3C-7

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

野田 五十樹 †
(独) 産業技術総合研究所 サービス工学研究センター
JST, CREST

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。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扱う形式的な方法を提案してきた。本稿では、その形式化を基に、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を調べ、エージェントの総数が最適 exploration 率の決定に寄与しないことを示す。さらに、その関係をいくつかの実験によって確認する。

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 [2, 1]。

定理 2.1
各エージェントの学習誤差の下限は以下の式で与えられる。

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

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。

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

□

3 エージェント総数と最適 Exploration 率

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。この最

Relation between Agent Population and Optimal Exploration Ratio of Multiagent Learning for Nonstationary Environments

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

適 ϵ を解析的に求めるのはまだ困難であるため、最適 ϵ と他のパラメータがどのような関係にあるかを解析する。特に本稿では、エージェントの総数 N に着目し、それと最適 ϵ の関係を調べる。

まずそのために、いくつかの仮定を置く。

- 行動 c の報酬 r_c は、c に関わらず一定の単調現象報酬関数 ψ により、$r_c(d_c) = \psi\left(\frac{d_c}{\gamma_c}\right)$ により決まるものとする。ただし、γ_c は正の定数であり、d_c は c を選択しているエージェント数である。
- 学習の平衡状態では、各行動 c の報酬はすべて同じ (均衡) であるとする。その結果、各優勢確率も同じであるとする。
- exploration による分布 $\boldsymbol{d}$ のゆらぎ Δd_c による報酬のゆらぎは $\Delta r_c = \frac{\Delta d_c}{\gamma_c}\bar{\psi}'$ で近似できるとする。
- 分布 d_c のゆらぎによる優勢確率のゆらぎ $\frac{\partial \rho_a(c')}{\partial d_c}$ は以下の式で表せるとする。

$$\frac{\partial \rho_a(c')}{\partial d_c} \propto \begin{cases} \frac{\bar{\psi}'}{\gamma_c} \cdot \mathcal{P}\left(\Delta r_{c'} = 0\right) & ;c' = c\text{の時} \\ \frac{-\bar{\psi}'}{(K-1)\gamma_c} \cdot \mathcal{P}\left(\Delta r_{c'} = 0\right) & ;c' \neq c\text{の時} \end{cases}$$

これらをもとに、$\frac{\partial \mathcal{L}}{\partial \epsilon} = 0$ を展開すると、次の式が得られる。

$$\frac{1}{T}\frac{\partial}{\partial \epsilon}\left(\frac{Q}{\epsilon}\right) + \frac{\partial}{\partial \epsilon}\left(\epsilon\left(2 - \frac{K+1}{K}\epsilon\right)\right) = 0$$

$$Q = \mathrm{tr}\left(\boldsymbol{R}^{-1}\right) = \mathrm{tr}\left([R_{ij}]\right)$$

$$R_{ij} = \sum_{c \in \mathbf{C}} \frac{\kappa_{ic}\kappa_{jc}}{\gamma_i \gamma_j H_c(\epsilon)}$$

ここで重要なのは、この式の中にエージェント数 N が含まれていない点である。すなわち、最適 ϵ は N に依存せず決定できることが、この式からわかる。

4 実験による検証

上記で得られた最適 ϵ と N の非依存性を示すために、ある **PG** を用いて学習実験を行い、エージェント数と ϵ のみを変化させてどのように学習誤差が変化するかを調べた。その結果を図 **??**に示す。この図からわかるように、最適 ϵ はエージェント数 N によらず一定であることが確認できる。

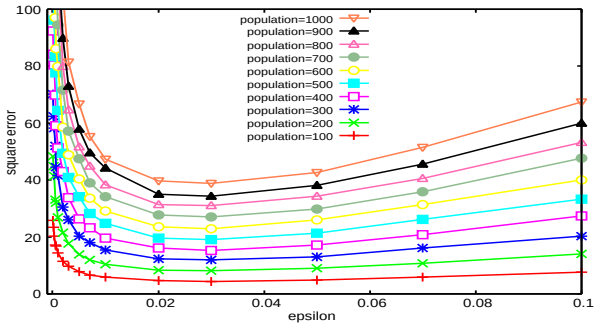

図 1: 報酬が $r_c(d_c) = \gamma_c/d_c$ の時の学習誤差の変化

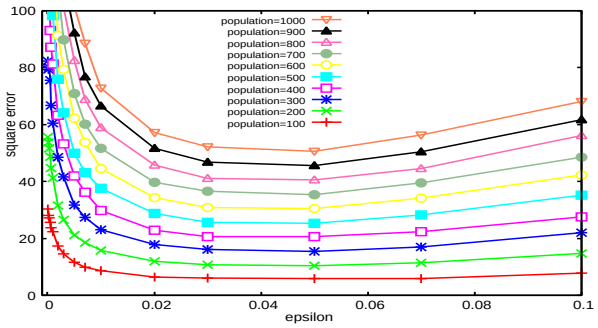

図 2: 報酬が $r_c(d_c) = \sqrt{\gamma_c/d_c}$ の時の学習誤差の変化

5 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、エージェント総数と最適 Exploration 率の関係を調べ、ある条件下で両者の間が非依存であることを理論的に示した。また、それを実験により確認した。

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

[1] Itsuki Noda. Limitations of simultaneous multiagent learning in nonstationary environments. In *Prof. of 2013 IEEE/WIC/ACM International Conference on INtelligent Agent Technology (IAT 2013)*, pages paper–13. IEEE, Nov. 2013.

[2] 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, 9 月 2013.