4H1-2

マルチエージェント強化学習の最適 Exploration 率と各種パラメータの関連の実験的考察

Experimental Investigation of Relation between Exploration Ratio and Environmental Parameters in Multiagent Reinforcement Learning

野田 五十樹[*1]
Itsuki NODA

[*1](独) 産業技術総合研究所 サービス工学研究センター, JST, CREST
Center for Service Research, AIST and CREST, JST

Experimental investigation of relations among optimal learning and environmental parameters are reported. In multiagent learning (MAL) for non-stationary environment, several learning parameters affect learning performance in combinatorial ways. In order to figure out effects of each parameters, I carried out several MAL experiments to find mainly optimal exploration ratio. Based on the results, I try to illustrate relations among learning parameters.

1. まえがき

非定常環境マルチエージェント学習において重要となる Exploration 率について、エージェントの総数がどのように関係するかを分析する。エージェントの学習で必須の Exploration が相互の学習に影響しあうマルチエージェント環境に於いては、Exploration を行う割合を適切に設定しておく必要がある。Exploration の割合についてはこれまで、静的な環境における学習での分析が主に行われてきた [Zhang 06, Martinez-Cantin 09, Rejeb 05, Tokic 10, Reddy 11]。しかし、動的な環境でのマルチエージェント学習という設定での分析はあまり行われてきていない。筆者はこれまで、Exploration 率と学習の精度の間のトレードオフの関係を分析する形式的な枠組みを提案してきた。本稿ではその枠組みを基に、最適 exploratin 率が他のパラメータからどのような影響をうけるかについて、式展開と実験結果を元に議論・検討していく。

2. 形式化と定理

本稿では、マルチエージェント環境として population game (**PG**) を取り上げる。**PG** は $\langle \mathbf{A}, \mathbf{C}, \boldsymbol{r} \rangle$ で定義される。ここで、$\mathbf{A} = \{a_1, a_2, \cdots, a_N\}$ はエージェント集合、$\mathbf{C} = \{c_1, c_2, \cdots, c_K\}$ はエージェントの行動集合、$\boldsymbol{r} = \{r_a | a \in \mathbf{A}\}$ は各エージェントに対する報酬関数である。この報酬関数 $r_a(c; \boldsymbol{d}_{\bar{a}})$ は、おなじ行動を選んだエージェントの数に応じて決定される点が、**PG** の最大の特徴付けとなる。行動ごとにそれを選んだエージェント数を**分布**と呼ぶ。また、あるエージェント a 以外のエージェントについての分布を $\left[\ d_{\bar{a},c} | c \in \mathbf{C} \ \right]$ として表す。また、報酬関数 r_a の返す値はは確率的に決定されるとする。

この **PG** に対し、あるエージェント a がある分布の条件下 $\boldsymbol{d}_{\bar{a}}$ で各行動 c を選択した際に他の行動に比べ最大の報酬が得られる確率を**優勢確率 (AP)** と呼ぶ。

$$\rho_a(c; \boldsymbol{d}_{\bar{a}}) \quad = \quad \mathcal{P}\left(\forall c' \in \mathbf{C} : r_a(c; \boldsymbol{d}_{\bar{a}}) \geq r_a(c'; \boldsymbol{d}_{\bar{a}})\right)$$

ここで、各エージェントは優勢確率が最大となる行動を選ぶことを理想状態と考え、また、エージェントの学習は、その理想状態に近づくために真の優勢確率を求めることであるとみなす。この学習を経験により進める方法として ϵ-greedy による強化学習を用いると仮定する。すなわち、学習を行うエージェントは、優勢確率最大の行動を選びつつ (Exploitation)、ある確率 ϵ でそれ以外の行動を選ぶ (Exploration) ことで、各選択肢の報酬の値と優勢確率を修正していくものとする。

この形式で学習を進める多数のエージェントからなる集団において、動的な環境での学習精度について、以下の定理が知られている [野田 13, Noda 13]。

定理 2..1
各エージェントの平均学習誤差の下限は以下の式で与えられる。

$$Error \quad \geq \quad T\sigma^2 + \frac{K\tilde{g}_a}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon), \qquad (1)$$

ただし、$\tilde{g}_a$ は以下のような AP のフィッシャー情報行列の逆行列の跡 *(*$tr(\boldsymbol{G}_a)$*)* である。

$$\boldsymbol{G}_a^{-1} \quad = \quad \left[E\left[\frac{\partial \log \rho_a}{\partial d_{\bar{a},i}} \cdot \frac{\partial \log \rho_a}{\partial d_{\bar{a},j}} \right]_{ij} \right]$$

また、T は学習の時間間隔、σ は環境の変化率 *(*ランダムウォークモデルの変動サイズ*)*、K は選択行動 *(*共有資源*)* の数である。 □

2.1 最適 Exploration 率とエージェント数

上記の定理に基づき [野田 13, 野田 14] では、ある一定の条件下ではエージェントの総数 N が変化しても、最適な Exploration 率 ϵ は変化しないことを、解析的方法および実験的方法により示している。ここで、(1) 式に示された学習誤差の下限 ($\mathcal{L}(\epsilon)$ と表す) が最小値となる ϵ を最適であるとする。この時、$\mathcal{L}(\epsilon)$ を ϵ で微分をゼロにする式は、以下のような形に展開される。

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \epsilon} \quad &= \quad \frac{1}{T}\frac{\partial}{\partial \epsilon}\left(\frac{Q}{\epsilon}\right) + \frac{\partial}{\partial \epsilon}\left(\epsilon\left(2 - \frac{K+1}{K}\epsilon\right)\right) \\ &= \quad 0 \end{aligned} \qquad (2)$$

この内、Q は各行動選択 (資源) の報酬を決める容量パラメータと ϵ のみに依存する値である。この式の中にエージェント数 N が含まれていないことから、最適 ϵ は N に依存せず決まることを示すことができる。

連絡先: 野田五十樹, 産業技術総合研究所, つくば市梅園 1-1-1, 029-861-3298，029-862-6548，i.noda@aist.go.jp

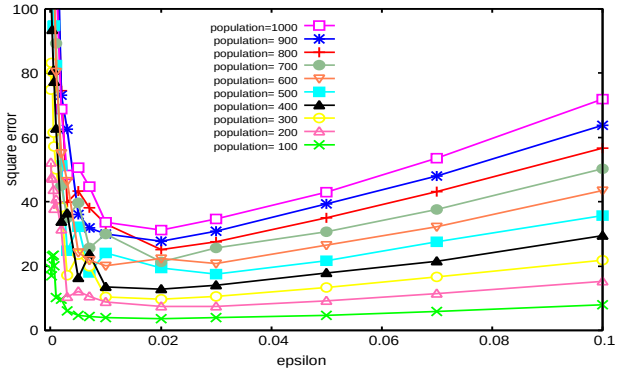

図 1: 報酬が $r_c(d_c) = B - (d_c/\gamma_c)$ の時の学習誤差の変化

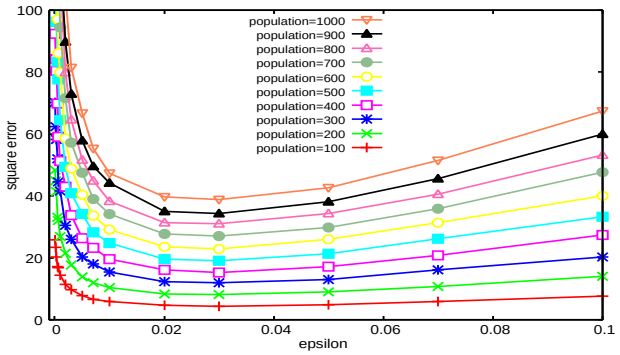

図 2: 報酬が $r_c(d_c) = \gamma_c/d_c$ の時の学習誤差の変化

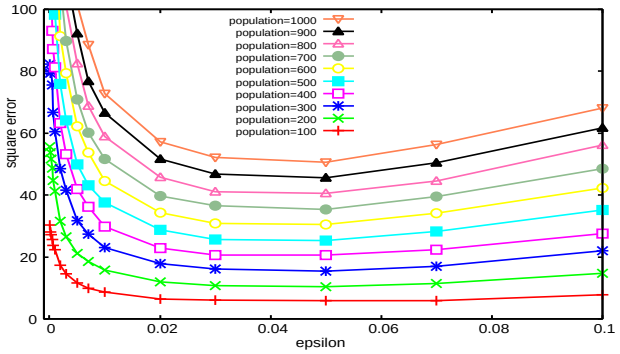

図 3: 報酬が $r_c(d_c) = \sqrt{\gamma_c/d_c}$ の時の学習誤差の変化

この性質は実験によっても確認できる。図 1 図は、ある **PG** をプレーする学習エージェントについて、その資源選択の分布の誤差 (理想の分布からの乖離) が ϵ に対しどのように変化するかを示している。この図から、この変化が下に凸の曲線を描くこと、さらには、エージェントの総数が増えるとそれに比例して誤差の大きさが増えていく事も示している。しかしその一方で、誤差を最小とする ϵ の値は、エージェント総数 N にかかわらずほぼ一定であることも示されている。これが、(2) 式で示されている、最適 ϵ の N 非依存性である。

3. 環境の変化率およびステップサイズとの関係

ここで、N 以外のパラメータにも注目してみる。

(1) 式あるいは (2) 式からわかるように、環境の変化率 σ も、最適 ϵ の決定には影響を及ぼさない。(1) 式で示されるように、変化率は $T\sigma^2$ の形で学習誤差に加えられているだけなので、誤差の大きさのみに影響する。これは実験によっても確認できる。図 4 は、強化学習のステップサイズパラメータ α [*1] を 0.001 から 0.3 と様々に変化させた時に、様々な変化率 σ (図中では fluct として表現) における、平均学習誤差の ϵ に対する変化を示している。図 1〜図 3 と同様に、この図から、変化率 σ の違いにより誤差の大きさに差は出るものの、いずれのケースでも、その誤差を最小化する 最適 ϵ の値はほとんど変化していないことがわかる。

次に、学習時間間隔 T あるいは学習のステップサイズパラメータ α と最適 ϵ との関係を調べてみる。図 4 に示した実験結果の見方を変え、変化率 σ を固定して、様々な α 毎に学習誤差平均の ϵ に対する変化をプロットしたものが図 5 である。この図からわかるように、最適 ϵ は α の値により大きく変化している。全体的な傾向としては、α が大きくなるに従ってより小さな ϵ を選ぶ必要があることがわかる。これは、α が大きい (学習時間間隔が短い) 場合には、1 つの経験に学習が大きく影響されるため、ノイズ成分となる exploration を抑える必要があることに相当する。また、α が小さければ、より多く exploration を行なっても良いことも示されている。

ここでさらに、α と ϵ を同時に最適化することを考える。図 5 で示している場合では、誤差が最小となるのは、$\alpha = 0.3$ で $\epsilon = 0.01$ 程度となる。このように α をできるだけ大きく、ϵ をできるだけ小さくすれば、全体の誤差を最小化できる事が読み取れる。ただ、これは万能ではなく、 exploration 以外の外乱が報酬に入る場合、α をある程度小さく保つ必要が出てくる。その場合には *epsilon* をある程度大きくすべきことになる。この関係がわかれば、学習により α を調整する手法 [Noda 09, George 06] と連動させ、ϵ を調整することが可能となる可能性がある。

4. おわりに

本稿では、非定常環境におけるマルチエージェント同時学習において、環境の変動率および学習のステップサイズパラメータと最適 Exploration 率の関係について、各パラメータの組み合わせの網羅的な探索によって分析を試みた。

謝辞本研究は科研費 24300064 および JST CREST の助成を受けたものである。

参考文献

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

[Martinez-Cantin 09] Martinez-Cantin, R., Freitas, de N., Brochu, E., Castellanos, J. A., and Doucet, A.: A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot., *Auton. Robots*, pp. 93–103 (2009)

[Noda 09] Noda, I.: Recursive Adaptation of Stepsize Parameter for Unstable Environments, in Taylor, M. and Tuyls, K. eds., *Proc. of ALA-2009*, pp. Paper–14 (2009)

*1 ステップサイズパラメータ α と学習時間間隔 T の間には、$T = 2/\alpha - 1$ の関係がある。

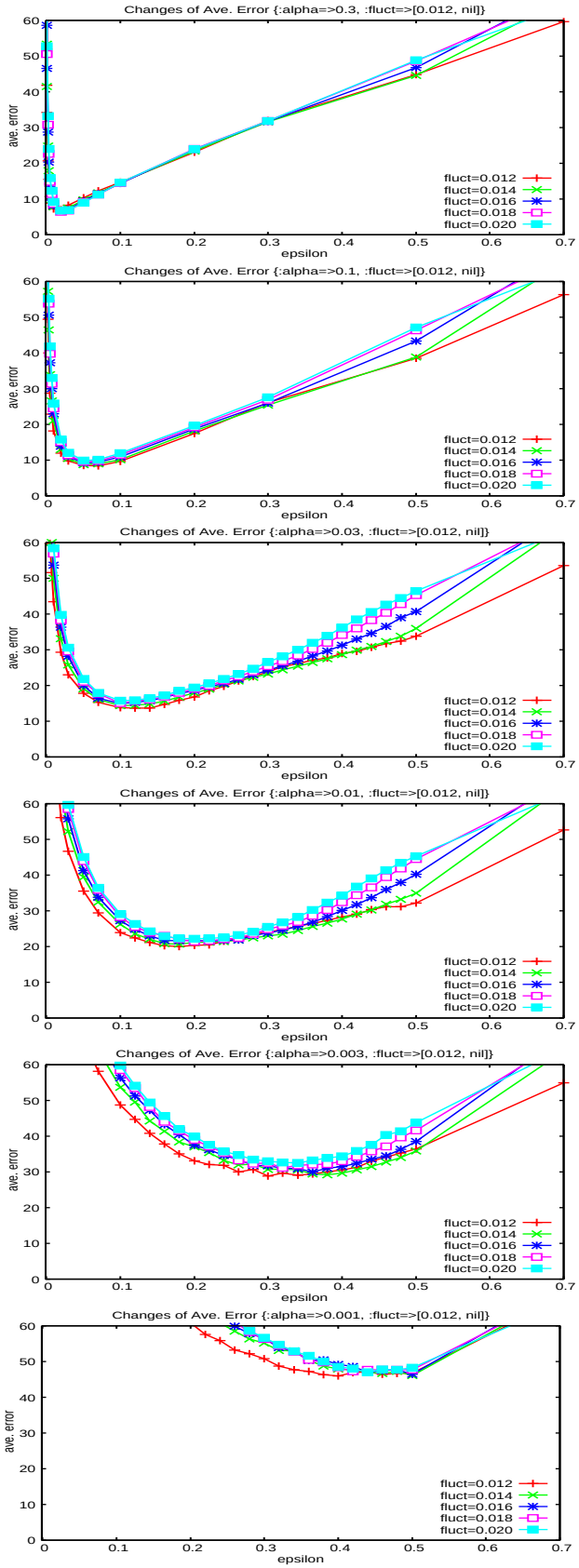

図 4: 各ステップサイズにおける学習誤差の変化

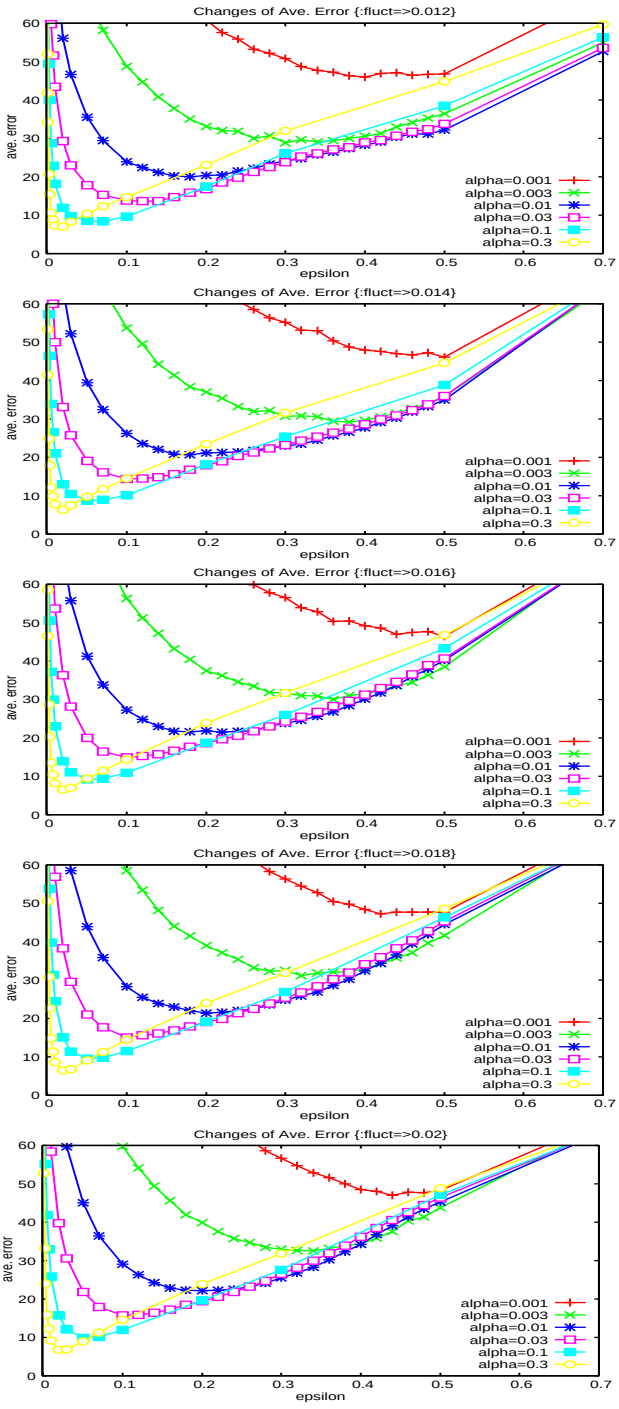

図 5: ステップサイズとの関係

[Noda 13] Noda, I.: Limitations of Simultaneous Multiagent Learning in Nonstationary Environments, in *Prof. of 2013 IEEE/WIC/ACM International Conference on INtelligent Agent Technology (IAT 2013)*, pp. paper–13, IEEE (2013)

[Reddy 11] Reddy, P. P. and Veloso, M. M.: Learned Behaviors of Multiple Autonomous Agents in Smart Grid Markets, in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAA) (2011)

[Rejeb 05] Rejeb, L., Guessoum, Z., and M'Hallah, R.: The Exploration-Exploitation Dilemma for Adaptive Agents, in *Proceedings of the Fifth European Workshop on Adaptive Agents and Multi-Agent Systems* (2005)

[Tokic 10] Tokic, M.: Adaptive e-greedy exploration in reinforcement learning based on value differences, in *Proceedings of the 33rd annual German conference on Advances in artificial intelligence (KI'10)*, Springer-Verlag (2010)

[Zhang 06] Zhang, K. and Pan, W.: The Two Facets of the Exploration-Exploitation Dilemma, in *Proceedings of the IEEE/WIC/ACM international conference on Intelligent Agent Technology (IAT-06)*, pp. 371–380, Washington, DC, USA (2006), IEEE Computer Society

[野田 13] 野田五十樹：動的環境におけるマルチエージェント同時学習における最適 Exploration に関する考察, in *JAWS 2013*JAWS2013 実行委員会 (2013)

[野田 14] 野田五十樹：非定常環境マルチエージェント学習におけるエージェント数と最適 Exploration 率の関係, 情報処理学会全国大会予稿集, pp. 3C–7 情報処理学会 (2014)