強化学習エージェントと報酬頻度に関する考察

Relation Between Stepsize Parameter and Stochastic Reward on Reinforcement Learning

野田五十樹[*1*2]
Itsuki Noda

[*1](独) 産業技術総合研究所 情報技術研究部門
Information Technology Research Institute, AIST

[*2]北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

This article reports an investigation about relations between stepsize parameters and stochastic rewards under TD-learning of reinforcement learning agents.

1. はじめに

TD 学習など強化学習を行うエージェントでは、ステップサイズパラメータなどの学習パラメータを、与えられた環境に適切に設定する必要がある。特にマルチエージェント環境など、環境条件が徐々に変化する非定常的な環境では、定常状態を仮定した学習手法とは異なるアプローチが必要となってくる [Gorge 06, 野田 08, Sato 01, Benveniste 90, Douglas 95, Bowling 02]。特にステップサイズパラメータについては、指数移動平均というかたちで、エージェントが得る報酬に含まれる雑音成分を除去する機能を制御するが、平均という統計処理と変化への追従という相反する問題を解くために、与えられている環境に適応してその値を定める必要がある [Gorge 06, 野田 08, Benveniste 90, Douglas 95]。

ただし、上記のステップサイズパラメータに対する分析や手法は単純な対象の指数移動平均を前提としており、学習エージェントが仮定するような状態遷移を考慮していなかった。本稿ではこの状態遷移でモデル化された環境における TD 学習について、ステップサイズパラメータと獲得される期待報酬の関係を調べていく。

2. 状態遷移と期待報酬の学習

図 1 に示すような、3 つの状態 $\{A, B, C\}$ の間の状態遷移があるとする。つまり、状態 B からは A あるいは C へ、確率 p および $1-p$ で遷移し、A および C からは、確率 1 で B へ遷移する。報酬は、B から A への遷移の時のみ、ある頻度で r が与えられるものとする。[*1]

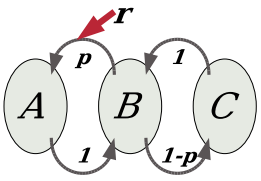


図 1: 単純な 3 状態の遷移

連絡先: 野田五十樹、(独) 産業技術総合研究所、つくば市梅園 1-1-1、029-862-6517、i.noda@aist.go.jp

*1 A への遷移確率 p が十分小さければ、「ある頻度」と等価となるとみなすことにする。

この時、割引率 γ を用いて、各状態の期待報酬 $v(A), v(B), v(C)$ を TD 学習する場合を考える。

$$\begin{aligned} v_{t+1}(A) &= (1-\alpha)v_t(A) + \alpha\gamma v_t(B) \\ v_{t+1}(B) &= (1-\alpha)v_t(B) + \alpha(pr + \gamma(pv_t(A) + (1-p)v_t(C)) \\ v_{t+1}(C) &= (1-\alpha)v_t(C) + \alpha\gamma v_t(B) \end{aligned}$$

ここで、モデルの対称性から $v(A) = v(C)$ なので、次のように簡約できる。

$$\begin{aligned} v_{t+1}(A) &= (1-\alpha)v_t(A) + \alpha\gamma v_t(B) \\ v_{t+1}(B) &= (1-\alpha)v_t(B) + \alpha(pr + \gamma v_t(A)) \end{aligned}$$

これを行列形式に書き直しておく。

$$\begin{aligned} \boldsymbol{v}_{t+1} &= \begin{bmatrix} v_{t+1}(A) \\ v_{t+1}(B) \end{bmatrix} \\ &= \begin{bmatrix} 1-\alpha & \alpha\gamma \\ \alpha\gamma & 1-\alpha \end{bmatrix} \begin{bmatrix} v_t(A) \\ v_t(B) \end{bmatrix} + \delta_{t,t'} \begin{bmatrix} 0 \\ r \end{bmatrix} \\ &= \boldsymbol{A}\boldsymbol{v}^{t+1} + \delta_{t,t'}\boldsymbol{r} \end{aligned} \quad (1)$$

(1) 式 を直行分解すれば次のように変形できる。

$$\begin{aligned} \boldsymbol{v}_{t+1} &= \begin{bmatrix} 1/2 & -1 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1-\alpha+\alpha\gamma & 0 \\ 0 & 1-\alpha-\alpha\gamma \end{bmatrix}^{t+1} \\ &\quad \cdot \begin{bmatrix} \alpha r \\ (1/2)\alpha r \end{bmatrix} \end{aligned} \quad (2)$$

2.1 単調に減衰する報酬に対する指数平滑移動平均

周期 T で立ち上がり、減衰する入力 x を考える。

$$x_t = \begin{cases} x_0\lambda^t & ; \quad 0 \leqslant t < T \\ x_{t \bmod T} & ; \quad \text{othewise} \end{cases} \quad (3)$$

これの指数平滑移動平均 ξ_t を考える。

$$\begin{aligned} \xi_t &= (1-\alpha)\xi_{t-1} + \alpha x_t \\ &= \beta\xi_{t-1} + \alpha x_t \end{aligned}$$

ただし、α はステップパラメータである。

これを、 $0 \leqslant t < T$ の範囲で展開すると以下のようになる。

$$\begin{aligned}
\xi_t &= \beta\xi_{t-1} + \alpha x_{t-1} \\
&= \beta^2\xi_{t-2} + \beta\alpha x_{t-2} + \alpha x_{t-1} \\
&= \beta^t\xi_0 + \alpha\sum_{\tau=0}^{t-1}\beta^{t-1-\tau}x_\tau \\
&= \beta^t\xi_0 + \alpha\beta^{t-1}x_0\sum_{\tau=0}^{t-1}\left(\frac{\lambda}{\beta}\right)^\tau \\
&= \beta^t\xi_0 + \alpha x_0\left(\frac{\beta^t-\lambda^t}{\beta-\lambda}\right)
\end{aligned}$$

x の周期性から、 $\xi_T = \xi_0$ が成り立つとする。

$$\begin{aligned}
\xi_T = \beta^T\xi_0 + \alpha x_0\left(\frac{\beta^t-\lambda^t}{\beta-\lambda}\right) &= \xi_0 \\
\alpha x_0\left(\frac{\beta^t-\lambda^t}{\beta-\lambda}\right) &= (1-\beta)\xi_0
\end{aligned}$$

$$\begin{aligned}
\xi_0 &= \frac{\alpha}{1-\beta^T}\frac{\beta^t-\lambda^t}{\beta-\lambda}x_0 \\
&= \left(\sum_{\tau=0}^{T-1}\beta^\tau\right)^{-1}\left(\sum_{\tau=0}^{T-1}\lambda^\tau\beta^{T-1-\tau}\right)x_0
\end{aligned}$$

次に、予測誤差 $\epsilon_t = x_t - \xi_t$ を求める。

$$\begin{aligned}
\epsilon_t &= x_t - \xi_t \\
&= x_0\left[\lambda^t - \beta^t\frac{1-\beta}{1-\beta^T}\left(\frac{\beta^T-\lambda^T}{\beta-\lambda}\right) - (1-\beta)\frac{\beta^t-\lambda^t}{\beta-\lambda}\right] \\
&= \frac{x_0}{(1-\beta^T)(\beta-\lambda)} \\
&\quad\cdot\Big[(1-\beta^T)(\beta-\lambda)\lambda^t - \beta^t(1-\beta)(\beta^T-\lambda^T) \\
&\quad -(1-\beta)(1-\beta^T)(\beta^t-\lambda^t)\Big]
\end{aligned}$$

上式の [ ] 内を整理すると、

$$\begin{aligned}
[\cdot] &= \beta\lambda^t - \lambda^{t+1} - \beta^{T+1}\lambda^t + \beta^T\lambda^{t+1} \\
&\quad -\beta^{T+t} + \beta^t\lambda^T + \beta^{t+T+1} - \beta^{t+1}\lambda^T \\
&\quad -\beta^t + \beta^{t+1} + \beta^{t+T} - \beta^{t+T+1} \\
&\quad +\lambda^t - \beta\lambda^t - \beta^T\lambda^t + \beta^{T+1}\lambda^t \\
&= \lambda^t(1-\lambda) - \beta^T\lambda^t(1-\lambda) - \beta^t(1-\lambda^T) + \beta^{t+1}(1-\lambda^T) \\
&= \lambda^t(1-\lambda)(1-\beta^T) - \beta^t(1-\beta)(1-\lambda^T)
\end{aligned}$$

よって、誤差 ϵ_t は以下のようになる。

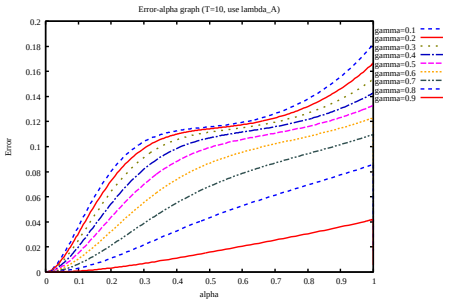
$$\begin{aligned}
\epsilon_t &= \frac{x_0}{(1-\beta^T)(\beta-\lambda)}\left[\lambda^t(1-\lambda)(1-\beta^T) - \beta^t(1-\beta)(1-\lambda^T)\right] \\
&= \frac{x_0}{(1-\beta^T)(\beta-\lambda)}A_t
\end{aligned}$$

これを使って平均二乗誤差 E を求める。

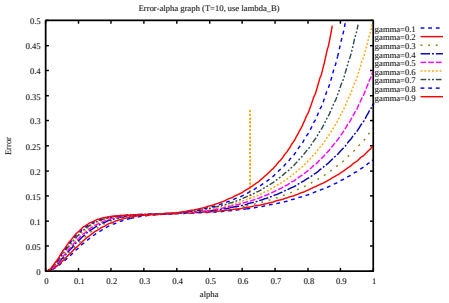
$$E = \frac{1}{T}\sum\tau = 0^{T-1}\epsilon_t^2 \qquad (4)$$

これの α による微分を求めて見る。

$$\begin{aligned}
\frac{\partial E}{\partial\alpha} &= \frac{1}{T}\sum_{\tau=0}^{T-1}\frac{\partial E}{\partial\epsilon_\tau}\frac{\partial\epsilon_\tau}{\partial\beta}\frac{\partial\beta}{\partial\alpha} \\
&= \frac{1}{T}\sum_{\tau=0}^{T-1}2\epsilon_\tau\frac{\partial\epsilon_\tau}{\partial\beta}(-1)
\end{aligned}$$



(a) $\lambda = 1 - \alpha + \alpha\gamma$



(b) $\lambda = 1 - \alpha - \alpha\gamma$

図 2: ステップパラメータ α による平均二乗誤差 E の変化 ($T = 10$)

3. 数値計算による分析

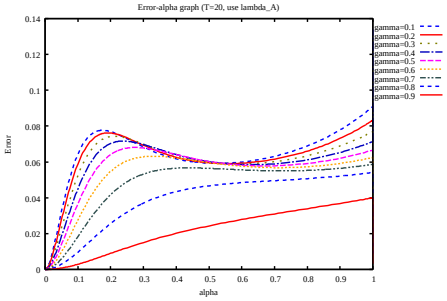
今仮に、(4) 式 において、λ を (2) 式 の解に置き換えてみる。すなわち、

$$\lambda = 1 - \alpha \pm \alpha\gamma \qquad (5)$$

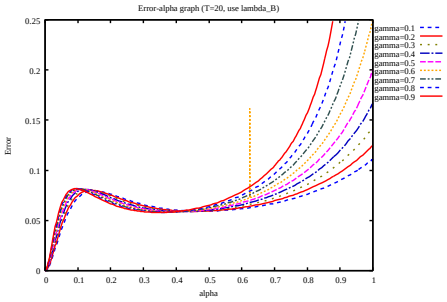
とする。この時、様々な周期 T および割引率 γ について、α が変化したときに平均二乗誤差 E がどう変化するかをプロットしたものが図 2 〜図 10 である。

これらの数値計算により、以下のことが言える。

- いずれのケースに置いても、$\alpha = 0$ で二乗平均誤差 E が最低になる。つまり、たまに得られる報酬が状態遷移を伝搬することで、期待報酬は減衰していくことになるが、その学習を最適化するためには、$\alpha = 0$ としなければならない。しかし一方、$\alpha = 0$ では学習が進まないので、わずかでも正の数である必要がある。
- また、$\alpha = 0$ において $E = 0$ となっているが、本当は直接もらえた報酬分の誤差があるはずである。しかし、いずれの場合も直接報酬をもらった場合には同程度の大きな誤差が生じるので、これを除外してもそれほど大きな違いは生じない。
- T および γ の組合わせによれば、$E - \alpha$ の描く曲線に local minimum が生じる。これは、山登り的に逐次的に計算すると、その local minimum から逃れられない可能性があることを示している。local minimum は、T が大きくなるほど広く深くなり、その局所最低点も大きい値になる。(図 6)
- λ と α を独立と見なすと、グラフの形はかなり変る。この場合、$\alpha = 0$ が必ずしも最低点にならない。(図 7 〜図 10)

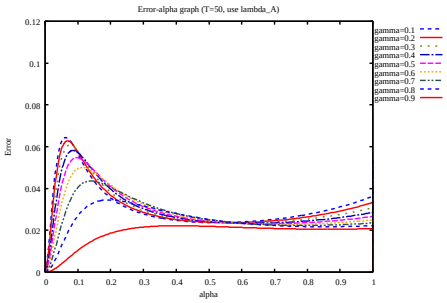


(a) $\lambda = 1 - \alpha + \alpha\gamma$

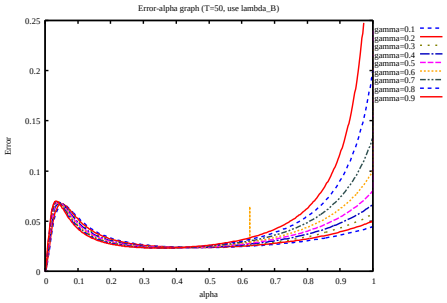


(b) $\lambda = 1 - \alpha - \alpha\gamma$

図 3: ステップパラメータ α による平均二乗誤差 E の変化 ($T = 20$)

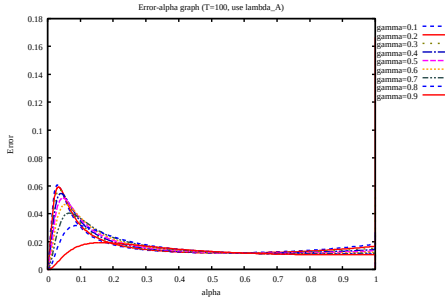


(a) $\lambda = 1 - \alpha + \alpha\gamma$

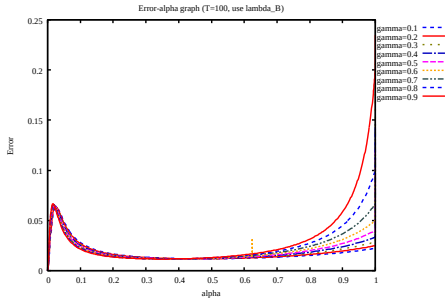


(b) $\lambda = 1 - \alpha - \alpha\gamma$

図 4: ステップパラメータ α による平均二乗誤差 E の変化 ($T = 50$)

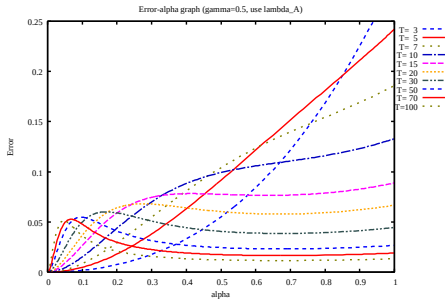


(a) $\lambda = 1 - \alpha + \alpha\gamma$

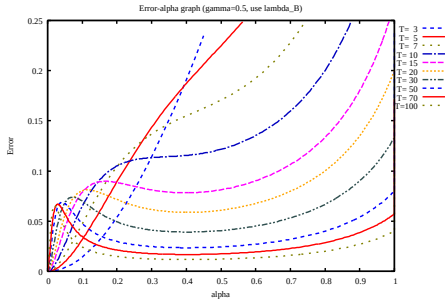


(b) $\lambda = 1 - \alpha - \alpha\gamma$

図 5: ステップパラメータ α による平均二乗誤差 E の変化 ($T = 100$)



(a) $\lambda = 1 - \alpha + \alpha\gamma$



(b) $\lambda = 1 - \alpha - \alpha\gamma$

図 6: ステップパラメータ α による平均二乗誤差 E の変化 ($\gamma = 0.50$)

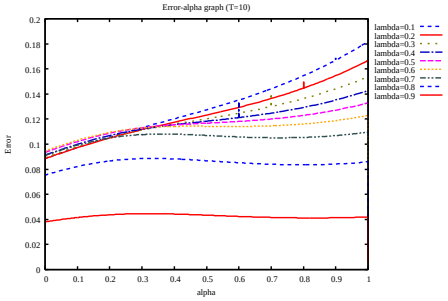


図 7: λ を固定した場合の、ステップパラメータ α による平均二乗誤差 E の変化 ($T = 10$)

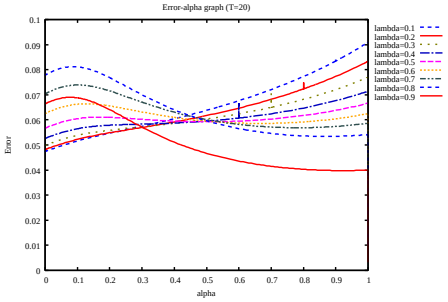


図 8: λ を固定した場合の、ステップパラメータ α による平均二乗誤差 E の変化 ($T = 20$)

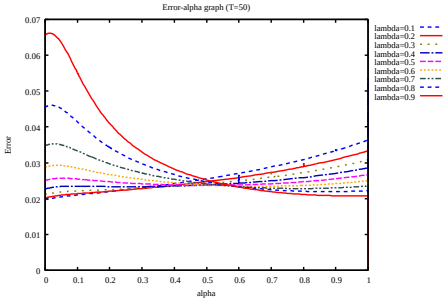


図 9: λ を固定した場合の、ステップパラメータ α による平均二乗誤差 E の変化 ($T = 50$)

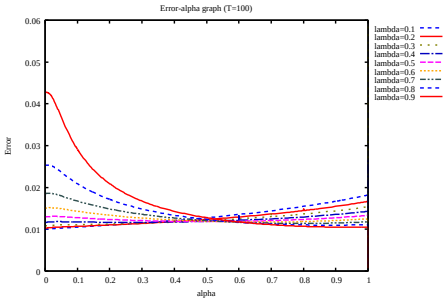


図 10: λ を固定した場合の、ステップパラメータ α による平均二乗誤差 E の変化 ($T = 100$)

4. おわりに

本稿では、状態遷移とその上での割引期待報酬を TD 学習で獲得する学習エージェントについて、ステップサイズパラメータと学習される期待報酬の関係について数値計算をもとに解析を行った。本稿で示した TD 学習の性質は多くの近似や仮定を前提としており、実際の問題に直接適用できるとは限らない。ただ、期待報酬についてステップサイズに局所最適解が存在しうることは重要な性質であり、ステップサイズパラメータを適応的に求める際には注意を要することがわかる。今後は、ここで明らかになった性質を考慮したステップサイズの適応手法を確立していく必要がある。

参考文献

[Benveniste 90] Benveniste, A., Metivier, M., and Priouret, P.: *Adaptive Algorithms and Stochastic Approximations*, Springer (1990)

[Bowling 02] Bowling, M. and Veloso., M.: Multiagent learning using a variable learning rate, *Artificial Intelligence*, Vol. 136, pp. 215–250 (2002)

[Douglas 95] Douglas, S. C. and Mathews, V. J.: Stochastic gradient adaptive step size algorithms for adaptive filtering, in *Proc. International Conference on Digital Signal Processing*, pp. 142–147 (1995)

[Gorge 06] Gorge, A. P. and Powell, W. B.: Adaptive step-sizes for recursive estimation with applications in approximate dynamic programming, *Machine learning*, Vol. 65, No. 1, pp. 167–198 (2006)

[Sato 01] Sato, M., Kimura, H., and Kobayashi, S.: TD Algorithm for the Variance of Return and Mean-Variance Reinforcement Learning (in Japanese), *Transactions of the Japanese Society for Artificial Intelligence*, Vol. Vol. 16, No. No. 3F, 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)