2F1-4

動的環境におけるマルチエージェント同時学習に関する考察

Discusssion on Limitation of Simultaneous Multiagent Learning in Unstationary Environment

野田五十樹[*1*2*3]
Itsuki Noda

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

In MAL, exploration of an agent affect to others' learning as a noise, while exploration is necessary to catch-up changes of the unstationary environment. We formalize the MAL situation from the viewpoint of learning of probability distribution, in which the purpose of the learning is defined as maximization of probability to choose the right that acquires more benefit than other actions. An agent need to explore all actions to check and confirm the best action especially in an unstationary environment, in which the best action may change over time even if other agents do not change their policies. Based on the formalization, we investigate a simple case of resource sharing problems to show existence of the boundary of learning performance that limit the convergence of learning to the right policy.

1. はじめに

動的環境における行動選択の制御は強化学習に置いて重要な問題である。特に、exploration 率は学習エージェントの性能を左右する基本的パラメータである。エージェントが動的環境に対しても迅速に適応するためには、exploration を多く用いる必要がある。一方、マルチエージェント環境での同時学習を考えると、exploration 率はより大きな問題をエージェント群に与える。マルチエージェント環境下ではあるエージェントは他のエージェントにとっては環境の一部であるため、そのエージェントの行動選択機構の変化は他エージェントの環境変化として相互に影響する。

exploration 率の制御方法については、これまでいくつかの研究が行われてきている [Zhang 06, Martinez-Cantin 09, Rejeb 05, Tokic 10, Reddy 11] が、それらの多くは静的環境あるいは単エージェント環境を仮定している。一方、実世界問題の多くは、例えば省エネと利便性を両立させるスマートシティなどのアプリケーションのように、問題設定は動的環境・マルチエージェント環境であり、従来の研究を単純に適用することはできない。

本稿では、この問題に取り組むため、最適行動選択確率の視点を導入して問題の新たな定式化を行う。

2. 問題の定式化

2.1 Population Game

本節ではまず、マルチエージェントゲームの一つである population game を定式化する。population game は多数のエージェントが参加するゲームであり、各エージェントは有限の選択肢の中から 1 つずつ選択するものとする。各エージェントは、同じ選択肢を選んだエージェント数に応じて報酬を得る。その選択は同時に行われ、それが無限に繰り返される。

population game **PG** は以下のように定義される。

$$\mathbf{PG} \quad = \quad \langle \mathbf{A}, \mathbf{C}, \boldsymbol{r} \rangle \tag{1}$$

連絡先: 野田五十樹、産業技術総合研究所 サービス工学研究センター、茨城県つくば市梅園 1-1-1、Tel:029-861-3298、E-mail: I.Noda@aist.go.jp

ただし、$\mathbf{A} = \{a_1, a_2, \cdots, a_N\}$ はエージェント集合、$\mathbf{C} = \{c_1, c_2, \cdots, c_K\}$ は行動選択肢集合、$\boldsymbol{r} = \{r_a | a \in \mathbf{A}\}$ は報酬関数集合である。報酬関数 $r_a(c; \boldsymbol{d}_a^-)$ はエージェント a が行動 c を選択し、かつ、エージェント a 以外のその時の行動選択分布が $\boldsymbol{d}_a^-$ である場合の報酬を示す。ここで、行動選択分布が $\boldsymbol{d}_a^-$ とは、行動 c' を選んだ a 以外のエージェントの数 $d_{a,c'}^-$ のベクトル $\left[\ d_{a,c'}^- | c' \in \mathbf{C} \ \right]$ であるとする。また、ここでは、実際にエージェントが得られる報酬値 r_a は確率的であり、実際には各報酬関数は報酬値の確率分布を与えるものとする。

2.2 優勢確率

ここで、各エージェント a の**優勢確率** (*advantageous probability*, AP) $\rho_a(c; \boldsymbol{d}_a^-)$ を導入する。これは、エージェント a が行動 c を選択した場合、その選択が、他のエージェントが行動を変えない (行動選択分布 $\boldsymbol{d}_a^-$ が変化しない) 条件で、他のどの選択を行った場合よりも高い報酬得を得る確率と定義される。すなわち、AP は以下のように定義される。

$$\rho_a(c; \boldsymbol{d}_a^-) \quad = \quad \mathcal{P}\left(\forall c' \in \mathbf{C} : r_a(c; \boldsymbol{d}_a^-) \geq r_a(c'; \boldsymbol{d}_a^-)\right), \tag{2}$$

ただし、$\mathcal{P}$ (⟨ 条件 ⟩) は '⟨ 条件 ⟩' が成立する確率を示している。

2.3 最適優勢選択

次に、この AP を使って、**最適優勢選択** を以下のように定義する。すなわち、行動 $\mathring{c}$ が**最適優勢選択**であるとは、その AP $\rho_a(\mathring{c}_a)$ が行動選択集合 $\mathbf{C}$ の中で最大となることをいう。

$$\mathring{c}_a \quad = \quad \arg\max_{c \in \mathbf{C}} \rho_a(c) \tag{3}$$

もちろん、各エージェントにとっては、他の全エージェントの行動、あるいは行動選択分布 $\boldsymbol{d}_a^-$ を知ることは一般にできない。よって、学習エージェントの目的は、各々のエージェントの行動報酬に基づいて、各行動の AP を求め、それに基づき最適優勢選択を選べるようになることになる。この学習途中の各行動 c の AP を**学習** AP $\tilde{\rho}_a(c)$ と表す。

2.4 理想行動選択分布

この学習 AP を用いて、exploitation/exploration を次のように定義する。エージェント a が exploit しているとは、その

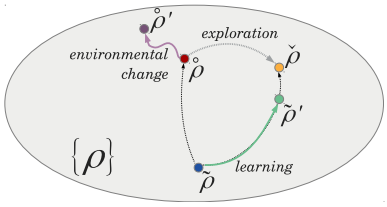


図 1: 各優勢確率の関係

エージェントが学習 AP$\tilde{\rho}_a$ の中で最大となる行動 $\hat{c}$ を選んでいる場合を指し、それ以外の行動を選択している場合を explore していると呼ぶ。

また、全エージェントがその学習結果に基づいて最適行動を選択している場合の**理想行動選択分布** $\mathring{\boldsymbol{d}}$ を以下のように定義する。

$$\mathring{\boldsymbol{d}} \quad = \quad \left[\ \mathring{d}_c | c \in \mathbf{C} \ \right]$$

$$\mathring{d}_c \quad = \quad \text{exploit により行動 } c \text{ を選択しているエージェントの数}$$

また、エージェント a 以外についての理想行動分布は、$\mathring{\boldsymbol{d}}^-_a$ と記述することにする。

これらを用いて、エージェント a にとっての**理想** AP を以下のように定義する。

$$\mathring{\rho}_a(c) \quad = \quad \rho_a(c; \mathring{\boldsymbol{d}}^-_a) \tag{4}$$

3. Exploration と学習の相互作用

本稿におけるエージェントの学習の目的は、各エージェント a が、学習 AP$\tilde{\rho}_a$ をできるだけ理想 AP $\mathring{\rho}_a$ に近づけることである。そして、もし、全エージェントが $\tilde{\rho}_a = \mathring{\rho}_a$ を達成できた場合、エージェントの行動選択の学習は均衡に達する。

この学習目的を達成するためには、各エージェントは可能なすべての行動選択を試す (explore) 必要がある。また、1. 節で述べているように、**PG** の報酬関数が時間とともに変化するような動的環境を仮定すると、エージェントは、たとえある事典で均衡に達したとしても、絶えず explore を続ける必要がある。そこでここでは、exploration 率が一定のまま保たれると仮定して以下の議論をすすめる。

3.1 実行行動選択分布

マルチエージェント学習環境下では、同時に複数のエージェントが explore する可能性があるため、行動選択分布 $\boldsymbol{d}$ はその理想 $\mathring{\boldsymbol{d}}$ からずれることになる。この実際の行動選択分布を**実効行動選択分布** と呼び、$\check{\boldsymbol{d}}$ と表すことにする。またこれまでと同様に、エージェント a 以外の実効行動選択分布を $\check{\boldsymbol{d}}^-_a$ と記述することにする。

これらを使い、新たに、学習環境下でのエージェント a の AP である **実効** AP を以下のように定義する。

$$\check{\rho}_a(c) \quad = \quad \rho_a(c; \check{\boldsymbol{d}}^-_a) \tag{5}$$

3.2 各優勢確率の関係

理想 AP $\mathring{\rho}$、学習 AP $\tilde{\rho}$、 実効 AP $\check{\rho}$ の関係を表したものが図 1 である。すなわちある時点において、全エージェントが exploit している場合に、それによって得られるエージェント a にとっての理想 AP$\mathring{\rho}$ が、確率分布関数空間上に存在する。同時に、その時点でエージェント a が学習結果として獲得している AP の推定値 $\tilde{\rho}$ も同じ空間上にマップされる。上記したように、学習の目的は $\tilde{\rho}$ を $\mathring{\rho}$ に近づけることであるが、同時学習であるため、いくつかのエージェントは explore することになり、エージェント a が実際に得られる報酬サンプルのもととなる実効 AP$\check{\rho}$ は、理想 AP$\mathring{\rho}$ からずれることになる。このため、そのサンプルを用いた学習によりの修正方向は $\tilde{\rho}$ を $\check{\rho}$ に近づけることになり、新たな学習結果として $\tilde{\rho}'$ を得る。一方で、その学習の間に動的環境では理想 AP は変化し、$\mathring{\rho}'$ のようにずれていってしまう。

よって、学習全体がどの程度均衡に近づいていけるかは、exploration による理想 AP と実効 AP のズレ、学習 AP が目標 (実効 AP) に近づく速さ、および動的環境における理想 AP の移動の速さとの関係に依存することになる。そこで以下では、各々の距離および速度についての考察をすすめる。

3.3 学習における社会的対称性と学習精度

ここで、エージェントの学習に関する下記のような社会的対称性を仮定する。すべてのエージェントは、その学習に際し、同じ ϵ 値を用いた ϵ-greedy により行動選択をするものとする。すなわち、任意のエージェント a は、確率 $(1-\epsilon)$ で最適優勢選択 $\hat{c}_a$ を選択し (exploit)、それ以外の時は、すべての可能な行動を当確率で選択する (explore) ものとする。

このような仮定のもとでは、a にとっての実効行動分布 $\check{\boldsymbol{d}}^-_a$ は、その理想行動分布 $\mathring{\boldsymbol{d}}^-_a$ から以下のようにずれる。上記の社会的対称性より、行動 c を選択するエージェントの数 $\mathring{d}^-_{a,c}$ は、exploit を続けるエージェントの分布 $\mathcal{B}(d; \mathring{d}^-_{a,c}, 1-\epsilon)$ と explore して行動 c を選択するエージェントの分布 $\mathcal{B}(d; N, \epsilon/K)$ の畳み込みとなる。ここで、$\mathcal{B}(x; N, p)$ は、総数 N、確率 p による二項分布である。よって、$\mathring{d}^-_{a,c}$ は以下のように近似することができる。

$$\begin{aligned} \mathcal{P}\left(\mathring{d}^-_{a,c} = d\right) \quad &= \quad \mathcal{B}(d; \mathring{d}^-_{a,c}, 1-\epsilon) * \mathcal{B}(d; n, \frac{\epsilon}{K}) \\ &\approx \quad \mathcal{G}(d; [\mathring{d}^-_{a,c} + \epsilon(\frac{N}{K} - \mathring{d}^-_{a,c})], \\ &\qquad [\epsilon(1-\epsilon)\mathring{d}^-_{a,c} + \frac{\epsilon}{K}(1-\frac{\epsilon}{K})N]), \end{aligned} \tag{6}$$

ただし、$\mathcal{G}(x; \mu, \sigma^2)$ は平均 μ、 分散 σ^2 の正規分布である。

これにより、$\mathring{d}^-_{a,c}$ と $\check{d}^-_{a,c}$ の平均二乗誤差は次のようになる。

$$\begin{aligned} \mathrm{E}\left[(\mathring{d}^-_{a,c} - \check{d}^-_{a,c})^2\right] \quad &= \quad \epsilon^2(\frac{N}{K} - \mathring{d}^-_{a,c})^2 + \epsilon(1-\epsilon)\mathring{d}^-_{a,c} \\ &\qquad + \frac{\epsilon}{K}(1-\frac{\epsilon}{K})N] \end{aligned} \tag{7}$$

この式は、$\mathring{d}^-_{a,c}$ と $\check{d}^-_{a,c}$) の差 (の二乗) が exploitation 率 ϵ とその二乗 ϵ^2 の線形和となることを示している。

一方、学習による学習 AP の変化速度については、以下のように考えることができる。まず、各エージェント a はある単位学習時間 T 内で、最適優先選択意外の行動については explore している間に $\frac{\epsilon T}{K}$ 回ずつ報酬値をサンプリングできる。すなわち、それに比例した回数だけ実効 AP$\tilde{\rho}_a$ をサンプリングでき、それをもとに、エージェント a は学習 AP$\tilde{\rho}_a$ を変化させ $\tilde{\rho'}_a$ へと AP のパラメータを変化させる。ここで、$\tilde{\boldsymbol{d'}}^-_a$ を確率 $\tilde{\rho'}_a$ のパラメータとみなせば、Cramer-Rao の定理により、以下の不等式を導出できる。

$$\mathrm{E}\left[(\check{d}^-_{a,c} - \tilde{d'}^-_{a,c})^2\right] \quad \geq \quad \mathrm{Var}\left(\tilde{d'}^-_{a,c}\right) \geq \frac{K}{\epsilon T \tilde{g}_{a,c}} \tag{8}$$

$$\tilde{g}_{a,c} \quad = \quad \frac{\partial \log \tilde{\rho'}_a}{\partial \tilde{d'}^-_{a,c}}$$

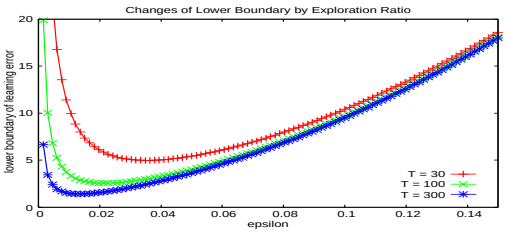


図 2: 学習精度の下限の Exploration Ratio ϵ による変化

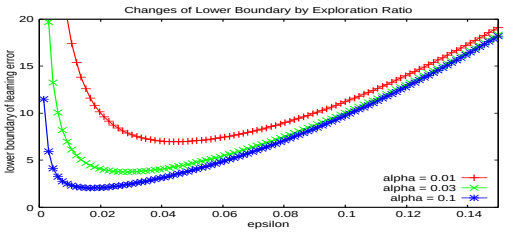


図 3: EMA を用いた場合の学習精度の下限

すなわち、$\check{d}^-_{a,c}$ と $\tilde{d'}^-_{a,c}$ の平均二乗距離の下界、つまり $\frac{\epsilon T}{K}$ 回のサンプリング学習による関数近似誤差の下界は、$\frac{1}{\epsilon}$ に比例して増加することになる。

以上の式 (7) と (8) をまとめると、次の定理を導ける。

定理

PG において全エージェントが ϵ-greedy で同時学習を行う場合、それで得られる学習誤差の二乗平均 $\mathrm{E}\left[\left\|\mathring{\boldsymbol{d}}^-_a - \tilde{\boldsymbol{d}'}^-_a\right\|^2\right]$ は、$\frac{1}{\epsilon}$ と ϵ と ϵ^2 の線形和で表される。 □

証明

$$\begin{aligned}
&\mathrm{E}\left[\left\|\mathring{\boldsymbol{d}}^-_a - \tilde{\boldsymbol{d}'}^-_a\right\|^2\right] \\
=\ &\mathrm{E}\left[\left\|\mathring{\boldsymbol{d}}^-_a - \check{\boldsymbol{d}}^-_a\right\|^2\right] + \mathrm{E}\left[\left\|\check{\boldsymbol{d}}^-_a - \tilde{\boldsymbol{d}'}^-_a\right\|^2\right] \\
\geq\ &\frac{K}{\epsilon T \tilde{g}_{a,c}} + \epsilon^2\left(\frac{N}{K} - \mathring{d}^-_{a,c}\right)^2 \\
&+\epsilon(1-\epsilon)\mathring{d}^-_{a,c} + \frac{\epsilon}{K}\left(1 - \frac{\epsilon}{K}\right)N \qquad (9)
\end{aligned}$$

■

3.4 学習精度の下限と学習速度

上記の定理の中の式 (9) で示した下限の典型的な変化を示したのが図 2 である。このグラフの中で、各曲線は異なる学習単位時間 T に対する平均学習誤差の下限を表している。これからわかるように、exploration 率を制御するパラメータ ϵ について、下限値を最小化する最適な値が存在することがわかる。また、最適な ϵ をとったとしても、かならず正の平均学習誤差が残ることもわかる。また、学習単位時間 T について見てみると、長い T であるほど下限値が小さくなることがわかる。これは、学習をゆっくり時間かけて行えば、より学習 AP を理想 AP に近づけることができることを示している。

しかし、最適で比較的小さな ϵ と長い T を用いた場合、動的環境の変化に追従できないという問題が生じる。例えば、動的環境の変化が連続的であり、理想 AP $\mathring{\rho}_a$ を決めるパラメータ $\mathring{\boldsymbol{d}}^-_a$ がランダムウォークで変化する場合を考える。ここで図 1 に示しているように、時刻 t の理想 AP を $\mathring{\rho}_a$、T 回のランダムウォーク後である時刻 $t+T$ の理想 AP を $\mathring{\rho}'_a$ と書くとし、各々のパラメータを $\mathring{\boldsymbol{d}}^-_a$、$\mathring{\boldsymbol{d}'}^-_a$ とする。これらのパラメータの変化がランダムウォークと仮定すれば、その性質により、$\mathring{\boldsymbol{d}}^-_a$ と $\mathring{\boldsymbol{d}'}^-_a$ の平均二乗距離は以下のようになる。

$$\mathrm{E}\left[\left\|\mathring{\boldsymbol{d}}^-_a - \mathring{\boldsymbol{d}'}^-_a\right\|^2\right] = T\sigma^2, \qquad (10)$$

ここで、σ^2 はランダムウォークの各サイクルのステップの分散である。

本節の最初に述べているように、各エージェントの学習の目的は、最新の理想 AP $\mathring{\boldsymbol{d}'}^-_a$ を近似することである。よって、時刻 $t+T$ における学習による誤差は、下記のように、式 (9) にランダムウォーク分の偏差を加える形となる。

$$\mathrm{Error} = \mathrm{E}\left[\left\|\mathring{\boldsymbol{d}'}^-_a - \tilde{\boldsymbol{d}'}^-_a\right\|^2\right] \qquad (11)$$

寄って、学習誤差の下限は以下のようになる。

$$\begin{aligned}
\mathrm{Error} &= \mathrm{E}\left[\left\|\mathring{\boldsymbol{d}'}^-_a - \mathring{\boldsymbol{d}}^-_a\right\|^2\right] + \mathrm{E}\left[\left\|\mathring{\boldsymbol{d}}^-_a - \check{\boldsymbol{d}}^-_a\right\|^2\right] \\
&\quad +\mathrm{E}\left[\left\|\check{\boldsymbol{d}}^-_a - \tilde{\boldsymbol{d}'}^-_a\right\|^2\right] \\
&\geq T\sigma^2 + \frac{K}{\epsilon T \tilde{g}_{a,c}} + \epsilon^2\left(\frac{N}{K} - \mathring{d}^-_{a,c}\right)^2 \\
&\quad +\epsilon(1-\epsilon)\mathring{d}^-_{a,c} + \frac{\epsilon}{K}\left(1 - \frac{\epsilon}{K}\right)N \qquad (12)
\end{aligned}$$

ここで、この下限には学習時間 T が入っている。つまり、学習時間がながければ、誤差が大きくなることを示していることになる。これが、先に述べたように、小さな ϵ と長い T による時間をかけた学習が動的環境ではうまく機能しないことに対応する。

3.5 指数移動平均による学習精度の下限

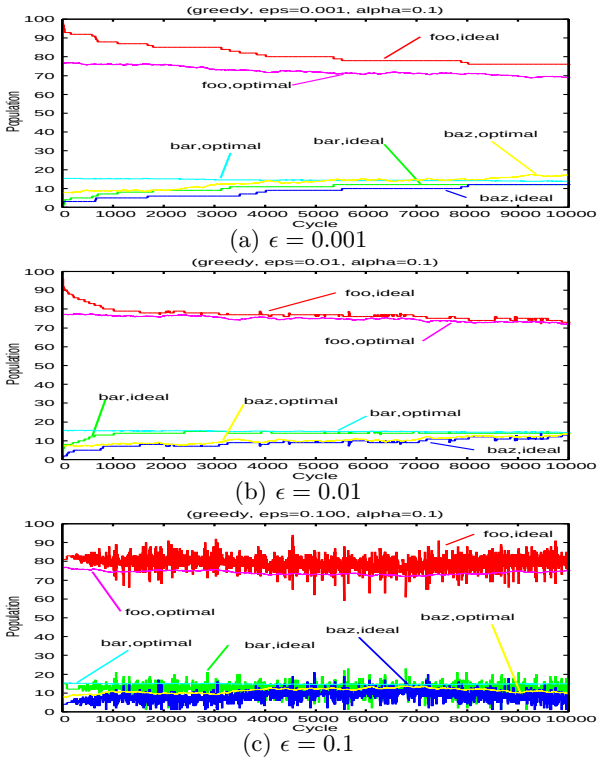
以上では、学習がある単位学習時間 T の間にサンプリングしたものをまとめてバッチ的の処理したものとして計算してきたが、より連続的な学習方式として、指数移動平均 (exponential moving average, EMA)、$\bar{x}_{t+1} \leftarrow (1-\alpha)\bar{x}_t + \alpha x_t$ を使っても同じような結果を導き出すことができる。ステップサイズパラメータ α を用いた指数移動平均を用いた強化学習は、バッチ的な学習の時間窓幅 τ と $\tau = 2/\alpha - 1$ のような関係が有ることがわかっている [Noda 09b, Noda 09a]。この関係を使えば $\epsilon T = 2/\alpha - 1$ とみなせるので、学習誤差の下限として以下のような式を導き出すことができる。

$$\begin{aligned}
\mathrm{Error} \geq\ &\frac{(2-\alpha)\sigma^2}{\alpha\epsilon} + \frac{\alpha K}{(2-\alpha)\tilde{g}_{a,c}} + \epsilon^2\left(\frac{N}{K} - \mathring{d}^-_{a,c}\right)^2 \\
&+\epsilon(1-\epsilon)\mathring{d}^-_{a,c} + \frac{\epsilon}{K}\left(1 - \frac{\epsilon}{K}\right)N \qquad (13)
\end{aligned}$$

この EMA を用いた場合の誤差の下限曲線を示したものが図 3 である。図 2 と同様の形状を描くが、式 (3) の第二項のため、全体として下限値はもとの曲線よりも上にずれることになり、最適な ϵ も大きめとなる。これが、動的環境における explration 率のトレードオフである。

4. 実験および考察

最後に、前節までで示してきた下限値の構造、すなわち、ϵ に最適値が有ることの実際の学習への影響を確認するために、



(a) $\epsilon = 0.001$

(b) $\epsilon = 0.01$

(c) $\epsilon = 0.1$

図 4: ϵ-greedy による実験結果

以下のような単純な population game を仮定し、学習実験を行った。

$$
\begin{aligned}
\mathbf{PG} &= \langle \mathbf{A}, \mathbf{C}, \boldsymbol{r} \rangle \\
\mathbf{A} &= \{a_1, a_2, \cdots, a_{100}\} \\
\mathbf{C} &= \{\mathrm{foo}, \mathrm{bar}, \mathrm{baz}\} \\
\boldsymbol{r} &= \{r_a | \forall a \in \mathbf{A}, \forall c \in \mathbf{C} : \quad r_a(c) = r(c) = B - (d_c/\gamma_c)\} \\
B &: \text{報酬のオフセット (定数)} \\
\gamma_c &: \text{選択肢 } c \text{ の容量} \\
&\quad \gamma_{\mathrm{foo}} = 100; \gamma_{\mathrm{bar}} = 20; \\
&\quad \gamma_{\mathrm{baz}} = 10 \quad \text{at beginning;}
\end{aligned}
$$

また、このゲームに動的性を持たせるために、パラメータ γ_{baz} はランダムウォークするものとした。ただし、この変化は一様分布 $[-0.01, 0.01]$ に従って決められるものとする。

各エージェントは、各々期待報酬関数をもち、強化学習によってそれを修正していくものとする。エージェントはランダムに explore と exploit を行い学習をすすめるが、その戦略としては、ϵ-greedy を用いた。

ステップサイズパラメータについては $\alpha = 0.1$、exploration 率は $\epsilon \in \{0.0, 0.001, 0.01, 0.1\}$ として各場合の学習実験を行った。まず、図 4 は ϵ が 0.001、 0.01、 0.1 の各場合における ϵ-greedy エージェントによる学習結果である。学習 AP に対応する分布 $(\tilde{d}_c)$(つまり行動 c を最適とみなしているエージェントの数) と、問題設定より求めた真の最適分布を示している。

これらのグラフの中で、図 4-(a) は学習 AP に対する分布は、真の最適分布に追従しようとしているが、ϵ が小さすぎるため、追従しきれていない状態を示している。また、図 4-(c) では ϵ が大きすぎるため、エージェント相互の explore の影響が大きく、ノイズ成分が大きくなってしまっている。図 4-(b) は比較的最適な ϵ に近いと見られ、環境変化に追従しつつ、相互のノイズも小さい状態が保たれているといえる。

いずれにしても、ϵ が小さすぎる、あるいは大きすぎる場合には、学習が適切には進まないことが、これらの結果から見て取れる。

5. まとめ

本稿では、動的環境におけるマルチエージェント学習における、exploration 率と学習誤差の関係についての理論的分析を行った。その結果、平均学習誤差の下限が exploration 率とその逆数および逆数の 2 乗の線形和で表されることを示した。これにより、学習誤差を最適にする exploration 率は正の値であること、また、その場合においても有限の平均誤差が残ることを示した。

3. 節で導入した学習目的の定義は、従来のエージェント学習の定義とは少し異なっている。通常、強化学習などのエージェント学習の目的は、期待報酬 (average reward, AR) の最大化である。一方、本稿では以下の議論のため、あくまで他の選択肢より多くの報酬をもらえる確率を最大化することという目的に変更している。これは主に以下ですすめる議論を完結にするためのものである。今後は、この違いについても解析をすすめる必要がある。

謝辞
本研究は科研費 24300064 の助成を受けたものである。

参考文献

[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 09a] Noda, I.: Adaptation of Stepsize Parameter for Non-Stationary Environments by Recursive Exponential Moving Average, in *Prof. of ECML 2009 LNIID Workshop*, pp. 24–31ECML (2009)

[Noda 09b] 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)

[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