2P4-4in1

# 逐次的自然方策勾配推定のための適応的学習率

An adaptive step size for incremental natural policy gradient estimation and stability analysis

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 Graduate School of Engineering, Osaka University

*2玉川大学工学部 脳科学研究所 Brain Science Institute, Tamagawa University

The step size is one of the crucial factors in incremental learning algorithms. The performance of the natural policy gradient (NPG) estimation is also affected dramaticaly. In this paper, we propose to apply online importance weight aware update to incremental NPG estimation. We also theoretically show the instability of conventional method and the stability of proposed method. We confirm the usefulness of the proposed method in the classical benchmark.

## 1. はじめに

自然方策勾配法 [Kakade 01] は強化学習手法の一つであり，自然勾配 [Amari 98] を利用することで，収益を局所的に最大化する方策パラメータを，プラトーを回避しながら探索できる．深層学習を利用する大規模な問題においても，自然方策勾配法は有用である [Duan 16]．一方，自然勾配を逐次的かつパラメータの次元に線形な計算量とメモリで推定するために，様々な学習則が提案されてきた [Morimura 05, Bhatnagar 09, Degris 12, Thomas 14] が，自然方策勾配の推定は大量の学習サンプルを必要とし，さらに学習率などのメタパラメータに非常に敏感である．これらの収束の遅さと数値的な不安定性が，自然方策勾配の逐次推定を大規模な問題に適用することを困難にしている．

本研究では，Online Importance Weight Aware Update [Karampatziakis 11] と呼ばれる適応的学習率を，自然方策勾配の推定に適用する．さらに，従来法と提案法を理論解析し，学習の安定性について議論する．古典的なベンチマーク課題において，提案法の頑健さを示す．

## 2. 自然方策勾配法

マルコフ決定過程 (Markov Decision Process, MDP) における局所最適方策を獲得する問題を扱う．MDP は $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \pi)$ の組によって特定される．$\mathcal{S}, \mathcal{A}$ をそれぞれ可能な状態と行動の集合とする．離散時刻 $t \in \mathbb{N}_{\geq 0}$ 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，行動 $\boldsymbol{a}_t \in \mathcal{A}$ を選択する．環境の状態は，状態遷移確率 $\mathcal{P}$ に従って次状態 $\boldsymbol{s}_{t+1}$ に遷移する．エージェントは，有界な報酬関数 $\mathcal{R}$ に従って環境から報酬 $r_t \in \mathbb{R}$ を得る．エージェントの意思決定は，パラメータ $\boldsymbol{\theta} \in \mathbb{R}^d$ で表現される確率的方策 $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}) \triangleq \Pr(\boldsymbol{a}_t = \boldsymbol{a}|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{\theta})$ に従う．方策 $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$ は，全ての状態 $\boldsymbol{s}$ と行動 $\boldsymbol{a}$ において，パラメータ $\boldsymbol{\theta}$ について微分可能であるとする．状態価値関数 $V^\pi(\boldsymbol{s}) \triangleq \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, \boldsymbol{s}_0 = \boldsymbol{s}\right]$ は，方策 $\pi_{\boldsymbol{\theta}}$ のもとである状態 $\boldsymbol{s}$ に期待される割引収益である．同様に，行動価値関数 $Q^\pi(\boldsymbol{s}, \boldsymbol{a}) \triangleq \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, \boldsymbol{s}_0 = \boldsymbol{s}, \boldsymbol{a}_0 = \boldsymbol{a}\right]$ は，ある状態 $\boldsymbol{s}$ において行動 $\boldsymbol{a}$ をとり，その後方策 $\pi_{\boldsymbol{\theta}}$ に従った場合に期待される割引収益である．ただし，$\gamma \in [0, 1]$ は割引率であり，$\mathop{\mathbb{E}}_{\boldsymbol{\theta}}[\cdot]$ は方策 $\pi_{\boldsymbol{\theta}}$ のもとでの期待値を表す．エージェントの目的は，平均報酬 $J(\boldsymbol{\theta}) \triangleq \lim_{T\to\infty} \frac{1}{T} \mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\sum_{t=0}^{T} r_t\right]$ を極大化する局所最適な方策パラメータ $\boldsymbol{\theta}^*$ を獲得することである．

方策のスコア関数 (適格度) を基底とする線形関数近似器 $f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a})$ を次のように定義する:

$$f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a}) \triangleq \boldsymbol{w}^\top \boldsymbol{\psi} = \boldsymbol{w}^\top \nabla_{\boldsymbol{\theta}} \ln \pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}).$$

ただし，$|\boldsymbol{w}| = |\boldsymbol{\theta}|$ である．$b(\boldsymbol{s})$ を状態依存の任意な関数として，以下の方程式が成立するとする [Sutton 99]:

$$\mathop{\mathbb{E}}_{\boldsymbol{\theta}}\left[\left(Q^\pi(\boldsymbol{s}, \boldsymbol{a}) - b(\boldsymbol{s}) - f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a})\right) \nabla_{\boldsymbol{w}} f_{\boldsymbol{w}}(\boldsymbol{s}, \boldsymbol{a})\right] = 0. \quad (1)$$

このとき，線形近似器のパラメータ $\boldsymbol{w}$ が自然方策勾配に一致する [Kakade 01]:

$$\tilde{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \simeq \boldsymbol{w}.$$

式 (1) を満たす $\boldsymbol{w}$ を推定することを目的として，様々なアルゴリズムが提案されてきた [Morimura 05, Bhatnagar 09, Degris 12, Thomas 14]．本研究では特に Incremental Natural Actor Critic (INAC) [Bhatnagar 09] に着目する．状態価値関数 $V^\pi(\boldsymbol{s})$ が所与であると仮定すると，INAC は，学習率 $\alpha$ を用いて以下の手続きで実行される:

$$\begin{aligned} \delta_t^\pi &= r_t + \gamma V^\pi(\boldsymbol{s}_{t+1}) - V^\pi(\boldsymbol{s}_t), \\ \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t + \alpha\left(\delta_t^\pi - f_{\boldsymbol{w}}\right)\boldsymbol{\psi}_t. \end{aligned} \quad (2)$$

## 3. 適応的学習率

### 3.1 更新則

Online importance weight aware update (OIWAU) [Karampatziakis 11] は，線形近似器のための適応的学習率として提案された．与えられた学習サンプルに対し，無限小の学習率でパラメータを無限回更新するという極限を考えることで，各時刻での更新が局所最適点を超えないことを保証する．損失関数として二乗損失を用いると，OIWAU によるパラメータの更新は閉形式で記述できる．OIWAU を自然方策勾配の逐次推定 (2) に適用することで次式を得る:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \frac{1 - \exp(-\alpha\|\boldsymbol{\psi}_t\|^2)}{\|\boldsymbol{\psi}_t\|^2}\left(\delta_t^\pi - f_{\boldsymbol{w}}\right)\boldsymbol{\psi}_t. \quad (3)$$

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

### 3.2 理論解析

次に，従来法と提案法を理論解析し，安定性について議論する．式 (2) と (3) は，それぞれ次のように変形できる:

$$\boldsymbol{w}_{t+1} = \left(I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)\boldsymbol{w}_t + \alpha\delta_t^\pi\boldsymbol{\psi}_t \tag{4}$$

$$\boldsymbol{w}_{t+1} = \left(I - \beta_t\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)\boldsymbol{w}_t + \beta_t\delta_t^\pi\boldsymbol{\psi}_t. \tag{5}$$

ただし，$\beta_t = \left(1-\exp(-\alpha\|\boldsymbol{\psi}_t\|^2)\right)/\|\boldsymbol{\psi}_t\|^2$ とした．簡単のため，$\delta^\pi = 0$ と仮定する．時刻 $T \in \mathbb{N}_{\geq 0}$ における $\boldsymbol{w}$ は，初期値を $\boldsymbol{w}_0$ とするとそれぞれ以下のように書ける:

$$\boldsymbol{w}_T = \prod_{t=0}^{T-1}\left(I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)\boldsymbol{w}_0, \tag{6}$$

$$\boldsymbol{w}_T = \prod_{t=0}^{T-1}\left(I - \beta_t\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)\boldsymbol{w}_0. \tag{7}$$

行列のノルムを $\|\cdot\|_2$ とする．(6) において，全ての $t$ に対して $\|I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\|_2 \leq 1$ であれば，$\|\boldsymbol{w}_T\|_2 \leq \prod_{t=0}^{T-1}\|I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\|_2\|\boldsymbol{w}_0\|_2$ は有界であることが保証される．(7) についても同様である．ここで，

$$\begin{aligned}\|I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\|_2 &= \left(I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)^\top\left(I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\right)\\ &= I + \left(\alpha^2\|\boldsymbol{\psi}_t\|^2 - 2\alpha\right)\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top \triangleq I + A.\end{aligned}$$

行列 $A$ は明らかに階数 1 の実対称行列である．よって，行列 $A$ は $d-1$ 個のゼロ固有値と，固有値 $(\alpha\|\boldsymbol{\psi}_t\|^2)^2 - 2\alpha\|\boldsymbol{\psi}_t\|^2$ をもち，行列 $I+A$ は $d-1$ 個の固有値 1 と，固有値 $\left(\alpha\|\boldsymbol{\psi}_t\|^2 - 1\right)^2$ をもつ．提案法に対しても同様にして，それぞれ

$$\begin{aligned}\|I - \alpha\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\|_2 &= \max\{1, |\alpha\|\boldsymbol{\psi}_t\|^2 - 1|\} \geq 1, \\ \|I - \beta_t\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top\|_2 &= \max\{1, \exp(-\alpha\|\boldsymbol{\psi}_t\|^2)\} = 1.\end{aligned} \tag{8}$$

すなわち，従来法による学習では，$\|\boldsymbol{\psi}_t\|^2$ の値が大きくなれば $\|\boldsymbol{w}\|_2$ は発散するが，提案法による学習では，$\|\boldsymbol{\psi}_t\|^2$ の値によらず $\|\boldsymbol{w}\|_2$ は有界である．

## 4. 数値実験

提案法と従来法を倒立振子の振り上げ・安定化問題 [Doya 00, Morimura 05] に適用した．状態は振子の関節角度を $q$，角速度を $\dot{q}$ として，$\boldsymbol{s} = (q, \dot{q})^\top$ である．行動は振子の関節に印可されるトルク $5a = \tau \in [-5, 5]$ である．報酬関数は $\mathcal{R}(\boldsymbol{s}) = \cos(q) - (\dot{q}/15\pi)^2$ と定義した．方策は正規分布 $\pi(a|\boldsymbol{s};\boldsymbol{\theta}) = 1/\sqrt{2\pi\sigma_{\boldsymbol{\theta}}^2(\boldsymbol{s})}\exp\left(-(a-\mu_{\boldsymbol{\theta}}(\boldsymbol{s}))^2/2\sigma_{\boldsymbol{\theta}}^2(\boldsymbol{s})\right)$ によって表現した．$\mu_{\boldsymbol{\theta}}(\boldsymbol{s})$ と $\sigma_{\boldsymbol{\theta}}(\boldsymbol{s})$ は三層のニューラルネットワークの出力であり，それぞれ tanh とシグモイド関数を活性としてもつ．隠れ層はシグモイド関数を活性にもつ 10 のユニットからなる．入力は $(\cos(q), \sin(q), \dot{q})^\top$ である．状態価値関数は正規化基底関数ネットワーク [Doya 00] で表現し，ガウシアンカーネルを状態空間に $10 \times 10$ の格子状に配置した．状態価値の学習には TD($\lambda$) を用い，$\lambda = 0.95$ とした．1 エピソードは 1000 ステップとした．各エピソードの初期状態は $\boldsymbol{s}_0 = (q_0, 0)^\top$ であり，$q_0$ は乱数により決定した．$\boldsymbol{\theta}$ と $\boldsymbol{w}$ の学習率をグリッドサーチにより探索し，それぞれ $\{10^{-4}, 5\cdot10^{-5}, 10^{-5}, \ldots, 10^{-7}\}, \{10^{-1}, 5\cdot10^{-2}, 10^{-2}, \ldots, 10^{-4}\}$ から選択した．すべての学習率の組に対する学習曲線の 10 試行の平均を図 1 に示す．1 つの試行であっても推定値が発散した場合，それ以降の学習曲線は排除した．従来法では多くの学習試行が発散しているが，提案法は頑健に学習していることがわかる．

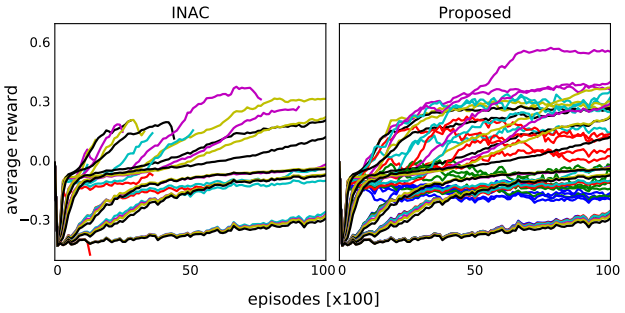


図 1: 学習率の各組に対する平均報酬の変化．10 試行の平均．

## 5. おわりに

本研究では，OIWAU を自然方策勾配の逐次推定に利用することを提案した．理論解析により，従来法の不安定性と提案法の安定性を示した．数値実験により，提案法の頑健さを示した．INAC の更新行列のノルムが，学習率と適格度の関数として解析的に得られたため (式 (8))，OIWAU に限らず，適応的学習率を設計するための新たな指針となる．適格度の履歴を利用する学習則 [Morimura 05] においても，同様の適応的学習率を提案・適用することで頑健な学習になることが期待できる．

## 参考文献

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

[Bhatnagar 09] Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M.: Natural Actor-Critic Algorithms, *Automatica*, Vol. 45, No. 11 (2009)

[Degris 12] Degris, T., Pilarski, P. M., and Sutton, R. S.: Model-free reinforcement learning with continuous action in practice, in *Proceedings of the 2012 American Control Conference* (2012)

[Doya 00] Doya, K.: Reinforcement Learning in Continuous Time and Space, *Neural Computation*, Vol. 12, (2000)

[Duan 16] Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P.: Benchmarking Deep Reinforcement Learning for Continuous Control, in *Proceedings of the 33rd International Conference on Machine Learning* (2016)

[Kakade 01] Kakade, S.: A natural policy gradient, in *Advances in Neural Information Processing Systems*, Vol. 14 (2001)

[Karampatziakis 11] Karampatziakis, N. and Langford, J.: Online Importance Weight Aware Updates, in *Uncertainty in Artificial Intelligence* (2011)

[Morimura 05] Morimura, T., Uchibe, E., and Doya, K.: Utilizing natural gradient in temporal difference reinforcement learning with eligibility traces, in *International Symposium on Information Geometry and Its Applications*, pp. 256–263 (2005)

[Sutton 99] Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, in *Advances in Neural Information Processing Systems*, Vol. 12, pp. 1057–1063 (1999)

[Thomas 14] Thomas, P. S.: Bias in Natural Actor-Critic Algorithms, in *Proceedings of The 31st International Conference on Machine Learning* (2014)