1A3-1

マルチエージェント環境下における 強化学習のステップサイズパラメータの適応

Adaptation of Stepsize Parameter in Reinforcement Learning for Multiagent Environments

野田五十樹[*1]
Itsuki Noda

[*1](独) 産業技術総合研究所 情報技術研究部門
National Institute of Advanced Industrial Science and Technology

A method to adjust a stepsize parameter in exponential moving average (EMA) based on Newton method to minimize square errors is proposed. In the most of situation of reinforcement learning, the target value of learning is generally supposed to be stable. Therefore, several learning parameters are determined according to this assumption. For example, the stepsize is decreased to be zero during learning. However, such assumption is violated under unstable environment, where target values like expected rewards may change over time. In order to adapt stepsize parameters, we proposed a framework to acquire higher-order derivatives of learning values by the stepsize parameter. Based on this framework, we propose a method to determine the best stepsize using Newton method to minimize EMA of square error of learning. The method is confirmed by mathematical theories and by results of experiments.

1. まえがき

マルチエージェント環境における学習では、他のエージェントの振る舞いや環境の変化への追従と、気まぐれな振る舞いなどによる雑音成分への耐性をどう両立するかが重要となる。一方、通常の強化学習などでは以下の式のような行動価値学習が行われることが多いが、このうちのステップサイズパラメータ α は学習を通じて 0 に漸近させることが多い。[Even-dar 03]。これは、環境の変化への追従をあきらめ、雑音成分への耐性を強化するという方策をとっていることに相当する。

$$\begin{aligned} Q_{t+1}(s_t,a_t) &= (1-\alpha)Q_t(s_t,a_t) \\ &\quad +\alpha(r_t+\gamma\max_{a'}Q_t(s_{t+1},a')) \end{aligned} \tag{1}$$

しかし、実際の応用場面では環境や他のエージェントの変化は無視できず、上記のステップサイズパラメータは適切に設定されなければならない。

この問題に対し、筆者はこれまで再帰的ステップサイズパラメータ適応法 (Recursive Adaptation of Stepsize Parameters, RASP) を提案し、これにより予測誤差を漸近的に最小化する手法を構築してきた [Noda 09]。しかしこの方法ではステップサイズが大きく変動する際に収束に時間がかかるという問題があった。

本稿ではこの問題を解決するために、Newton 法を用いて最適なステップサイズパラメータを求める方法を導出し、マルチエージェント環境における学習に適応し、その有効性を確認する。

2. 再帰的指数移動平均法

式 (1) にあげたような強化学習の更新式を一般化すると、以下の指数移動平均 $\tilde{x}_{t+1}=(1-\alpha)\tilde{x}_t+\alpha x_t$ と見做すことができる。ここで、x_t および $\tilde{x}_t$ は経験によって実際に観測された値 (報酬 r_t など) およびその推定値であり、時刻 t によって更新されていく。

これに対し、再帰的指数移動平均法では、以下のような再帰的な指数移動平均 (REMA) $\xi_t^{\langle k\rangle}$ を導入する。

$$\begin{aligned} \xi_t^{\langle 0\rangle} &= x_t \\ \xi_{t+1}^{\langle 1\rangle} &= \tilde{x}_{t+1}=(1-\alpha)\tilde{x}_t+\alpha x_t \\ \xi_{t+1}^{\langle k\rangle} &= (1-\alpha)\xi_t^{\langle k\rangle}+\alpha\xi_t^{\langle k-1\rangle} \end{aligned} \tag{2}$$

この REMA を用いると、推定値 $\tilde{x}_t$ の α による偏微分について、以下の公式が成立する。

$$\frac{\partial\xi_t^{\langle k\rangle}}{\partial\alpha} = \frac{k}{\alpha}(\xi_t^{\langle k\rangle}-\xi_t^{\langle k+1\rangle}) \tag{3}$$

$$\frac{\partial^k\tilde{x}_t}{\partial\alpha^k} = (-\alpha)^{-k}k!(\xi_t^{\langle k+1\rangle}-\xi_t^{\langle k\rangle}) \tag{4}$$

3. 2 乗誤差の指数移動平均とニュートン法

与えられた時系列 $\{x_t\}$ と、それの EMA の系列 $\{\tilde{x}_t\}$ の誤差 $\varepsilon_t=\tilde{x}_t-x_t$ と 2 乗誤差 $\mathcal{E}_t=(1/2)\varepsilon_t^2$ を考える。**再帰的指数平滑移動平均によるステップサイズ勾配降下法** (GDASS)[野田 08] ではこの $\mathcal{E}_t$ を減少させる方向 $(-\frac{\partial\mathcal{E}_t}{\partial\alpha})$ に α を漸近的に変化させる方法をとっていた。これに対し、本稿では、式 (4) で求まる高階の微係数を利用して Newton 法により最適の α を求める方法を考える。

まず、2 乗誤差の偏微分について、以下の定理が成り立つ。

定理 1

2 乗誤差、$\mathcal{E}_t$ の α による k 次偏微分は次のように求めることができる。

$$\frac{\partial^k\mathcal{E}_t}{\partial\alpha^k} = \sum_{i=0}^{k-1}\frac{(k-1)!}{(k-1-i)!i!}\frac{\partial^i\varepsilon_t}{\partial\alpha^i}\frac{\partial^{k-i}\varepsilon_t}{\partial\alpha^{k-i}} \tag{5}$$

ただし、$\frac{\partial^0\varepsilon_t}{\partial\alpha^0}=\varepsilon_t$ とする。 □

連絡先: 野田五十樹，(独) 産業技術総合研究所 情報技術研究部門，つくば市梅園 1-1-1，Tel: 029-862-6517，E-mail: i.noda@aist.go.jp

次に、各時刻における 2 乗誤差 $\mathcal{E}_t$ の指数的移動平均 $\tilde{\mathcal{E}}_{t+1} = (1-\beta)\tilde{\mathcal{E}}_t + \beta\mathcal{E}_t$ を考える。ただし、β は 2 乗誤差のためのステップサイズパラメータである。この $\tilde{\mathcal{E}}_t$ は、[佐藤 01] において提案されている期待報酬値の分散の予測値と同じである。

この $\tilde{\mathcal{E}}_t$ について、もとのステップサイズパラメータ α により微分を求めると、以下のようになる。

$$\frac{\partial\tilde{\mathcal{E}}_{t+1}}{\partial\alpha} = (1-\beta)\frac{\partial\tilde{\mathcal{E}}_t}{\partial\alpha} + \beta\frac{\partial\mathcal{E}_t}{\partial\alpha} \quad (6)$$

$$\frac{\partial^2\tilde{\mathcal{E}}_{t+1}}{\partial\alpha^2} = (1-\beta)\frac{\partial^2\tilde{\mathcal{E}}_t}{\partial\alpha^2} + \beta\frac{\partial^2\mathcal{E}_t}{\partial\alpha^2} \quad (7)$$

$$\frac{\partial^k\tilde{\mathcal{E}}_{t+1}}{\partial\alpha^k} = (1-\beta)\frac{\partial^k\tilde{\mathcal{E}}_t}{\partial\alpha^k} + \beta\frac{\partial^k\mathcal{E}_t}{\partial\alpha^k} \quad (8)$$

これらの式から、次のことが分かる。

式 (6)〜 (8) より、2 乗誤差の EMA $\tilde{\mathcal{E}}_t$ の α による高次の偏微分は、2 乗誤差 $\mathcal{E}_t$ の高次偏微分を用いて EMA により逐次的に求めることができる。また、$\mathcal{E}_t$ の高次偏微分も定理 1 により REMA $\xi_t^{\langle k\rangle}$ から求めることができる。よって、$\tilde{\mathcal{E}}_t$ の高次偏微分は REMA から逐次的に求めることができる。

この高次偏微分を用いれば、2 次の Taylor 展開

$$\tilde{\mathcal{E}}_t(\Delta\alpha) = \tilde{\mathcal{E}}_t(0) + \frac{\partial\tilde{\mathcal{E}}_t}{\partial\alpha}\Delta\alpha + \frac{1}{2}\frac{\partial^2\tilde{\mathcal{E}}_t}{\partial\alpha^2}\Delta\alpha^2$$

を用いて、以下の式のように、$\tilde{\mathcal{E}}_t$ を最小化する α を Newton 法で推定することができるようになる。

$$\Delta\alpha^* = \left(\frac{\partial\tilde{\mathcal{E}}_t}{\partial\alpha}\right)/\left(\frac{\partial^2\tilde{\mathcal{E}}_t}{\partial\alpha^2}\right) \quad (9)$$

$$\alpha^* = \alpha - \Delta\alpha^* \quad (10)$$

この方法を Rapid Recursive Adaption of Stepsize Parameter by Newton's method (RRASP-N) と呼ぶ。

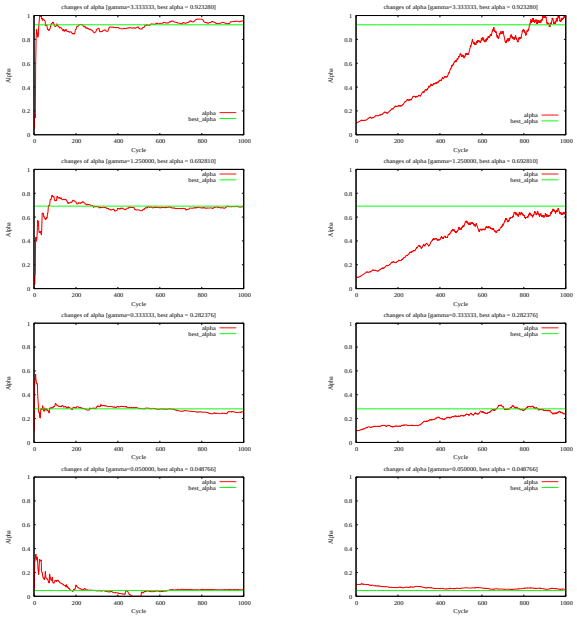
4. 実験

RRASP-N 法によりステップパラメータ α を効果的に修正できるかを確認するため、いくつかの実験を行った。

なお、以下の実験では、以下のような α の修正アルゴリズムを用いている。

S-1. $\bar{x}$ を更新する。
S-2. 式 (2) に従って $\xi^{\langle k\rangle}$ を更新する。
S-3. 式 (5) に従って $\frac{\partial^k\mathcal{E}}{\partial\alpha^k}$ を求める。
S-4. 式 (6)〜 (7) に従って $\frac{\partial\tilde{\mathcal{E}}}{\partial\alpha}$ および $\frac{\partial^2\tilde{\mathcal{E}}}{\partial\alpha^2}$ を更新する。
S-5. $\frac{\partial^2\tilde{\mathcal{E}}}{\partial\alpha^2} \leq 0$ の場合は α は変更しない。(2 次曲線が上に凸になるため、最小値が存在しない)
S-6. 式 (9) に従い $\Delta\alpha^*$ を求める。
S-7. $\Delta\alpha^*$ の絶対値が α より大きい場合は、$\Delta\alpha^* = \pm\alpha$ とする。
S-8. $\alpha \leftarrow \alpha - (1/2)\Delta\alpha^*$ とする。ただし、新しい α が $[0:1]$ の区間からはみ出る場合には、0 もしくは 1 に正規化する。
S-9. α の修正量に応じて、$\xi^{\langle k\rangle}$、$\frac{\partial\tilde{\mathcal{E}}}{\partial\alpha}$、$\frac{\partial^2\tilde{\mathcal{E}}}{\partial\alpha^2}$ を、それらの微係数を用いて修正する。

なお、S-7. において α の絶対値で $\Delta\alpha^*$ の大きさを制限するのは以下の理由による。



(a) RRASP-N 法　(b)GDASS 法

図 1: 実験 1

式 (4) を用いて $\xi^{\langle k\rangle}$ の Taylor 展開を行う場合、$\left(\frac{\Delta\alpha}{\alpha}\right)^n$ の項が現れる。この値は α が小さい領域では大きな値になり、Taylor 展開の打ち切りによる誤差が大きくなる可能性がある。よって、$\Delta\alpha$ の絶対値を α で制限することで、この誤差による影響を低減し、α の適応を安定させる。

4.1 実験 1: 最適な α の学習

まず、α が適切な値に効果的に修正されることを示すため、ランダムウォーク $v_{t+1} = v_t + \Delta v_t$ で変化する値 v_t に雑音 ϵ_t が重畳した観測値 $x_t = v_t + \epsilon_t$ を入力として学習・適応させた。ただし ϵ_t および Δv_t は平均 0、標準偏差 σ_ϵ、σ_v の乱数とする。

このような観測値 x_t を用いて、RRASP-N 法および GDASS 法により α を適応させた結果を図 1 に示す。これらのグラフは各々、異なる σ_ϵ 毎に α の適応の様子を示したものである。また各グラフのなかの水平線は、最適ステップパラメータの理論値である。

この図から分かるように、RRASP-N 法では GDASS 法に比べ、α が理論的最適値に急速に適応していることが分かる。

4.2 実験 2: 調整ゲーム

次に、マルチエージェント環境での学習の挙動を調べるため、マトリクスゲームの 1 つである調整ゲーム (coordination game) をとりあげ、その学習過程を分析する。ただし、通常の調整ゲームではなく、雑音の多い動的環境とするため、以下のような仮定をおく。

- 利得行列を複数用意し、一定時間毎に使用する行列を入れ替える。
- 各エージェントが獲得する報酬は、利得行列で与えられた値に雑音を重畳したものとする。
- 各エージェントは利得行列の入れ替え時期や各時点での相手の選んだ手を知ることはできない。
- 各エージェントは手の期待利得を学習し、ϵ-greedy により各時点での手を選ぶ。

実際に実験で用いた利得行列は以下の通りである。

matrix-0

player-X, player-Y	A	B
strategy-A	1, 1	-1, -1
strategy-B	-1, -1	1, 1

matrix-1

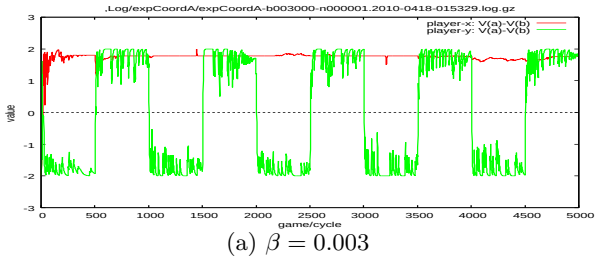
player-X, player-Y	A	B
strategy-A	-1, -1	1, 1
strategy-B	1, 1	-1, -1

この利得行列を 500 エポック毎に入れ替え、二人のエージェントにゲームを行なわせた。また重畳する雑音は、平均 0、標準偏差が 0.001 および 5.0 の正規分布に従う乱数とし、ϵ-greedy の ϵ の値は 0.1 とした。

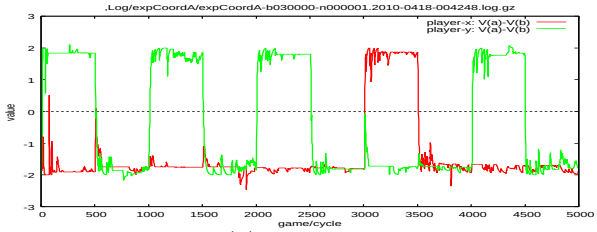
各エージェント X,Y は、各々が選べる手 (A,B) について、その期待報酬を EMA により強化学習で獲得するものとした。この際、ステップサイズパラメータを RRASP-N により変化させた。また、比較のために、ステップサイズを固定して学習を行なわせる実験も行なった。

図 2 と図 3 は雑音の影響が小さい (標準偏差が 0.001) の場合の、各エージェントの選択肢 A,B の期待報酬の差の学習による変化を示している。このグラフから分かるように、RRASP-N 法の場合は期待報酬の変化がほぼ矩形となっていることが分かる。また、雑音や ϵ-greedy の exprolation による揺らぎがほぼキャンセルされ、期待報酬の変化がほぼなくなっている部分（図 2 中でほぼ水平になっている部分）も見られる。これは RRASP-N により α がほぼ 0 となり、外乱を学習に反映させなくなっている部分である。実際、α の変化を見てみると、図 4 のように、環境の変化に追従する部分と外乱をキャンセルする部分により、α をうまく変化させていることがわかる。一方、α を固定した場合は、α が小さい場合は期待報酬の変化がなだらかであり、環境の変化 (利得行列の変化) に十分には追従できていない。一方 α が大きい場合は環境変化にはそこそこ追従するものの、期待報酬の値が外乱の影響を大きく受けてしまうことになる。

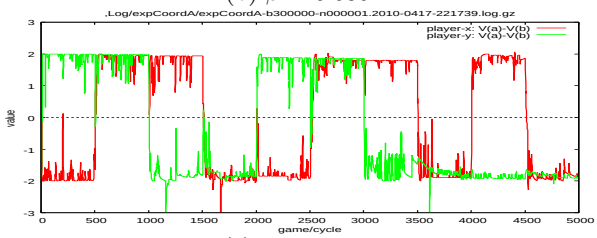
これらの性質は雑音成分が大きくなった場合でも維持される。図 5 および図 6 は標準偏差が 5.0 の雑音を利得に重畳した場合の学習結果である。さすがにきれいな矩形の変化にはならないものの、RRASP-N 法では、外乱をできるだけキャンセルしつつ、変化には敏感に追従していることがわかる。



(a) $\beta = 0.003$

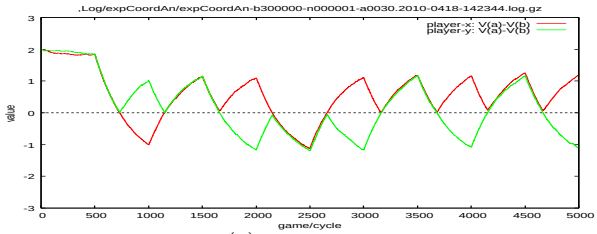


(b) $\beta = 0.030$

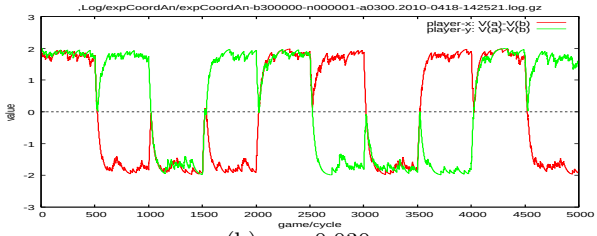


(c) $\beta = 0.300$

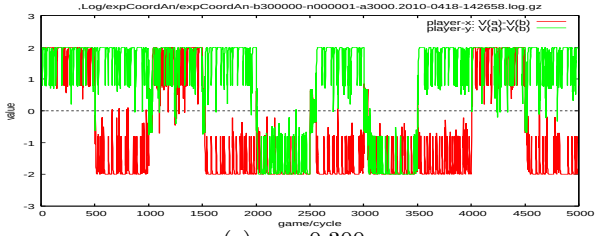
図 2: 実験 1: RRASP-N 法による調整ゲーム学習: 選択肢 A と 選択肢 B の期待報酬の差 (雑音の標準偏差=0.001)



(a) $\alpha = 0.003$



(b) $\alpha = 0.030$



(c) $\alpha = 0.300$

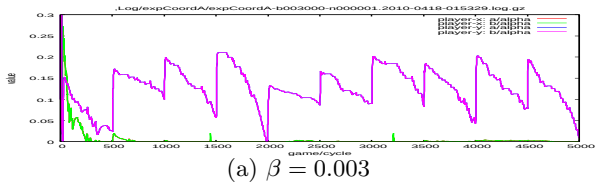
図 3: 実験 1: 固定ステップサイズによる調整ゲーム学習: 選択肢 A と 選択肢 B の期待報酬の差 (雑音の標準偏差=0.001)

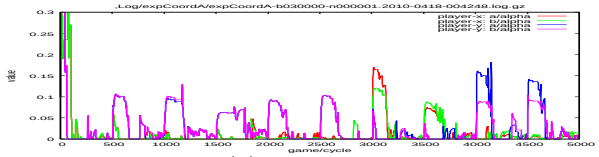
5. おわりに

本稿では、RASP の手法を拡張して、Newton 法を用いて 2 乗誤差の指数時間平均を最小化する手法、RRASP-N を提案した。この手法の特徴は、RASP により求めることができる高次の偏微分の値を用いて、2 乗誤差の指数時間平均のステップサイズパラメータによる高次偏微分を逐次的に求め、それをもとに Newton 法で最適なステップサイズの値を決めるというものである。

実験による動作確認では、提案手法が従来手法の GDASS よりも迅速に最適なステップサイズに到達できると同時に、調整ゲームのようなマルチエージェント環境においても、環境の変化に追従しつつ、相手エージェントの挙動による外乱や雑音にロバストに期待報酬を学習できることが示された。

一方、2 乗誤差の指数時間平均を求めるための新たなステップサイズパラメータ β の導入など、新たなパラメータの導入があり、それらの調整方法や、他の学習パラメータとの関係を今後検討していく必要がある。


(a) $\beta = 0.003$


(b) $\beta = 0.030$

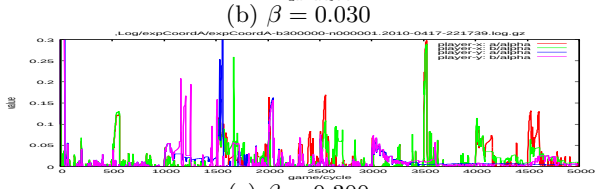
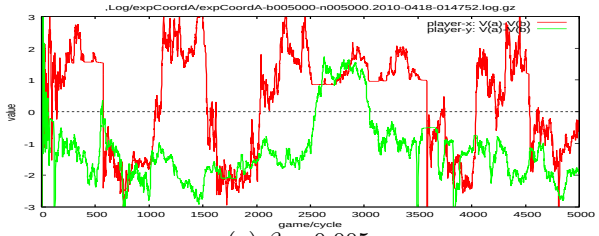

(c) $\beta = 0.300$

図 4: 実験 1: RRASP-N 法による調整ゲーム学習: ステップサイズの変化 (雑音の標準偏差=0.001)


(a) $\beta = 0.005$

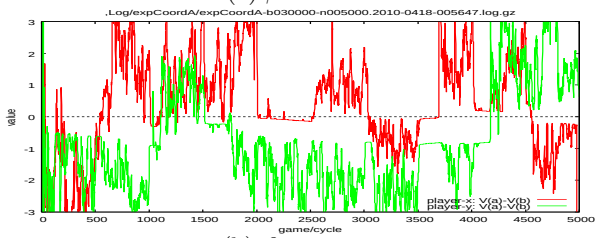
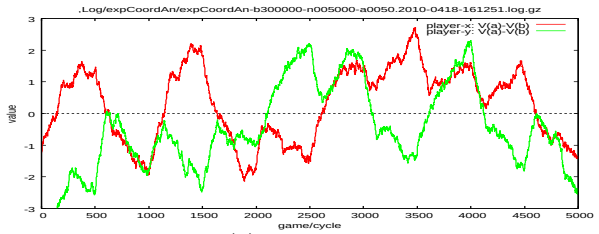

(b) $\beta = 0.030$

図 5: 実験 1: RRASP-N 法による調整ゲーム学習: 選択肢 A と 選択肢 B の期待報酬の差 (雑音の標準偏差=5.00)


(a) $\alpha = 0.005$

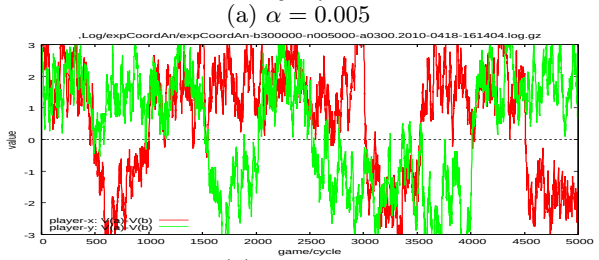

(b) $\alpha = 0.030$

図 6: 実験 1: 固定ステップサイズによる調整ゲーム学習: 選択肢 A と 選択肢 B の期待報酬の差 (雑音の標準偏差=5.00)

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

[Even-dar 03] Even-dar, E. and Mansour, Y.: Learning rates for Q-learning, *Journal of Machine Learning Research*, Vol. 5, p. 2003 (2003)

[Noda 09] Noda, I.: Recursive Adaptation of Stepsize Parameter for Non-stationary Environments, in Yang, J.-J., Yokoo, M., Takayuki Ito, Z. J., and Scerri, P. eds., *Principles of Practice in Multi-Agent Systems (Proc. of 12th International Conference, PRIMA 2009)*, pp. 525–533, Heidelberg (2009), Springer

[佐藤 01] 佐藤誠, 木村元, 小林重信：報酬の分散を推定する TD アルゴリズムと Mean Variance 強化学習法の提案, 人工知能学会論文誌, 16 巻, 3 号 F, pp. 353–362 (2001)

[野田 08] 野田 五十樹：動的環境における強化学習のステップサイズパラメータ調整法, 合同エージェントワークショップ＆シンポジウム 2008(JAWS2008) 予稿集 (2008)

A 定理 1 の証明

まず、以下の補題を示す。

補題 1
誤差 ε_t の α による偏微分は $\frac{\partial \varepsilon_t}{\partial \alpha} = \frac{\partial \bar{x}_t}{\partial \alpha}$ と書ける。また、一般の k 次偏微分は $\frac{\partial^k \varepsilon_t}{\partial \alpha^k} = \frac{\partial^k \bar{x}_t}{\partial \alpha^k}$ と書ける。 □

しかるに、式 (4) により $\frac{\partial^k \bar{x}_t}{\partial \alpha^k}$ は REMA $\xi_t^{\langle k \rangle}$ を用いて求めることができる。よって、REMA により $\frac{\partial^k \varepsilon_t}{\partial \alpha^k}$ は求めることができる。

つぎに、$\mathcal{E}_t$ の偏微分に注目する。

$j \leq k$ において、$\mathcal{E}_t$ の α による j 次偏微分が式 (5) を満たすとする。すなわち、

$$\frac{\partial^j \mathcal{E}_t}{\partial \alpha^j} = \sum_{i=0}^{j-1} \frac{(j-1)!}{(j-1-i)!i!} \frac{\partial^i \varepsilon}{\partial \alpha^i} \frac{\partial^{j-i} \varepsilon}{\partial \alpha^{j-i}} \tag{11}$$

この時、$k+1$ 次偏微分は次のようになる。

$$\frac{\partial^{k+1} \mathcal{E}_t}{\partial \alpha^{k+1}} = \sum_{i=0}^{k-1} \frac{(k-1)!}{(k-1-i)!i!} \cdot \left[\frac{\partial^i \varepsilon}{\partial \alpha^i} \frac{\partial^{j-i+1} \varepsilon}{\partial \alpha^{j-i+1}} + \frac{\partial^{i+1} \varepsilon}{\partial \alpha^{i+1}} \frac{\partial^{j-i} \varepsilon}{\partial \alpha^{j-i}} \right] \tag{12}$$

ここで、$\frac{\partial^i \varepsilon}{\partial \alpha^i} \frac{\partial^{k-i+1} \varepsilon}{\partial \alpha^{k-i+1}}$ の項で整理すると、その係数 a_i は、$i = 0$ または $i = k$ の場合、

$$a_i = 1 = \frac{((k+1)-1)!}{((k+1)-1-i)!i!} \tag{13}$$

$0 < i < k$ の場合、

$$\begin{aligned} a_i &= \frac{(k-1)!}{(k-1-(i-1))!(i-1)!} + \frac{(k-1)!}{(k-1-i))!i!} \\ &= \frac{((k+1)-1)!}{((k+1)-1-i)!i!} \end{aligned} \tag{14}$$

よって、$k+1$ 次の偏微分においても、式 (5) が成立する。

よって、任意の $k > 0$ において、式 (5) が成立する。 ■