1E-3

ニュートン法による強化学習ステップサイズパラメータの調整法

野田 五十樹 †
(独) 産業技術総合研究所 情報技術研究部門

1 まえがき

強化学習による行動価値学習でよく用いられる以下の式において、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。[1]。

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

しかし、実際の強化学習の適用場面では、期待報酬や状態遷移は時間を通じて変化することが多く、その場合、上記のステップサイズパラメータを単純に 0 に漸近させることはできなくなる。

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 [2]。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提案する。

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。

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

Adaptation of Stepsize Parameters in Reinforcement Learning by Newton Method

† Itsuki Noda, ITRI, AIST <i.noda@aist.go.jp>

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。

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

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

3 平均2乗誤差の指数移動平均

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。この時、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。

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

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

次に、各時刻における 2 乗誤差 $\mathcal{E}_t$ の指数的移動平均 $\tilde{\mathcal{E}}_t = (1-\beta)\tilde{\mathcal{E}}_t + \beta\mathcal{E}_t$ 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、[3] 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。

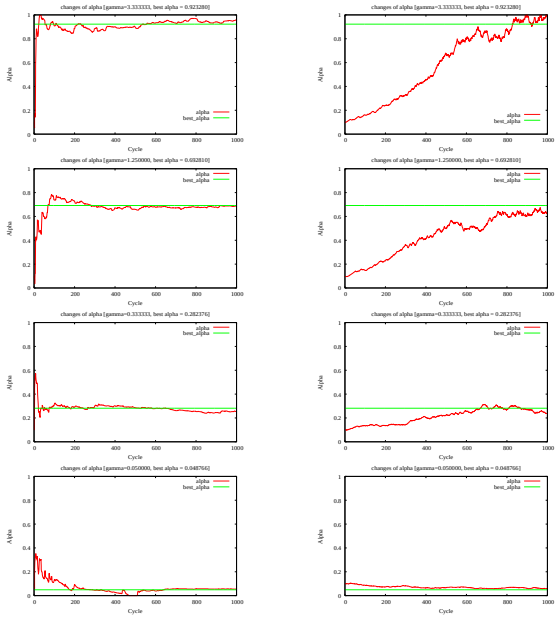
$$\frac{\partial \tilde{\mathcal{E}}_{t+1}}{\partial \alpha} = (1-\beta)\frac{\partial \tilde{\mathcal{E}}_t}{\partial \alpha} + \beta\frac{\partial \mathcal{E}_t}{\partial \alpha} \quad (6)$$

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

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

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



(a) RRASP-N 法　　(b)GDASS 法

図 1: 実験 1b

ことができることがわかる。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

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

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。

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

$$\alpha^* \quad = \quad \alpha - \Delta\alpha^* \tag{10}$$

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の実験を行った。この実験では、α が適切な値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 法により α は理論的最適値に急速に適応していることが分かる。

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、真の信号の周期性などの性質を安定して抽出できる機能を持っていることが示された。

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

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

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

[3] 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 (2001), 353–362.