more stably than the conventional method.

1. はじめに

強化学習とは，報酬という評価値を手掛かりに，システムが自律的に制御則を獲得するための手法である．学習主体であるエージェントは，将来得られる報酬を最大化する方策すなわち制御則を学習することを目的とする．エージェントは，将来得られる期待収益を価値という形式で推定し，学習に利用する．学習則や報酬の設計次第で，制御対象である環境についての事前知識がなくとも，エージェントは設計者の設定したタスクを達成するための方策を学習できる．

しかし，強化学習を実問題に適用する場合，方策や価値関数の学習に必要な時間が膨大になることが知られている．この原因として，従来の強化学習法では，入力分布に対する方策パラメータの敏感さや相関を無視していることが考えられる．自然方策勾配法 [Kakade 02] は，方策空間の幾何学的構造を学習に利用することでこれらの問題を緩和し，プラトーと呼ばれる学習の停滞現象を回避することができる．しかし，オンラインでの自然方策勾配推定は，学習が不安定になりやすいことが指摘されている [木村 07]．

本研究は，自然方策勾配をオンラインで効率良く推定できる学習則を提案することを目的とする．まず，状態価値関数の推定値に依存せず自然方策勾配を不偏推定可能な学習則を提案する．次に，自然方策勾配推定の分散最小化を目的関数とする状態価値関数の学習則を提案する．提案手法を，単純化した倒立振子の振り上げ問題に適用し，頑健に学習できることを示す．

2. 自然方策勾配法

2.1 方策勾配型強化学習法

強化学習問題として，離散時間マルコフ決定過程（MDP）を考える [Sutton 98]．環境の可能な状態の集合を $\mathcal{S}$，エージェントの選択可能な行動の集合を $\mathcal{A}$，実数の集合を $\mathbb{R}$ とする．各離散時刻 t において，エージェントは環境の状態 $\boldsymbol{s}_t \in \mathcal{S}$ を観測し，n 次元の方策パラメータ $\boldsymbol{\theta} \in \mathbb{R}^n$ によって規定される確率的方策 $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) \equiv \Pr(\boldsymbol{a}|\boldsymbol{s},\boldsymbol{\theta})$ に従い，行動 $\boldsymbol{a}_t \in \mathcal{A}$ を選択する．環境は，状態遷移確率 $\mathcal{T}^a_{ss'} \equiv \Pr(\boldsymbol{s}_{t+1} = \boldsymbol{s}'|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a})$ に従って新しい状態 $\boldsymbol{s}_{t+1}$ に遷移し，有界な報酬関数 $\mathcal{R}^a_{ss'} \equiv \mathbb{E}[r_t|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}, \boldsymbol{s}_{t+1} = \boldsymbol{s}']$ によって定まる期待値に従い，エージェントに即時報酬 $r_t \in \mathbb{R}$ を与える．

エージェントは，以下で定義される期待収益 $\rho(\boldsymbol{\theta})$ を最大化する方策パラメータ $\boldsymbol{\theta}$ を獲得することを学習目的とする．

$$\rho(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,\middle|\, \boldsymbol{s}_0\right]$$

ただし，$\gamma \in [0,1]$ は割引率であり，将来の報酬が現在においてどれだけの価値があるかを決定する．ここで，行動価値関数 $Q^\theta(\boldsymbol{s},\boldsymbol{a})$ を以下のように定義する．

$$Q^\theta(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\sum_{\tau=t}^{\infty}\gamma^{\tau-t} r_\tau \,\middle|\, \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\right]$$

行動価値関数 $Q^\pi(\boldsymbol{s},\boldsymbol{a})$ は，状態 $\boldsymbol{s}$ で行動 $\boldsymbol{a}$ をとり，その後方策 π に従った場合に得られる期待収益である．一般に，行動価値関数 $Q^\pi(\boldsymbol{s},\boldsymbol{a})$ は，状態価値関数 $V^\pi(\boldsymbol{s})$ とアドバンテージ関数 $A^\pi(\boldsymbol{s},\boldsymbol{a})$ を用いて，

$$Q^\pi(\boldsymbol{s},\boldsymbol{a}) = V^\pi(\boldsymbol{s}) + A^\pi(\boldsymbol{s},\boldsymbol{a})$$

と表される．状態価値関数 $V^\pi(\boldsymbol{s})$ は，状態 $\boldsymbol{s}$ の価値を表現し，アドバンテージ関数 $A^\pi(\boldsymbol{s},\boldsymbol{a})$ は，ある状態 $\boldsymbol{s}$ における行動 $\boldsymbol{a}$ の相対的な価値を表現する．

ここで，近似行動価値関数 $\hat{Q}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{w})$ を，近似状態価値関数 $\hat{V}(\boldsymbol{s};\boldsymbol{v})$ と，方策のスコア関数（対数勾配）$\nabla_{\boldsymbol{\theta}} \ln\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$ を用いて，以下のように表す．

$$\hat{Q}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{w}) = \hat{V}(\boldsymbol{s};\boldsymbol{v}) + \left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})\right)^{\mathrm{T}}\boldsymbol{w} \quad (1)$$

ただし，

$$\hat{A}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta},\boldsymbol{w}) = \left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})\right)^{\mathrm{T}}\boldsymbol{w} \quad (2)$$

連絡先: 岩城諒，ryo.iwaki@ams.eng.osaka-u.ac.jp

は近似アドバンテージ関数，$\boldsymbol{v}$ は状態価値関数の近似パラメータ，$\boldsymbol{w}$ はアドバンテージ関数の近似パラメータである．ここで，近似行動価値関数 $\hat{Q}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta},\boldsymbol{v},\boldsymbol{w})$ が局所最適解へ収束したとき，方策勾配定理 [Sutton 00] により，以下が成り立つ．

$$\frac{\partial \rho(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \int_{\mathcal{S}} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \int_{\mathcal{A}} \frac{\partial \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \hat{A}(\boldsymbol{s},\boldsymbol{a})\, d\boldsymbol{a} d\boldsymbol{s} \tag{3}$$

ただし，$d_{\boldsymbol{\theta}}(\boldsymbol{s})$ は方策 $\pi(\boldsymbol{\theta})$ のもとでの状態の分布を表す．

2.2 自然方策勾配法

式 (3) に式 (2) を代入することで，次式が成立する．

$$\boldsymbol{w} = G^{-1}(\boldsymbol{\theta}) \frac{\partial \rho(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{4}$$

ただし，$G(\boldsymbol{\theta})$ はフィッシャー情報行列であり，以下で定義される．

$$G(\boldsymbol{\theta}) = \int_{\mathcal{S}} d_{\boldsymbol{\theta}}(\boldsymbol{s}) \int_{\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})\right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})\right)^{\mathrm{T}} d\boldsymbol{a} d\boldsymbol{s}$$

フィッシャー情報行列 $G(\boldsymbol{\theta})$ は，パラメータ $\boldsymbol{\theta}$ を座標系とする空間の構造を規定する．式 (4) の右辺は自然勾配 [Amari 98] と呼ばれ，パラメータを座標系とする空間上での最急方向を表す．すなわち，式 (1) による関数近似が局所最適を達成すれば，アドバンテージ関数の近似パラメータ $\boldsymbol{w}$ は，方策の自然勾配に一致する [Kakade 02]．

2.3 アドバンテージ関数の近似方法

木村 [木村 07] は，SARSA(λ) に基づくアドバンテージ関数の近似方法を提案した．エージェントは，方策に従って行動を実行する Actor と呼ばれる部分と，価値関数を用いて方策を評価する Critic と呼ばれる部分から構成される．学習の手続きを Alg.1 に示す．

Alg.1 はオンラインで自然方策勾配を推定できるが，[木村 07] では学習則の理論的な妥当性が示されていない．また，学習後期に方策が決定的になるにつれて，方策のスコア関数が極端に大きな値を取り，学習が不安定になることが報告されている．

Algorithm 1 [木村 07]

Input:
- Parameterized policy
  $\pi(\cdot\,|\,\cdot\,;\boldsymbol{\theta})$
- Parameterized value function
  $\hat{V}(\cdot\,;\boldsymbol{v})$

Initialization:
- Policy, value function and advantage parameters
  $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{v} = \boldsymbol{v}_0, \boldsymbol{w} = \boldsymbol{w}_0$
- Stepsizes and discount rates
  $\alpha_\theta, \alpha_v, \alpha_w, \beta, \gamma$
- Draw initial state
  $\boldsymbol{s}_0 \sim p(\boldsymbol{s}_0)$
- Draw initial action
  $\boldsymbol{a}_0 \sim \pi(\boldsymbol{a}_0|\boldsymbol{s}_0;\boldsymbol{\theta}_0)$

for $t = 0, 1, 2, \ldots$ **do**

Execution:
- Observe next state
  $\boldsymbol{s}_{t+1} \sim \mathcal{T}^a_{ss'}$
- Observe reward
  $r_t \sim \mathcal{R}^a_{ss'}$
- Draw action
  $\boldsymbol{a}_{t+1} \sim \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)$

Eligibility Traces:
$\overline{\boldsymbol{e}}_{t+1} = \gamma \overline{\boldsymbol{e}}_t + \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)$

Action Value:
$\hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta}_{t-1},\boldsymbol{v}_t,\boldsymbol{w}_t)$
$= \hat{V}(\boldsymbol{s}_t;\boldsymbol{v}_t) + \left(\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}_t|\boldsymbol{s}_t;\boldsymbol{\theta}_{t-1})\right)^{\mathrm{T}} \boldsymbol{w}_t$
$\hat{Q}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\boldsymbol{\theta}_t,\boldsymbol{v}_t,\boldsymbol{w}_t)$
$= \hat{V}(\boldsymbol{s}_{t+1};\boldsymbol{v}_t) + \left(\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)\right)^{\mathrm{T}} \boldsymbol{w}_t$

TD Error:
$\delta^V_t = r_t + \gamma \hat{V}(\boldsymbol{s}_{t+1};\boldsymbol{v}_t) - \hat{V}(\boldsymbol{s}_t;\boldsymbol{v}_t)$
$\delta^Q_t = r_t + \gamma \hat{Q}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\boldsymbol{\theta}_t,\boldsymbol{v}_t,\boldsymbol{w}_t)$
$\quad - \hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta}_{t-1},\boldsymbol{v}_t,\boldsymbol{w}_t)$

Advantage Update:
$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_w\, \delta^Q_t\, \overline{\boldsymbol{e}}_t$

Actor Update:
$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_\theta \boldsymbol{w}_t$

Critic Update:
$\overline{\boldsymbol{c}}_t = \gamma \overline{\boldsymbol{c}}_{t-1} + \nabla_{\boldsymbol{v}} \hat{V}(\boldsymbol{s}_t;\boldsymbol{v}_t)$
$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t + \alpha_v\, \delta^V_t\, \overline{\boldsymbol{c}}_t$

end for

3. 自然方策勾配の不偏推定

本章では，[木村 07] で提案された手法をもとに，自然方策勾配を不偏推定できるアドバンテージ関数の学習則を提案する．提案する学習則では，TD 誤差を次のように計算する．

$$\delta^Q_t = r_t + \gamma\, \hat{Q}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\boldsymbol{\theta}_t,\boldsymbol{v}_t,\boldsymbol{w}_t) - \hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta}_{t-1},\boldsymbol{v}_{t-1},\boldsymbol{w}_{t-1}) \tag{5}$$

既存の自然方策勾配推定法 [木村 07] と提案する更新則との違いは，TD 誤差の計算に用いる価値関数のパラメータを一時刻ずらしていることである．提案手法を用いてアドバンテージ関数の近似パラメータを学習した場合，固定方策のもとで漸近的に以下が成り立つ．

$$\begin{aligned}
&\sum_{t=0}^{\infty} \Delta \boldsymbol{w}_t \\
&= \sum_{t=0}^{\infty} \delta^Q_t \sum_{\tau=0}^{t} \gamma^{t-\tau} \boldsymbol{e}_\tau \\
&= \sum_{t=0}^{\infty} \boldsymbol{e}_t \sum_{\tau=t}^{\infty} \gamma^{t-\tau} \delta^Q_\tau \\
&= \sum_{t=0}^{\infty} \boldsymbol{e}_t \left(\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau - \hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta}_{t-1},\boldsymbol{v}_{t-1},\boldsymbol{w}_{t-1}) \right)
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}\Big[\Delta \boldsymbol{w}_t\Big] \\
&= \lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T}\Delta \boldsymbol{w}_\tau \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\ln\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_{t-1})\left(Q_t^{\theta}(\boldsymbol{s},\boldsymbol{a})-\hat{A}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{t-1},\boldsymbol{w}_{t-1})\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{w}}\hat{A}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{t-1},\boldsymbol{w}_{t-1})\left(Q_t^{\theta}(\boldsymbol{s},\boldsymbol{a})-\hat{A}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{t-1},\boldsymbol{w}_{t-1})\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{w}}\left(Q_t^{\theta}(\boldsymbol{s},\boldsymbol{a})-\hat{A}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{t-1},\boldsymbol{w}_{t-1})\right)^2\right] \qquad (6)
\end{aligned}
$$

式 (6) は，提案手法を用いると，アドバンテージ関数の更新方向の期待値が，未知の真の行動価値関数 Q_t^{θ} との二乗誤差の最急降下方向に一致すること，かつ状態価値関数の推定値には依存しないことを意味する．すなわち，価値推定の近似精度によらず，自然方策勾配を不偏推定する [*1]．

4. 自然方策勾配推定の分散最小化

式 (6) は，アドバンテージ関数近似パラメータの更新値 $\Delta\boldsymbol{\theta}$ の期待値は，状態価値関数の推定値 $\hat{V}(\cdot;\boldsymbol{v})$ に依存しないことを意味する．$\hat{V}(\cdot;\boldsymbol{v})$ の値が不正確でも，更新を繰り返すことで $\Delta\boldsymbol{w}$ の期待値は最適値に近づく．しかし，各時刻の更新は $\hat{V}(\cdot;\boldsymbol{v})$ に依存しているため，$\Delta\boldsymbol{w}$ は $\hat{V}(\cdot;\boldsymbol{v})$ の値に応じた分散を持っていると考えられる．従来の自然方策勾配法では，価値関数とは独立したベースライン関数 $b(\boldsymbol{s})$ を調節することで，方策勾配の分散を小さくする [Morimura 08]．一方本研究では，式 (6) より，$\Delta\boldsymbol{w}$ の期待値が状態価値関数の値に依存せず最適値に近づくことが保証されているため，ベースライン関数を導入せず，$\Delta\boldsymbol{w}$ の分散を最小化するように状態価値関数を学習させる．Critic は以下に示す目的関数を最小化することを学習目標とする．

$$
f = \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\sum_{t=0}^{\infty}\Delta \boldsymbol{w}_t\right\|^2\right] - \left\|\sum_{t=0}^{\infty}\mathbb{E}_{\boldsymbol{\theta}}\Big[\Delta \boldsymbol{w}_t\Big]\right\|^2\right) \qquad (7)
$$

式 (7) は，アドバンテージ関数近似パラメータの更新則 $\Delta\boldsymbol{w}$ の分散共分散行列のトレースを意味している．式 (7) を Critic のもつパラメータ $\boldsymbol{v}_t$ で偏微分することで勾配を求め，最急降下法により

$$
\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \alpha_v\frac{\partial f}{\partial \boldsymbol{v}_t}
$$

として学習する．式 (6) より，式 (7) の右辺第二項は $\boldsymbol{v}_t$ に依存しないことから，

$$
\begin{aligned}
\Delta \boldsymbol{v}_t &= \frac{\partial f}{\partial \boldsymbol{v}_t} = \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{C}_t\sum_{\tau=0}^{\infty}\Delta \boldsymbol{w}_\tau\right] \qquad (8)\\
\boldsymbol{C}_t &= \Big(\frac{\partial}{\partial \boldsymbol{v}_t}\hat{V}(\boldsymbol{s}_{t+1};\boldsymbol{v}_t)\Big)\Big(\frac{\partial}{\partial \boldsymbol{\theta}_t}\ln\pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)\Big)^{\mathrm{T}}
\end{aligned}
$$

式 (8) の期待値計算は，時間平均で置き換える．また，式 (8) の計算には未来（$\tau > t$）の情報が必要であるため，このままでは計算できない．よって，以下のように変形する．

$$
\sum_{t=0}^{\infty}\boldsymbol{C}_t\sum_{\tau=0}^{\infty}\Delta \boldsymbol{w}_\tau = \sum_{t=0}^{\infty}\left(\boldsymbol{C}_t\left(\sum_{\tau=0}^{t-1}\Delta \boldsymbol{w}_\tau\right) + \left(\sum_{\tau=0}^{t}\boldsymbol{C}_\tau\right)\Delta \boldsymbol{w}_t\right) \qquad (9)
$$

式 (9) の総和計算には，t と τ の様々な組み合わせが含まれている．方策を逐次更新する場合，時間を大きく隔てた $\Delta\boldsymbol{w}_t$ と $\boldsymbol{C}_t$ は，異なる方策のもとで得られた値を保持しているため，これらの組み合わせは学習に悪影響を及ぼす可能性がある．よって，$|t-\tau|$ の値が大きい組み合わせを割り引くための定数 $\beta\in(0,1)$ を導入し，式 (9) を以下のように修正する．

$$
\sum_{t=0}^{\infty}\left(\boldsymbol{C}_t\left(\sum_{\tau=0}^{t-1}\beta^{t-\tau}\Delta \boldsymbol{w}_\tau\right) + \left(\sum_{\tau=0}^{t}\beta^{t-\tau}\boldsymbol{C}_\tau\right)\Delta \boldsymbol{w}_t\right)
$$

提案手法の学習手続きを Alg.2 に示す．

5. 計算機実験

提案手法の優位性を示すため，単純化した倒立振子の振り上げ問題に適用した．振子の振り上げに十分な大きさのトルクは許容されず，振動させて勢いをつけてから振り上げる必要があるため，長い行動系列を学習する必要がある．初期状態は振子が下向きに静止した状態とし，振子の先端がゴール高さへ振り上げられるまでを 1 エピソードした [*2]．各時刻での報酬は 0 であり，リンクがゴール高さに達した時のみ報酬 $r=1$ を与える．本研究では，Actor と Critic を Normalized Radial Basis Function Network（NRBF-Net）[Moody 89] によって構成した．Actor は，状態 $\boldsymbol{s}=(q,\dot{q})^{\mathrm{T}}$ を観測し，確率的方策 $\pi(a|\boldsymbol{s};\boldsymbol{\theta})$ に従って行動 $a=\tau$ を出力する．確率的方策 $\pi(a|\boldsymbol{s};\boldsymbol{\theta})$ は，正規分布を用いて以下のように定義する．

$$
\pi(a|\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{\theta_\sigma\sqrt{2\pi}}\exp\left(-\frac{(a-y(\boldsymbol{s},\boldsymbol{\theta}_\kappa,\theta_b))^2}{2\theta_\sigma^2}\right) \qquad (10)
$$

ただし，式 (10) における $y(\boldsymbol{s},\boldsymbol{\theta}_\kappa,\theta_b)=\boldsymbol{\theta}_\kappa^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{s})+\theta_b$ は NRBF-Net の出力であり，$\boldsymbol{\phi}(\boldsymbol{s})$ は RBF を並べたベクトル，$\boldsymbol{\theta}_\kappa$ は結合荷重，θ_b はバイアスである．各 RBF は，状態空間に格子状に配置される．すなわち，エージェントが学習する方策パラメータは $\boldsymbol{\theta}=(\boldsymbol{\theta}_\kappa,\theta_b,\theta_\sigma)$ である．同様に，Critic も NRBF-Net を用いて $\hat{V}(\boldsymbol{s},\boldsymbol{v}_\kappa,v_b)=\boldsymbol{v}_\kappa^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{s})+v_b$ と表現される．すなわち，エージェントが学習する Critic のパラメータは $\boldsymbol{v}=(\boldsymbol{v}_\kappa,v_b)$ である．

エージェントの学習則を以下の 3 条件として性能を比較した．(a) アドバンテージ関数の学習に文献 [木村 07] で提案された学習則を用いる条件 (Alg.1)．(b) アドバンテージ関数の学習に提案手法を用いる条件．(c) アドバンテージ関数と Critic の学習の両方に提案手法を用いる条件 (Alg.2)．乱数の初期値を変えながら 100 試行の学習を行った．各学習則を用いた場合について，乱数の初期値は同じ値を用いた．学習の 1 試行は 3000 エピソードとした．5000 ステップ経過してもリンク先端がゴール高さまで振り上げられない場合，そのエピソードを終了し，次のエピソードを開始した．(a-c) の学習則全てにおいて，$\boldsymbol{v}$ と $\boldsymbol{w}$ の学習にトレースを利用しており，更新のスケールに大きな違いはないと考えられる．よって，学習のハイパーパラメータは，全ての条件で同じ値を用いた．

実験結果として，初期から方策の改善が見られないものを除いた学習試行の平均の学習曲線を図 1 に示す．図 1 の横軸

*1 ただし [Thomas 14] において，$\Delta\boldsymbol{w}\propto\frac{\partial}{\partial \boldsymbol{w}}\big(Q_t^{\theta}-\hat{A}\big)^2$ に従って学習した場合でも，$\gamma\neq 1$ であればバイアスが加わり，割引収益ではなく平均報酬を最大化する自然勾配の不偏推定となることが示されている．

*2 振子先端がゴール高さまで振り上げられるとそのエピソードは終了するため，Mountain Car に近い実験設定である．

Algorithm 2 提案手法

Input:
- Parameterized policy
  $\pi(\cdot|\cdot;\boldsymbol{\theta})$
- Parameterized value function
  $\hat{V}(\cdot;\boldsymbol{v})$

Initialization:
- Policy, value function and advantage parameters
  $\boldsymbol{\theta}=\boldsymbol{\theta}_0, \boldsymbol{v}=\boldsymbol{v}_0, \boldsymbol{w}=\boldsymbol{w}_0$
- Stepsizes and discount rates
  $\alpha_\theta, \alpha_v, \alpha_w, \beta, \gamma$
- Draw initial state
  $\boldsymbol{s}_0 \sim p(\boldsymbol{s}_0)$
- Draw initial action
  $\boldsymbol{a}_0 \sim \pi(\boldsymbol{a}_0|\boldsymbol{s}_0;\boldsymbol{\theta}_0)$

for $t=0,1,2,\ldots$ **do**

Execution:
- Observe next state
  $\boldsymbol{s}_{t+1} \sim \mathcal{T}^a_{ss'}$
- Observe reward
  $r_t \sim \mathcal{R}^a_{ss'}$
- Draw action
  $\boldsymbol{a}_{t+1} \sim \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)$

Eligibility Traces:
$$\overline{\boldsymbol{e}}_{t+1} = \gamma\overline{\boldsymbol{e}}_t + \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)$$

Action Value:
$$\hat{Q}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\boldsymbol{\theta}_t,\boldsymbol{v}_t,\boldsymbol{w}_t)$$
$$= \hat{V}(\boldsymbol{s}_{t+1};\boldsymbol{v}_t) + \left(\nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_t)\right)^{\mathrm{T}} \boldsymbol{w}_t$$

TD Error:
$$\delta^Q_t = r_t + \gamma\hat{Q}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\boldsymbol{\theta}_t,\boldsymbol{v}_t,\boldsymbol{w}_t) - \hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta}_{t-1},\boldsymbol{v}_{t-1},\boldsymbol{w}_{t-1})$$

Advantage Update:
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_w \delta^Q_t \overline{\boldsymbol{e}}_t$$

Actor Update:
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_\theta \boldsymbol{w}_t$$

Critic Update:
$$\boldsymbol{C}_t = \nabla_{\boldsymbol{v}} \hat{V}(\boldsymbol{s}_{t+1};\boldsymbol{v}_t)\, \boldsymbol{e}^{\mathrm{T}}_{t+1}$$
$$\overline{\boldsymbol{C}}_t = \beta\overline{\boldsymbol{C}}_{t-1} + \boldsymbol{C}_t$$
$$\overline{\Delta\boldsymbol{w}}_t = \beta\overline{\Delta\boldsymbol{w}}_{t-1} + \Delta\boldsymbol{w}_t$$
$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t + \alpha_v \left(\beta\boldsymbol{C}_t\overline{\Delta\boldsymbol{w}}_{t-1} + \overline{\boldsymbol{C}}_t\Delta\boldsymbol{w}_t\right)$$

end for

はエピソード数，縦軸は 1 エピソード中に倒立振子の振り上げに要したステップ数を表す．縦軸の値が低いほどより良い方策を獲得していると考えられる．赤い線が条件 (a)，水色の線が条件 (b)，藍色の線が条件 (c) の場合を表す．条件 (a) で学習した場合，多くの学習試行が，初期の方策から改善した後にそれぞれ異なるエピソード数で発散した．それらの独立した学習試行間の平均をとっているため，一見方策が徐々に悪化しているように見受けられる．学習率をより小さな値に設定した条件でも実験したが，局所解に陥ったためか，多くの学習試行で初期から方策の改善が見られなかった．一方で，提案手法を用いることで，学習後期での発散を防ぎ，効率よく頑健に学習できていることがわかる．

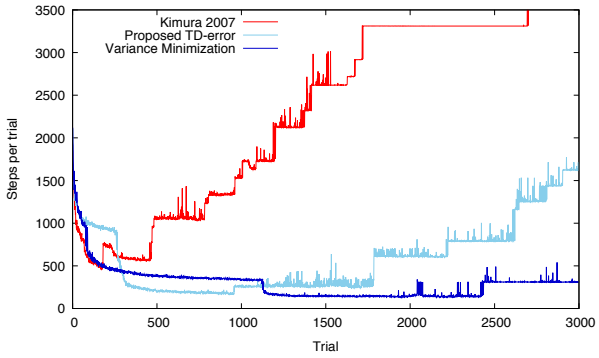

図 1: 平均の学習曲線

6. おわりに

本研究では，自然方策勾配を不偏推定する学習則と，自然方策勾配推定の分散を最小化する学習則を提案した．計算機実験により，提案手法が従来手法よりも学習後期で特に頑健であることを示した．しかし，今回提案した自然方策勾配の不偏推定・分散最小化ともに，実際は固定方策のもとでのみ成り立つ．よって今後の課題として，方策の違いによる影響を緩和可能な方策オフ学習に提案手法を拡張することが考えられる．

参考文献

[Amari 98] S. Amari, Natural Gradient Works Efficiently in Learning, *Neural Computation*, Vol. 10, No. 2, pp. 251-276 (1998).

[Kakade 02] S. Kakade, A Natural Policy Gradient, *Advances in Neural Information Processing Systems*, Vol. 14, pp. 227-242 (2002).

[Moody 89] J. Moody and C. J. Darken, Fast learning in networks of locally-tuned processing units, *Neural Computation*, Vol. 1, No. 2, pp. 281-294 (1989).

[Morimura 08] T. Morimura, E. Uchibe, and K. Doya. Natural actorcritic with baseline adjustment for variance reduction. *Artificial Life and Robotics*, Vol. 13, pp. 275279 (2008).

[Sutton 98] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*, MIT Press (1998).

[Sutton 00] R. S. Sutton, D. McAllester, S. Singh and Y. Mansour, Policy Gradient Methods for Reinforcement Learning with Function Approximation, *Advances in Neural Information Processing Systems*, Vol. 12, pp. 10571063 (2000).

[Thomas 14] P. Thomas, Bias in Natural Actor-Aritic Algorithms, *Proceedings of The 31st International Conference on Machine Learning*, pp. 441-448 (2014).

[木村 07] 木村 元，適正度の履歴を用いた自然勾配 Actor-Critic 法，第 *19* 回自律分散システムシンポジウム，pp. 6772 (2007).