1F3-4

# Exploration 率の進化計算的改善の可能性について

## Possibility of Evolutionary Methods for Optimization of Exploration Ratio

野田五十樹[*1*2*3]
Itsuki Noda

[*1]産業技術総合研究所 AIST　[*2]JST, CREST JST, CREST　[*3]東京工業大学 Tokyo Institute of Technology

I investigate a possibility to utilize selfish evolutionary methods to find optimal exploration ratio under multi-agent learning (MAL) environment. I conducted several experiments of MAL for repeated resource sharing of nonstationary environments. The results of the experiments tell that evolutionary search methods to adjust exploration ratio in selfish-way have difficulty to reach social optimal in exploration ratio.

## 1. はじめに

本稿では、マルチエージェント学習 (以下、MAL) における各エージェントの Exploration 率 (以下、探査率) を調整・改善する方法として進化的方法が有効かどうかを、MAL 環境下での繰り返し動的資源共有問題を用いた実験により検討する。

マルチエージェント学習の環境下では、各エージェントの探査方策は系全体の挙動を決める重要な要素であるにもかかわらず、その方策の選び方の系全体への影響は十分に分析されていない。例えば、$\epsilon$ が大きすぎれば学習エージェント間の影響が拡大し、一方、$\epsilon$ が小さすぎると学習速度が遅くなるという、「探査と収穫のジレンマ」[Sutton 98] があることも知られている。しかし、そのジレンマの数理的な定式化や分析の研究はまだ少数にとどまる。

動的環境下でのマルチエージェント学習ではこのジレンマはより深刻になる。動的環境下では、環境の変化に追従するため、エージェントは環境変化より速い学習が求められる。一方で、学習を速めるために $\epsilon$ を大きくすると、相互の学習への影響が無視できなくなる。このため、環境の動的要因に応じた $\epsilon$ の設定が必要になってくる。

このようなジレンマに対して、学習のダイナミクスという点で研究が行われてきている [Wunder 10, Kaisers 10] 。しかしこれらの研究では、最適な $\epsilon$ についての議論は行われていない。また、静的な環境に対しては、[Tokic 10, Tokic 11] などの研究があるが、単エージェントの学習にとどまっており、動的環境については検討はされていない。

本稿では、実世界への応用で重要となる動的環境での MAL の学習パラメータの最適設定を探るべく、探査率とエージェントの実利得の関係から、進化的方法が可能かを調べていく。

## 2. 繰り返し動的資源共有問題

本稿では、マルチエージェントの学習の対象として、次のような **繰り返し資源共有問題** (RNRSP) を取り上げる。

> 複数エージェントがいくつかの資源を共有し、各エージェントは 1 つの資源を選択する。この選択はエージェント全体で同時に行われ、繰り返される (図 1)。各資源では、それを選択したエージェント数に応じて、各エージェントが得る利得が決められる。

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

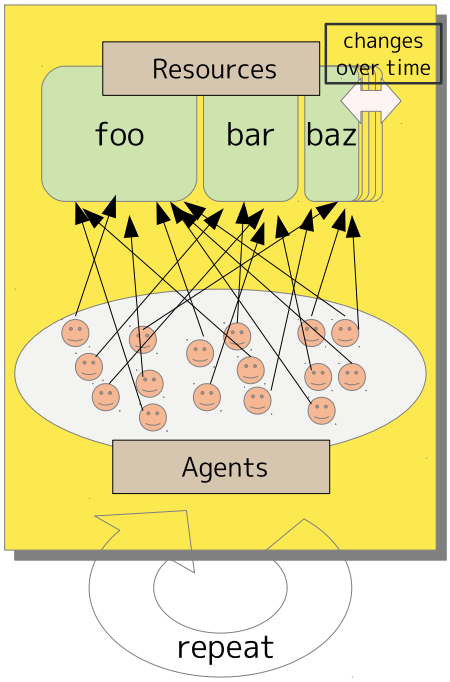


図 1: 繰り返し動的資源共有問題

> 利得は、各資源が各々持つ**容量** と、その資源を選択したエージェント数にのみ依存する。この容量は時間が経過するに従い徐々に変化するものとする。

RNRSP は、形式的には以下のようなタプルで表される。

$$\begin{aligned} \mathbf{NRSP} &= \langle \mathbf{A}, \mathbf{C}, r \rangle \\ \mathbf{A} &= \{a_1, a_2, \cdots, a_N\} \\ \mathbf{R} &= \{R_1, R_2, \cdots, R_M\} \\ r(R) &= f(d_R/\gamma_R) + \mathrm{noise}_t; \end{aligned} \quad (1)$$

ここで、$\gamma_R$ と $d_R$ は、資源 $R$ の容量と、資源 $R$ を選んだエージェント数を表している。また、各資源は異なる容量を持つものとする。一方で、利得を決める報酬関数 $f$ は全資源に共通とする。また、上記で述べているように、資源の容量 $\gamma_R$ は時間とともに変化するものとする。

なお実験では、この容量の変化は、各時刻に於いて各資源の基準容量から、ある確率で一定割合増加するという形に統一し

ている。以下ではこの変化する確率を、**変動率**と呼ぶ。

報酬関数 $f$ は、単調減少関数とする。すなわち、1 つの資源に対し、多くのエージェントがそれを選べば選ぶほど、それらのエージェントが得る報酬は減少していく。以下の実験では、この報酬関数としては $f(x) = \frac{1}{x}$ を用いる。

また、各エージェントは以下のような強化学習を行い資源選択方策を学習するものとした。エージェント $a$ は各々の資源 $R$ に対する期待報酬 $V_a(R)$ を独自に持っているものとする。エージェント $a$ は、この期待報酬に基づき、$\epsilon$-greedy 方策で資源選択を行う。つまり、$1-\epsilon$ の確率で期待報酬が最大となる資源を選択し、$\epsilon$ の確率でランダムに資源を選択するものとする。資源 $R$ を選んだ際に得られた報酬が $r$ であるとき、エージェントは資源 $R$ の期待報酬を $V_a(R) \leftarrow (1-\alpha)V_a(R) + \alpha r$ に従って修正する。このエージェントの資源選択と学習は、全エージェントが同時に行うものとする。

以下の実験評価では、エージェント学習の効率を測るため、エージェントの**資源選択分布**の均衡分布からのズレを用いる。以下ではこのズレを**学習誤差**と呼ぶことにする。ここで資源選択分布とは、各資源を選んでいるエージェントの数のベクトル $\boldsymbol{d} = \{d_R | R \in \mathbf{R}\}$ を指す。また、またその均衡分布とは、各エージェントの選択が Nash 均衡に達した段階資源選択分布を指すものとする。RNRSP の場合には、均衡分布 $\mathring{\boldsymbol{d}} = [\mathring{d}_R]$ は各資源の容量に応じた分布になる場合を均衡分布とした。分布の違いはベクトル $\boldsymbol{d}$ 同士のユークリッド距離で図るものとする。

以上のような設定について、これまでの研究により、以下のことがわかっている [Noda 13, 野田 15]。

- 学習誤差については、その下限が理論的に与えられている。その下限の探査率に対する変化は、下に凸の単峰性を成す。
- 実際の学習誤差は、上記の下限に類似の曲線を描く (最適な探査率が存在する)。

## 3. 不均一な探査率と均衡

上記の議論では、各エージェントは同じ探査率を用いているものとしている。一方、探査率を進化論的手法で改善する場合などを考えると、他と異なる探査率を持つエージェントを許す必要がある。よって、ここでは、一部のエージェントが異なる $\epsilon$ を使っている場合の、その各々のエージェントの学習効率の違いを分析していく。

### 3.1 平均利得の増減

エージェントが利己的であると仮定した場合、個々のエージェントの視点からみると、探査率をどの値にすれば、自らの学習効率、すなわち学習過程における平均利得が向上するかが最大の関心事となる。特に、探査率の調整に進化的アプローチをとったり、あるいはゲーム理論的手法で探査率の均衡点を探る場合、異なる探査率の平均利得の相対的優劣は重要な情報となる。

そこで $\epsilon$ 毎の平均利得の相対的優劣を調べるため、以下のような設定の実験を進めた。

- 探査率について、ある基準となる $\epsilon$ の値 ($\epsilon_c$) を定める。
- 180 エージェントについては、この $\epsilon_c$ を使って行動選択・学習を行うものとする。(基準値群)
- 10 エージェントについては、$0.5\epsilon_c$ となる探査率で行動選択・学習を行うものとする。(低値群)
- 10 エージェントについては、$1.5\epsilon_c$ となる探査率で行動選択・学習を行うものとする。(高値群)
- 資源の容量の変化は、ある一定の確率でランダムに資源が選択され、その資源の容量が倍増するとした。

その他の条件については、資源数を 10、資源選択および学習の回数は 10,000 回に固定して評価を行なっている。

上記の設定での各 $\epsilon_c$ に対する学習誤差の変化を図 2 に示す。この実験では、基準となる $\epsilon_c$ の値を $[0, 0.5]$ で決め、各値での学習誤差及び平均利得を求めた。また、ステップサイズ $\alpha$ については、$[0.001, 0.3]$ の間のいくつかの値について別々に実験を行った。また、環境の変化の度合いを示す変動率は、$[0.0001, 0.03]$ から選んでいる。この図では、各曲線は、各々の $\alpha$ の値に対応した学習誤差がプロットされている。これらの曲線は、概ね下に凸の単峰性となっており、その最低点となる最適探査率 $\epsilon^*$ が存在することがわかる。

次に、標準値群・低値群・高値群別にエージェントの平均報酬について調べてみる。シミュレーション結果の処理は以下のように行った。

1. 標準値群・低値群・高値群の各々のエージェント群において、その中での平均利得を求める。
2. 標準値群の平均利得を基準 1 とした際の、低値群・高値群の平均利得の相対利得率を求める。

図 3 は、基準 $\epsilon$ の変化に対する相対利得率の変化を示している。各曲線は環境の変動率の違いを表している。いずれの場合も、低値群の相対利得率は $\epsilon$ が大きくなるに従い大きくなり、逆に高値群では小さくなっていく事がわかる。そして、中間的な付近で 2 つの曲線が交差している。

### 3.2 相対利得率による均衡点と最適点

ここで、進化的手法で $\epsilon$ 値を修正する方法を検討する。

図 4 は、実験 2 で得られた高値群・低値群の相対利得率の変化を模式化したものである。3.1 節でも議論したように、低値群の利得率は 1 以下から 1 以上に変化し、高値群はその逆の動きをする。そして、真ん中辺りで 2 本の曲線は交差する。これにより、進化的アプローチでは、$\epsilon$ はこの交点に収束することになる。また、ゲーム理論的にも、この点が均衡点となる。つまり、他のエージェントの平均利得と探査率 $\epsilon$ がエージェント相互に参照できる環境下で各エージェントに利己的に $\epsilon$ 値を修正させると、系全体としてこの交点に $\epsilon$ が収束していくことが期待できる。

一方、系全体の性能の視点からこの収束点を見てみる。図 5 は、図 2 で示した平均学習誤差の変化を模式的にしたものであるが、これからわかるように、社会的最適に近づけるためには、この平均誤差が最小化されるような探査率 (最適点) を選ぶ必要がある。

そこで、これらの均衡点と最適点の関係を見て見るため、各実験設定ごとにこれらを求めプロットした (図 6)。また、環境の変化のさせ方を異なる設定にした場合の結果についても図 7 に示す。これら図からわかるように、最適点が小さい領域では、均衡点はより小さな値をとりがちであり、一方、最適点がある程度以上大きな当たりでは、均衡点は急に大きくなり、最適点より大きくなりがちである。このように、エージェントが利己的に探査率を変化させる場合は、社会的最適からは外れた探査率をエージェントはとりがちであることがわかる。

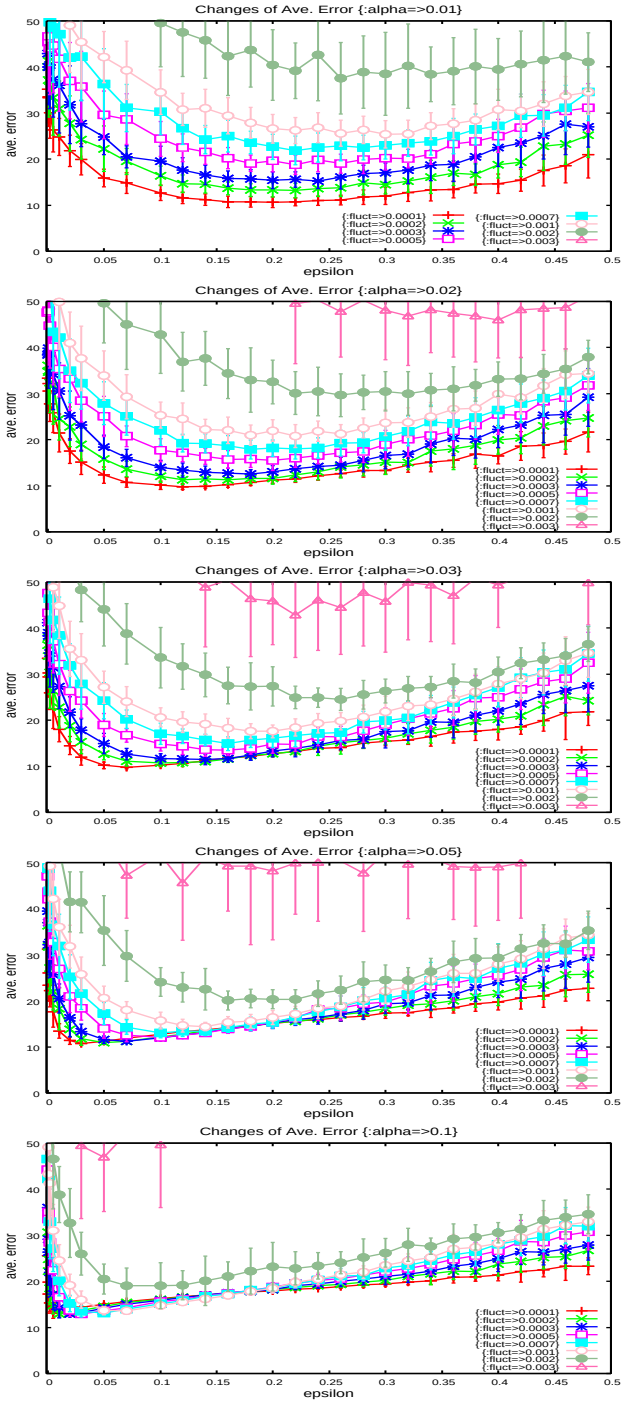


図 2: Average Error

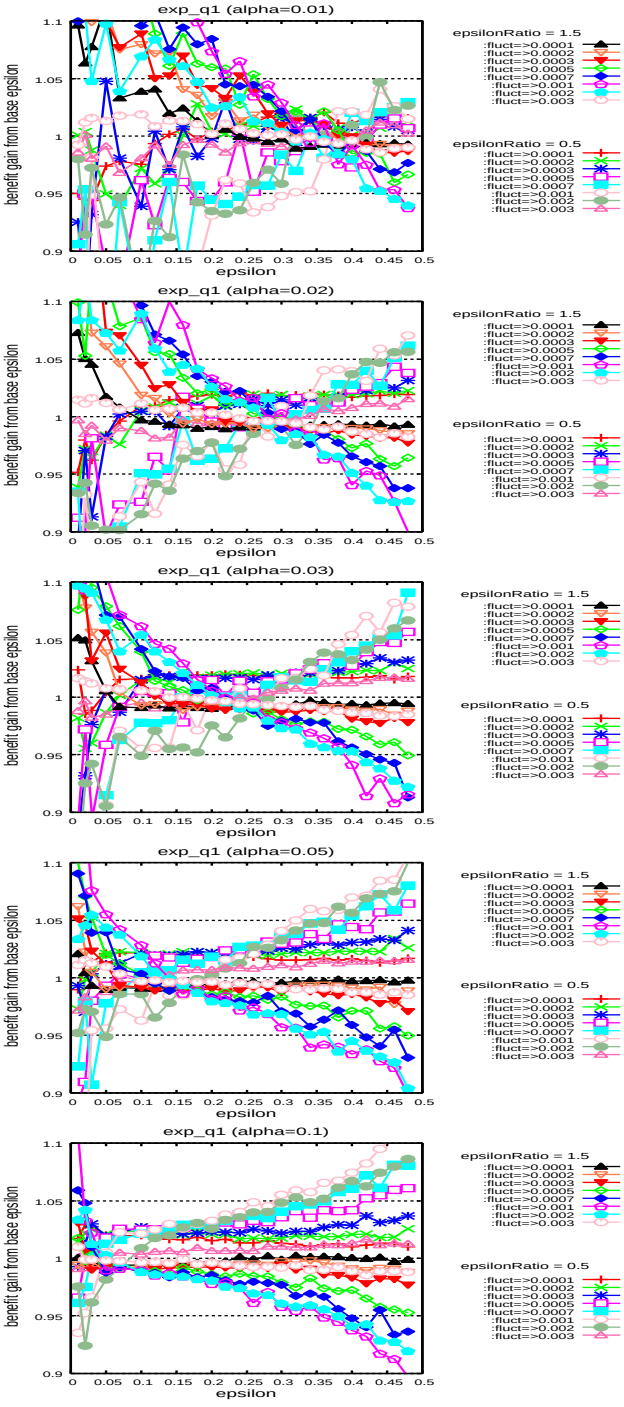


図 3: Average Benefit Gain

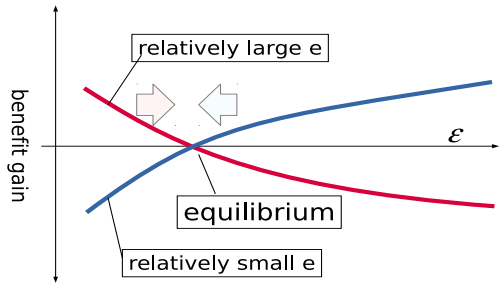


図 4: 高値群・低値群の相対利得率の変化の模式図と均衡点

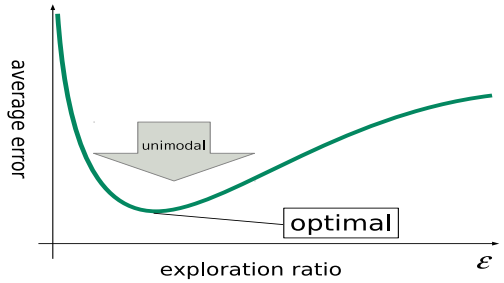


図 5: 平均誤差の模式図と最適点

## 4. まとめ

本稿では、マルチエージェント学習における重要な学習パラメータである探査率について、その最適性と均衡性の間の関係を分析し、進化的に探査率を改善する方法の可能性について検討した。その結果、残念ながら、単純な進化的手法で探査率を調整すると、社会的最適に適した探査率にはならない可能性が示された。

今後は、進化的手法に社会的最適に仕向ける社会的規範をエージェント学習や進化的手法に取り入れる方法を検討する必要がある。

**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

[Kaisers 10] Kaisers, M. and Tuyls, K.: Frequency Adjusted Multi-agent Q-learning, in *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pp. 309–315 (2010)

[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. 309 – 314, IEEE (2013)

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA (1998)

[Tokic 10] Tokic, M.: Adaptive epsilon-greedy exploration in reinforcement learning based on value differences, in *KI 2010 Proc. of 33rd annual German Conference on Advances in Artificial Intelligence*, pp. 203–210, Springer (2010)

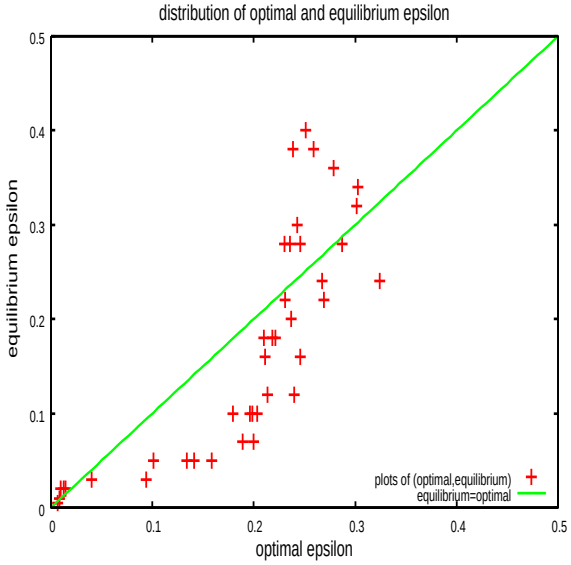


図 6: 均衡点と最適点の関係

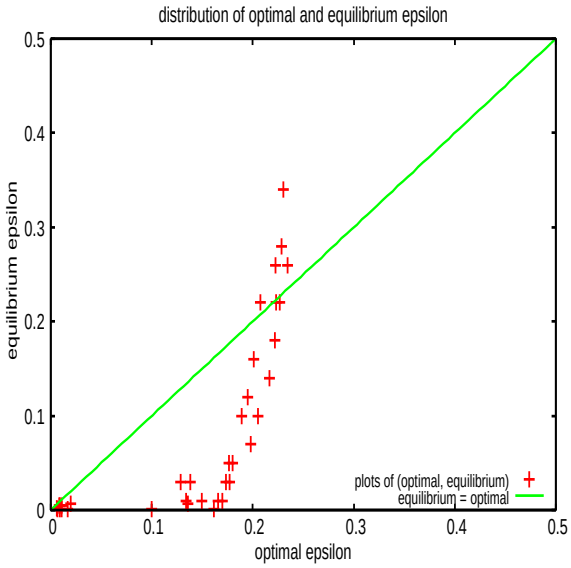


図 7: 均衡点と最適点の関係 (異なる設定)

[Tokic 11] Tokic, M. and Palm, G.: Value-Difference Based Exploration: Adaptive Control between Epsilon-Greedy and Softmax, in *KI 2011: Advances in Artificial Intelligence*, pp. 335–346, Springer (2011)

[Wunder 10] Wunder, M., Littman, M. L., and Babes, M.: Classes of Multiagent Q-learning Dynamics with epsilon-greedy Exploration, in Frnkranz, J. and Joachims, T. eds., *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1167–1174, Omnipress (2010)

[野田 15] 野田五十樹：探査率の最適性と均衡性に関する検討, 社会システムと情報技術研究ウィーク (2015)