3D-1

変化する環境に対する ステップサイズパラメータオンライン調整法

野田 五十樹 †
(独) 産業技術総合研究所 情報技術研究部門

1 指数平滑移動平均

Q 学習や TD 学習などの強化学習では、未知の環境における状態や行動の価値を、実際に行動して得られた報酬を元に推定し、そこで得られた推定価値をもとに行動を決定するかたちで学習が進む。この価値の推定で用いられる式を一般化すると、下記のような**指数平滑平均** (Exponential Moving Average, EMA) の式で表される。

$$\tilde{x}_{t+1} = (1-\alpha)\tilde{x}_t + \alpha x_t \quad (1)$$

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。また、α は**ステップパラメータ** である。この α は、直近の観測値 x_t をいかに重視するか、あるいは、どの程度長い時間の移動平均として推定値 $\tilde{x}_t$ を求めるかを示している。

ここで、観測値の系列 $\{x_t\}$ が下記のように、求めるべき真の値の系列 $\{s_t\}$ に雑音が重畳したものと考える。$x_t = s_t + \epsilon_t$ ただし ϵ_t は平均 0、標準偏差 σ_ϵ の乱数とする。

さらに、真の値の系列 $\{s_t\}$ は $s_{t+1} = s_t + v_t$ のようにランダムウォークで変化する値とする。ただし、v_t は平均 0、標準偏差 σ_v の乱数とする。

観察された時系列 $\{x_t\}$ が与えられている場合に、それに最も適した α を適応的に決定する方法を考える。

まず、以下のように (1) 式 を再帰的に適用した再帰的指数平滑移動平均 (Recursive Exponential Moving Average, REMA) $\xi_t^{\langle k\rangle}$ を導入する。

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

Recursive Adaptation of Stepsize Parameters in Reinforcement Learning for Dynamic Environments

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

この時、以下の補題が成立する。

補題 1
REMA $\xi_t^{\langle k\rangle}$ の α による *1* 階偏微分は次式で与えられる。

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

定理 1
EMA $\tilde{x}_t$ *(*$k=1$ *の REMA)* の k 階偏微分は下記で与えられる。

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

2 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

定理 1 により α による $\tilde{x}_t$ の導関数が求まるので、誤差 δ_t の二乗誤差を逐次的に極小化する勾配降下法を導くことができる。ただし、定理 1 では高階の導関数も求めることができるので、より精度の高い降下法を求めることができる。よって、この高階導関数を用いた勾配降下法を**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** (Recursive Adaptation of Stepsize Parameters, RASP) と呼ぶ。

今、仮に $\alpha \to \alpha + \Delta\alpha$ と変化させた場合の $\tilde{x}_t$ の変化分を $\Delta\tilde{x}_t$ とすると、

$$\Delta\tilde{x}_t = \sum_{k=1}^{\infty}(-1)^k \left(\frac{\Delta\alpha}{\alpha}\right)^k (\xi_t^{\langle k+1\rangle} - \xi_t^{\langle k\rangle}) \quad (5)$$

さらにまた、一般の $\xi_t^{\langle k\rangle}$ の変化についても、補題 1 の結果を用いて 1 次の Tailor 展開を行うと以下のようになる。[3]

$$\Delta\xi_t^{\langle k\rangle} \simeq k\left(\frac{\Delta\alpha}{\alpha}\right)(\xi_t^{\langle k\rangle} - \xi_t^{\langle k+1\rangle}) \quad (6)$$

[3]付録に示した高階導関数を用いれば、高次の Tailor 展開を求めることもできる。

具体的には、以下のような手順で学習を行う。

初期化: $\forall k \in \{0 \ldots k_{\max}-1\} : \xi^{\langle k \rangle} \leftarrow x_0$
while forever **do**
  観測データを x とする。
  for $k = k_{\max}-1$ to 1 **do**
    $\xi^{\langle k \rangle} \leftarrow (1-\alpha)\xi^{\langle k \rangle} + \alpha\xi^{\langle k-1 \rangle}$
  end for
  $\xi^{\langle 0 \rangle} \leftarrow x$
  $\delta \leftarrow \xi^{\langle 1 \rangle} - x$
  $\frac{\partial \xi^{\langle 1 \rangle}}{\partial \alpha}$ を (4) 式 により求める。
  δ および $\frac{\partial \xi^{\langle 1 \rangle}}{\partial \alpha}$ から、α の変化分 $\Delta\alpha$ を決定。
  for $k = 1$ to $k_{\max}-1$ **do**
    (5) 式 および (6) 式 により $\Delta\xi^{\langle k \rangle}$ を求める。
    $\xi^{\langle k \rangle} \leftarrow \xi^{\langle k \rangle} + \Delta\xi^{\langle k \rangle}$
  end for
  $\alpha \leftarrow \alpha + \Delta\alpha$
end while

この中で、$\Delta\alpha$ を決定する方法にはいくつか考えられる。通常の勾配降下法と同じく、基本的には $\delta\frac{\partial \xi^{\langle 1 \rangle}}{\partial \alpha}$ が正ならば $\Delta\alpha < 0$ 、負であれば $\Delta\alpha > 0$ とすればよい。

3 実験

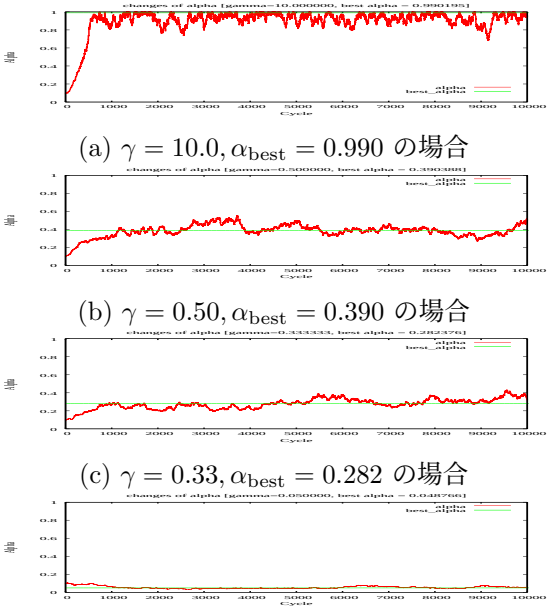
まず、上で述べた α の更新方法により、最適な値が得られるかどうかを、実験により確認する。

図 1 は、異なる γ を持つ観測データ x_t を用いて α を適応させた場合の実験結果である。これらのグラフは、学習回数 (横軸) が進むのに従って、α がどのように変化したかを示している。また、グラフ中に描かれている水平の直線は、ランダムウォークとノイズの分散から求めた最適の α_{best} である。この α の変化を見ると判るように、学習を経るに従い、γ に応じた最適なステップサイズを獲得できていることが判る。ただし、観測データに重畳されるノイズの影響により、最適値に収束していくのではなく、最適値の周辺で変動を続けるようになる。

以上のように、与えられた観測データに含まれる真の値のランダムウォークの大きさと雑音成分の大きさに応じたステップサイズを、RASP により獲得できていることがわかる。

4 まとめ

本稿では、徐々に変化する環境の中で、エージェントが学習を継続しながら環境の変化に適応していくという前提のもと、学習パラメータであるステップサイズ α と環境から得られる観測データの関係を明らかにした。その結果を元に、α を環境の変化に追随させる方法として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 (RASP) を提案した。この方法の特徴は、通常の勾配降下法と異なり、高階の導関数をシステマティックに求めることができる点である。このため、勾配法による α の変更量を比較的大きく取ることができ、さらに、修正の際に利用する値も同時に修正する方法を提供できる。[1]



(a) $\gamma = 10.0, \alpha_{\text{best}} = 0.990$ の場合

(b) $\gamma = 0.50, \alpha_{\text{best}} = 0.390$ の場合

(c) $\gamma = 0.33, \alpha_{\text{best}} = 0.282$ の場合

(d) $\gamma = 0.05, \alpha_{\text{best}} = 0.048$ の場合

図 1: 実験 1: 標準偏差比 γ の違いによる α の学習過程の変化

参考文献

[1] 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 (10 月 2008).