Robustness of Optimality of Exploration Ratio against Agent Population in Multiagent Learning for Nonstationary Environments

Itsuki Noda National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki, JAPAN
JST CREST, Tokyo Institute of Technology
i.noda@aist.go.jp

Abstract

In this article, I show the robustness of optimality of exploration ratio against the number of agents (agent population) under multiagent learning (MAL) situation in nonstationary environments. Agent population will affect efficiency of agents' learning because each agent's learning causes noisy factors for others. From this point, exploration ratio should be small to make MAL effective. In nonstationary environments, on the other hand, each agent needs explore with enough probability to catch-up changes of the environments. This means the exploration ratio need to be significantly large. I investigate the relation between the population and the efficiency of exploration based on a theorem about relations between the exploration ratio and a lower boundary of learning error. Finally, it is shown that the population of the agents does not affect the optimal exploration ratio under a certain condition. This consequence is confirmed by several experiments using population games with various reward functions.

Introduction

Exploration is an indispensable behavior for learning agents especially under nonstationary environments. The agent needs to explore in a certain ratio (exploration ratio) permanently to catch up changes of the environment. On the other hand, in multi-agent learning (MAL) situation, exploration of an agent causes noise to other agents. So, agents need to keep the exploration ratio as small as possible to help others to learn. So, there is a trade-off problem of choosing the "large-or-small" exploration ratio in MAL under nonstationary environments.

Focusing on real-world problems, we can find several applications of MAL under nonstationary environments. Resource allocations like traffic managements and smart-grid controls are typical problems of such applications. One of difficulties in such applications is *open-ness*, by which the environments may increase or decrease available resources continuously. Also, the population of agents may change over time. In order to handle such open-ness, we need to develop a method to choose suitable behavior parameters of agents like exploration ratio. And, as the first step to establish the method, we need to know relation among such parameters (especially, the exploration ratio), properties of the environments, and learning performance of agents.

Related Works

Choosing and controlling the exploration ratio has been studied mainly for stationary but noisy environments or for single learning agents (Zhang and Pan 2006; Martinez-Cantin et al. 2009; Rejeb, Guessoum, and M'Hallah 2005; Tokic 2010; Reddy and Veloso 2011). The most of these works focused on relation between efficiency of total performance of agents and learning speed in the balance of exploration and exploitation.

No-regret learning also provides a means for agents to learn and reach equilibrium in action-selection for multiagent and probabilistic environments (Gordon, Greenwald, and Marks 2008; Hu and Wellman 1998; Jafari et al. 2001; Greenwald and Jafari 2003). However, the most of these studies assume that the environments are stationary so that learning ends when agents reach equilibrium.

Minority games and its dynamical variations has been studies by (Challet and Zhang 1998; Catteeuw and Manderick 2011). They investigate the case of stationary environments and try to find relations among parameters and agent performances. For nonstationary setting, (Galstyan and Lerman 2002; Galstyan and andKristina Lerman 2003) investigate numerical analysis of behaviors of agents to changing resource capacities.

For MAL and nonstationary environments, (Noda 2013) proposed a formalization based on the concept of *advantageous probabilities*, and derived a theorem about the lower boundary of learning error for a given exploration ratio. In this article, I follow the result of this work, and investigate what factors in MAL will affect the optimal value of the exploration ratio under a kind of resource sharing problem called population games.

Formalization and Theorems

This section will provide a formalization of MAL in nonstationary environments.

Population Game

In this article, we focus on a set of simplified games called *population games* (PGs) in which multiple agents play and learn to make decisions.

In a PG, a large number of agents participate. Each agent selects one of the limited choices and gets a reward on the basis of the choice. The reward is decided only by the number of agents who select the same choice.

Formally, a population game **PG** is defined as follows:

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

where $\mathbf{A} = \{a_1, a_2, \cdots, a_N\}$ is a set of agents, $\mathbf{C} = \{c_1, c_2, \cdots, c_K\}$ is a set of choices, and $\boldsymbol{r} = \{r_a | a \in \mathbf{A}\}$ is a set of reward functions. A reward function $r_a(c; \boldsymbol{d}_{\bar{a}})$ determines the reward for agent a who selects choice c under the distribution of other agents $\boldsymbol{d}_{\bar{a}}$. The distribution $\boldsymbol{d}_{\bar{a}}$ is a vector $[\ d_{\bar{a},c} | c \in \mathbf{C}\ ]$ where $d_{\bar{a},c}$ is the number of other agents who select choice c. Under this definition, each reward function r_a is assumed to return stochastic values. In other words, the environment of the **PG** is stochastic.

Advantageous Probability

Here, *advantageous probability* (AP) $\rho_a(c; \boldsymbol{d}_{\bar{a}})$ for each agent a is introduced to define the probability that choice c will return a larger reward than any other choices under distribution $\boldsymbol{d}_{\bar{a}}$. Formally, AP is defined as follows:

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

where, $\mathcal{P}(\langle \text{condition} \rangle)$ indicates the probability that the '$\langle$condition$\rangle$' holds. Choice $\mathring{c}$ is defined as the *most advantageous choice* of ρ_a when the probability $\rho_a(\mathring{c}_a)$ becomes maximum over all choices in $\mathbf{C}$.

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

It is assumed that each agent cannot know the choices of other agents or their distribution $\boldsymbol{d}_{\bar{a}}$, but can learn the AP by its experiences on the basis of the receiving rewards. A probability function $\tilde{\rho}_a(c)$ indicates *learning AP*, i.e., the probability learned by agent a. Agent a is *exploiting* when a is selecting the most advantageous choice $\mathring{c}$ to learn its AP $\tilde{\rho}_a$, and agent a is *exploring* when a is not selecting $\mathring{c}$.

The *ideal distribution* $\mathring{\boldsymbol{d}}$ is defined as follows:

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

$$\mathring{d}_c \quad : \quad \text{number of agents who are exploiting with choice } c.$$

Similarly, the ideal distribution without agent a is denoted as $\mathring{\boldsymbol{d}}_{\bar{a}}$. Using these definitions, the *ideal AP* for agent a is defined as follows:

$$\mathring{\rho}_a(c) \quad = \quad \rho_a(c; \mathring{\boldsymbol{d}}_{\bar{a}}). \tag{4}$$

Learning and Exploration

Suppose that the purpose of each agent is to select the most advantageous choice, i.e., each agent tries to select a choice that maximizes the probability of obtaining a larger reward than other choices. Therefore the learning goal of each agent is to make its learned AP $\tilde{\rho}_a$ closer to the ideal AP $\mathring{\rho}_a$. If all agents reach $\tilde{\rho}_a = \mathring{\rho}_a$, the **PG** reaches the Nash equilibrium.

The above assumption is slightly different from the conventional formalization of the reinforcement learning, where the agents aim to maximize average reward (AR). AP is introduced instead of AR to avoid scaling and variation issues of reward functions. When reward values are directly handled, as is the case for AR, we need to introduce a framework to classify variations in reward functions. By introducing AP, we can simplify the reward structure as binary (large-or-small) relations of values and can keep the framework simple.

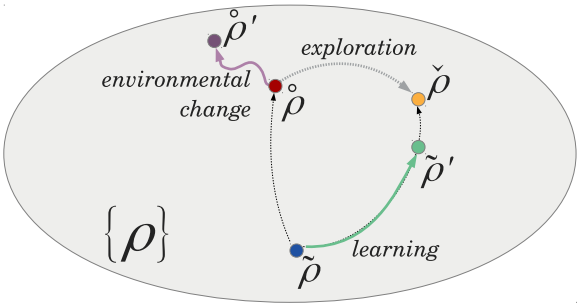

Figure 1: Relations among Ideal, Learning, and Practical Advantageous Probability

To learn successfully, each agent must explore all possible choices. In addition, when a **PG** is nonstationary and reward functions may change over time, agents need to continue to explore the environment beyond the equilibrium point, so that each agent in a **PG** continuously explores within a certain probability.

Because some agents explore simultaneously, the distribution $\boldsymbol{d}$ varies from the ideal distribution $\mathring{\boldsymbol{d}}$. A distribution under agents' exploration is called as the *practical distribution* and denote it as $\check{\boldsymbol{d}}$. Similarly, an explored distribution without agent a is denoted as $\check{\boldsymbol{d}}_{\bar{a}}$. Using these definitions for distribution, the *practical AP* for agent a is defined as follows:

$$\check{\rho}_a(c) \quad = \quad \rho_a(c; \check{\boldsymbol{d}}_{\bar{a}}). \tag{5}$$

Figure 1 illustrate the relationship among ideal AP $\mathring{\rho}$, learning AP $\tilde{\rho}$ and practical AP $\check{\rho}$. At a certain time, $\mathring{\rho}$ is determined by assuming that all agents exploit according to $\tilde{\rho}$. To adjust $\tilde{\rho}$ to $\mathring{\rho}$ through learning, all agents explore possible choices so that the practical AP $\check{\rho}$ separates from $\mathring{\rho}$. Because each agent can acquire a reward according to $\check{\rho}$, $\tilde{\rho}$ moves to $\tilde{\rho}'$ to approximate $\check{\rho}$ by learning. Because of the changes in $\tilde{\rho}$ and the environment during learning, the target AP $\mathring{\rho}$ also move to $\mathring{\rho}'$.

Lower Boundary of Learning Error

(Noda 2013) showed a theorem about relation between lower boundary of learning error and exploration ratio in a certain condition, as described below.

Let's consider the following assumptions:

- Each agent uses the ϵ-greedy policy so that the agent a selects the most advantageous choice $\mathring{c}_a$ with probability $(1 - \epsilon)$ (exploiting mode) and selects one of all possible choices c with probability ϵ (exploring mode). In the exploring mode, assume that all choices are selected with the same probability.

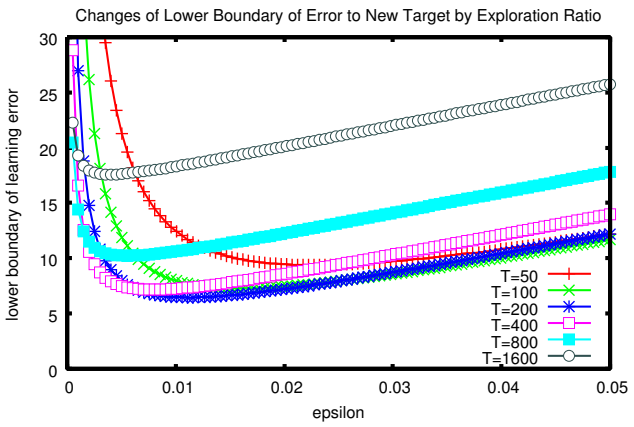

Figure 2: Changes in Lower Boundaries of Error to $\mathring{\boldsymbol{d}}'_{\bar{a}}$ by Exploration Ratio ϵ

- Each agent collects reward information for each choice by performing selections T times according to the above exploration policy. After then, the agent adapts its own learning APs using the reward information.
- The changes in the environment can be modeled as a random walk of $\mathring{\boldsymbol{d}}_{\bar{a}}$ in the parameter space of $\mathring{\rho}_a$, where the variance of each step in the random walk is denoted by σ^2. The original ideal AP at time t is denoted as $\mathring{\rho}_a$ and that at time $t+T$, i.e., after T cycles of random walk, as $\mathring{\rho}'_a$ as shown in figure 1. The parameters to determine these APs are denoted as $\mathring{\boldsymbol{d}}_{\bar{a}}$ and $\mathring{\boldsymbol{d}}'_{\bar{a}}$ respectively.

Under these assumptions, (Noda 2013) showed the following corollary about learning error in this MAL situation:

Corollary 1
The lower boundary of learning error of the above MAL situation is given as the following inequality:

$$\begin{aligned} \mathit{Error} &= E\left[\left\|\mathring{\boldsymbol{d}}'_{\bar{a}} - \tilde{\boldsymbol{d}}'_{\bar{a}}\right\|^2\right] && (6)\\ &\geq T\sigma^2 + \frac{K\tilde{g}_a}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon), && (7) \end{aligned}$$

where $\tilde{g}_a$ is a trail of the inversed Fisher information matrix of the AP ρ_a as follows:

$$\begin{aligned} \tilde{g}_a &= tr\left(\boldsymbol{G}_a\right)\\ \boldsymbol{G}_a^{-1} &= \left[E\left[\frac{\partial \log \rho_a}{\partial d_{\bar{a},i}} \cdot \frac{\partial \log \rho_a}{\partial d_{\bar{a},j}}\right]_{ij}\right] && (8) \end{aligned}$$

□

Figure 2 shows the relationships between the lower boundary and ϵ. Each curve corresponds to changes in the boundary for different values of T. As shown in this graph, there is a positive value for ϵ that minimizes the lower boundary of the squared learning error. However, a significant error remains even if ϵ is at the optimal value.

We can also revise equation (7) for incremental learning so that the agents use the exponential moving average(EMA) $\bar{x}_{t+1} \leftarrow (1-\alpha)\bar{x}_t + \alpha x_t$ to estimate parameters rather than the simple moving average $\bar{x}_{t+\tau} \leftarrow (1/T)\sum_{\tau=1}^{T} x_{t+\tau}$. It is known that EMA can approximate the simple moving average when $T = 2/\alpha - 1$ (Noda 2009b; 2009a). Therefore, we can get the following boundary:

$$\text{Error} \geq \frac{(2-\alpha)\sigma^2}{\alpha} + \frac{\alpha K\tilde{g}_a}{(2-\alpha)\epsilon} + \epsilon N(2 - \frac{K+1}{K}\epsilon) \quad (9)$$

Agent Population and Optimal Exploration Ratio

Based on the boundary relation shown in the previous section, I try to investigate relations among the optimal exploration ratio and parameters of a given PG.

First of all, we assume that the relation between the average error and the exploration ratio forms the same shape of equation (7) or equation (9). Then, define $\mathcal{L}(\epsilon)$ to be the boundary shown as equation (7):

$$\mathcal{L}(\epsilon) = T\sigma^2 + \frac{K\tilde{g}_a}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon). \quad (10)$$

From the definition given by equation (8), $\tilde{g}_a$ can be expanded as follows:

$$\begin{aligned} \tilde{g}_a &= \mathrm{tr}\left(\boldsymbol{G}_a\right)\\ \boldsymbol{G}_a^{-1} &= \left[\left(\sum_{c\in\mathbf{C}} \rho_a(c) \cdot \frac{\partial \log \rho_a(c)}{\partial d_{\bar{a},i}} \cdot \frac{\partial \log \rho_a(c)}{\partial d_{\bar{a},j}}\right)_{ij}\right] \end{aligned} \quad (11)$$

The purpose of the analysis here is to determine the optimal ϵ that makes $\mathcal{L}$ minimum. Because it is hard to solve it directly, we introduce the following assumptions:

- Reward r_c of a choice c is determined according to d_c by the following uniform function ψ with the capacity parameter γ_c:

$$\forall c \quad : \quad r_c(d_c) = \psi\left(\frac{d_c}{\gamma_c}\right), \quad (12)$$

  where, γ_c are positive constants, and ψ is a monotone decreasing function ($\psi' < 0$).
- Under the equilibrium situation of learning of all agents, the reward of each choice c is identical. In other words, if the learning reaches the equilibrium, the agent distribution $\boldsymbol{d}$ makes the reward r_c equal for any choice c. The identical reward is denoted by $\bar{r}$. From equation (12), $\boldsymbol{d}$ and $\bar{r}$ can be calculated as follows:

$$\begin{aligned} \forall c \quad : \quad d_c &= \frac{\gamma_c}{\Gamma}N\\ r_c = \bar{r} &= \psi\left(\frac{N}{\Gamma}\right), \end{aligned}$$

  where Γ is a sum of γ_c, that is, $\Gamma = \sum_c \gamma_c$.
- Under the equilibrium, APs for any agent a are also identical. Therefore,

$$\forall a \forall c \quad : \quad \rho_a(c) = \frac{1}{K}.$$

- The distribution $\boldsymbol{d}$ gets perturbation Δd_c by agents' exploration. Because of the perturbation, actual rewards r_c include noise Δr_c. Here, we suppose that Δd_c is small enough so that Δr_c can be approximated as follows:

$$\begin{aligned}\Delta r_c &= \Delta d_c \cdot \frac{\partial r_c}{\partial d_c} \\ &= \Delta d_c \cdot \psi'\left(\frac{N}{\Gamma}\right) \cdot \frac{1}{\gamma_c} \\ &= \frac{\Delta d_c}{\gamma_c}\bar{\psi}', \end{aligned} \tag{13}$$

where $\bar{\psi}'$ indicates $\psi'\left(\frac{N}{\Gamma}\right)$.

- When the distribution d_c and reward r_c for choice c gets a small perturbation, AP for any choice c' is also affected. Here, we suppose that the degree of change of AP for other choice c', denoted by $\frac{\partial \rho_a(c')}{\partial d_c}$, is in proportion to the probability density of the reward for the choice c' at the average:

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

Under these assumptions, I tried to calculate $\frac{\partial \mathcal{L}}{\partial \epsilon}$ to determine the optimal ϵ.

Here, consider a probability density function, $\mathcal{P}\left(\Delta d_c\right)$, for the perturbation Δd_c, which indicates a probability density of the case where the perturbation of the distribution d_c equals a certain value Δd_c. Because the perturbation is caused by agents' exploration, $\mathcal{P}\left(\Delta d_c\right)$ can be expanded as follows (see Appendix):

$$\begin{aligned}\mathcal{P}\left(\Delta d_c\right) \sim\ & \mathcal{G}(\Delta d_c; \epsilon N(\frac{1}{K} - \frac{\gamma_c}{\Gamma}), \\ & \epsilon N[(\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma})]), \end{aligned} \tag{14}$$

where, $\mathcal{G}(x; \mu, \sigma^2)$ is a Gaussian distribution with average μ and variance σ^2.

Based on equation (13) and equation (14), we can approximate $\mathcal{P}\left(\Delta r_c\right)$ as follows:

$$\begin{aligned}\mathcal{P}\left(\Delta r_c\right) =\ & \mathcal{G}(\Delta r_c; \epsilon N\bar{\psi}'(\frac{1}{K\gamma_c} - \frac{1}{\Gamma}), \\ & \epsilon N\frac{\bar{\psi}'^2}{\gamma_c^2}((\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma}))) \end{aligned}$$

Here, assume that the value of $\mathcal{P}\left(\Delta r_c = 0\right)$ can be approximated by the probability density Δr_c at the average ($r_c = \mathrm{E}\left[r_c\right]$). Then, we can expand $\mathcal{P}\left(\Delta r_c = 0\right)$ as follows:

$$\begin{aligned}\mathcal{P}\left(\Delta r_c = 0\right) &\sim \mathcal{G}(0; 0, \epsilon N\frac{\bar{\psi}'^2}{\gamma_c^2}((\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma}))) \\ &= \frac{1}{\sqrt{2\pi\epsilon N\frac{\bar{\psi}'^2}{\gamma_c^2}((\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma}))}} \end{aligned} \tag{15}$$

We denote the value of equation (15) as λ_c. This λ_c can be simplified by introducing H_c as follows:

$$\begin{aligned}\lambda_c &= \frac{1}{\bar{\psi}'\sqrt{N}\sqrt{H_c(\epsilon)}} \\ H_c(\epsilon) &= \frac{2\pi}{\gamma_c^2} \cdot ((\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma})) \end{aligned}$$

Using these values, (i, j)-th element of Fisher information matrix $\boldsymbol{I} = \boldsymbol{G}^{-1}$ can be calculated as follow (see Appendix):

$$\begin{aligned}\mathcal{I}_{ij} &= \mathrm{E}\left[\frac{\partial}{\partial d_i}\log\rho_a(c) \cdot \frac{\partial}{\partial d_j}\log\rho_a(c)\right] \\ &\propto \frac{K}{N}R_{ij}, \end{aligned} \tag{16}$$

where

$$\begin{aligned}R_{ij} &= \sum_{c\in\mathbf{C}} \frac{\kappa_{ic}\kappa_{jc}}{\gamma_i\gamma_j H_c(\epsilon)} \\ \kappa_{ic} &= \begin{cases} 1 & ;when\ c = i \\ \frac{1}{1-K} & ;when\ c \neq i \end{cases}. \end{aligned}$$

Using a matrix $\boldsymbol{R}$ whose (i, j)-th element is R_{ij}, we can get $\tilde{g}_a$ defined in equation (11) as follows (see Appendix):

$$\tilde{g}_a \propto \frac{N}{K}\mathrm{tr}\left(\boldsymbol{R}^{-1}\right), \tag{17}$$

From equation (10), $\mathcal{L}(\epsilon)$ can be calculated as follows (see Appendix):

$$\mathcal{L}(\epsilon) \propto T\sigma^2 + \frac{NQ}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon), \tag{18}$$

where

$$Q = \mathrm{tr}\left(\boldsymbol{R}^{-1}\right).$$

As a result of above derivations, we can get the following equation from equation (18):

$$\frac{\partial \mathcal{L}}{\partial \epsilon} \propto N\left(\frac{1}{T}\frac{\partial}{\partial \epsilon}\left(\frac{Q}{\epsilon}\right) + \frac{\partial}{\partial \epsilon}\left(\epsilon\left(2 - \frac{K+1}{K}\epsilon\right)\right)\right) \tag{19}$$

The optimal exploration ratio, under which the lower boundary of learning error $\mathcal{L}(\epsilon)$ become minimum, should make $\frac{\partial \mathcal{L}}{\partial \epsilon}$ zero. Therefore, the optimal ratio ϵ^* should satisfy the following equation:

$$\frac{1}{T}\frac{\partial}{\partial \epsilon^*}\left(\frac{Q}{\epsilon^*}\right) + \frac{\partial}{\partial \epsilon^*}\left(\epsilon^*\left(2 - \frac{K+1}{K}\epsilon^*\right)\right) = 0 \tag{20}$$

Unfortunately, the equation is still complex to determine the ϵ^* for given parameters. However, we can find the important relation between the agent population N and the optimal ratio ϵ^*. In equation (20), T and K are independent parameters to N. Also, Q is calculated only from ϵ, K and γ_i. Therefore, equation (20) does not include any factor of N. This means that the agent population N never affect to the optimal ratioϵ^*.

From this implication, we can derive the following pragmatic know-how:

When we can evaluate learning performance with a small number of agents and find the optimal exploration ratio for the problem, we can use the same ratio for the problem with the large number of agents.

Experiments

In order to confirm the implication of the previous section, I conducted experiments using several **PG** described below. The game **PG** is defined as follows:

$$
\begin{aligned}
\mathbf{PG} &= \langle \mathbf{A}, \mathbf{C}, \boldsymbol{r} \rangle \\
\mathbf{A} &= \{a_1, a_2, \cdots, a_{100}\} \\
\mathbf{C} &= \{\mathrm{foo}, \mathrm{bar}, \mathrm{baz}\} \\
\boldsymbol{r} &= \{r_a | \forall a \in \mathbf{A}, \forall c \in \mathbf{C} : r_a(c) = r(c)\} \\
r(c) &= B - (d_c/\gamma_c); \qquad (21) \\
B &= 10.0 \quad : \text{ constant offset} \\
\gamma_c &: \text{ capacity for choice } c \\
&\quad \gamma_{\mathrm{foo}} = 100; \\
&\quad \gamma_{\mathrm{bar}} = 20; \\
&\quad \gamma_{\mathrm{baz}} = 10 \quad \text{at beginning.}
\end{aligned}
$$

Nonstationary-ness is introduced to the game by allowing γ_{baz} to follow a random walk, where its value is changed for every time step. The change is taken from a uniform distribution in $[-0.01, 0.01]$.

Each agent has its own reward table that indicates an expected reward for each choice. In every cycle, each agent selects the best choice (in exploitation) or another possible choice (in exploration) on the basis of its own reward table. When the agent gets an actual reward because of its choice, the agent updates its table. In this experiment, we suppose that each agent applies ϵ-greedy policy for action selection.

In the first experiments, I changed the total population of agent. In the experiment, the agent populations are set from 100 to 1000. I run 10 times for each setting, and calculate the average error defined by equation (6). Figure 3 shows the result of the experiment. In this graph, the horizontal and vertical axes are exploration ratio ϵ and the average error. Each line corresponds to each agent population ($100 \sim 1000$). We can find that each line is scaled by the agent population and form similar shapes. The more important point of this result is that the optimal ϵ that makes the each error curve minimum is never changed by the agent population. In each line in figure 3, the error hits the bottom around $\epsilon = 0.02$. This supports the implication of the previous section.

I also conducted two more experiments using different reward functions in the same **PG** . Instead of equation (21), the following two functions are used:

$$r_b(c) = \frac{\gamma_c}{d_c} \qquad (22)$$

$$r_c(c) = \sqrt{\frac{\gamma_c}{d_c}} \qquad (23)$$

Figure 4 and 5 are the results of the experiments using the reward function of equation (22) and 23, respectively. Both graphs show the similar changes of errors as shown in figure 3. In the case of figure 4, each line hits the bottom around $\epsilon = 0.03$. In the case of figure 5, each line hits the bottom around $\epsilon = 0.05$. These results also support the above implication, that is,

the agent population N never affect to the optimal ϵ.

Discussion

In the analytic derivation of the relation between agent population N and the optimal exploration ratio ϵ^*, I introduced several assumptions. Here, I discuss about adequacy of the assumptions.

First of all, we assume that the average error curve form the same shape of its lower boundary. Also, we use AP as the learning target in the derivation, while the actual learning in the experiments tries to maximize AR. These assumptions are somewhat strong. Fortunately, however, the actual error curves shown in figure 2 is quite similar to the actual results shown in figure 3, 4 and 5. So, we will apply the results analytic derivation as a general investigation of actual learning phenomena.

We also assume that the reward function is uniform with capacity parameters. This seems reasonable in the most of the resource sharing problem, because such definitions of reward function can cover the situation of resource sharing widely. We will also be able to relax and generalize this condition by the further investigation.

From the viewpoint of prior coordinations in multiagent learning, the results of previous sections tell an interesting feature of MAL. The results say that the optimal exploration ratio is stable when the agent population increases. Therefore, as mentioned in section , we can start the learning with a small number of agents to determine the optimal exploration ratio, and increase the number of agents with the same ratio. Another way to utilize the results is that, we start the online learning with the fixed number of agents to find the optimal ratio, and make the learning system open for agents to join the system with restricting them to use the same exploration ratio.

Conclusion

In this article, I investigated what factors in MAL will affect the optimal value of the exploration ratio for a subset of population game. The investigation implies the optimal exploration ratio can be determined independent of the total population of the agents. This feature is confirmed by several experiments of MAL for a certain kind of population games with various reward functions. Using the feature, we can know that it is reasonable to use the same exploration ratio for MAL with the large population of agents when the ratio can be confirmed to be optimal for the game with the small population of agents.

There are several further issues of this work. We might be able to find another relation among the exploration ratio and other parameters like the number of resources K, learning speed (stepsize) α, or nonstationary factor σ^2. Also, there are several weaknesses in the derivations of the relation between agent population and optimal exploration ratio, for example, a strong assumption that the average error is equal to its lower boundary given by the corollarie.

Appendix: Derivations

Derivation of Equation (14)

The perturbation Δd_c can be divided into two factors, decreasing factor caused by exploration of agents who consider choice c is the best, and increasing factor caused by exploration of agents who consider other choice c' is the best. When we denote the probability densities of the both factors as $\mathcal{P}\left(\Delta^- d_c\right)$ and $\mathcal{P}\left(\Delta^+ d_c\right)$, respectively, $\mathcal{P}\left(\Delta d_c\right)$ can be expanded as follows:

$$
\begin{aligned}
\mathcal{P}\left(\Delta^- d_c\right) &= \left\langle \begin{array}{l} \textit{probability of the number of agents in} \\ \frac{\gamma_c N}{\Gamma} \textit{ agents do not choose } c \textit{ with the} \\ \textit{probability } \epsilon \end{array} \right\rangle \\
&= \mathcal{B}(-\Delta d_c; \epsilon, \frac{\gamma_c N}{\Gamma}) \\
&\sim \mathcal{G}(\Delta d_c; -\frac{\gamma_c N \epsilon}{\Gamma}, \frac{\gamma_c N}{\Gamma} \cdot \epsilon(1-\epsilon)) \\
\mathcal{P}\left(\Delta^+ d_c\right) &= \left\langle \begin{array}{l} \textit{probability of the number of agents in} \\ N \textit{ agents choose } c \textit{ with the probabil-} \\ \textit{ity } \frac{\epsilon}{K} \end{array} \right\rangle \\
&= \mathcal{B}(\Delta d_c; \frac{\epsilon}{K}, N) \\
&\sim \mathcal{G}(\Delta d_c; \frac{N\epsilon}{K}, N\frac{\epsilon}{K}(1-\frac{\epsilon}{K})) \\
\mathcal{P}\left(\Delta d_c\right) &= \mathcal{P}\left(\Delta^- d_c\right) * \mathcal{P}\left(\Delta^+ d_c\right) \\
&\sim \mathcal{G}(\Delta d_c; \epsilon N(\frac{1}{K} - \frac{\gamma_c}{\Gamma}), \\
&\qquad \epsilon N[(\frac{1}{K} + \frac{\gamma_c}{\Gamma}) - \epsilon(\frac{1}{K^2} + \frac{\gamma_c}{\Gamma})]),
\end{aligned}
$$

where, $B(x; p, n)$ is a Binomial distribution with success probability in each trial p and number of trials n, and $G(x; \mu, \sigma^2)$ is a Gaussian distribution with average μ and variance σ^2.

Derivation of Equation (16)

$$
\begin{aligned}
\mathcal{I}_{ij} &= \mathrm{E}\left[\frac{\partial}{\partial d_i}\log\rho_a(c) \cdot \frac{\partial}{\partial d_j}\log\rho_a(c)\right] \\
&= \mathrm{E}\left[\frac{1}{\rho_a(c)}\frac{\partial\rho_a(c)}{\partial d_i} \cdot \frac{1}{\rho_a(c)}\frac{\partial\rho_a(c)}{\partial d_j}\right] \\
&\propto \sum_c \rho_a(c)\left(\frac{1}{\rho_a^2(c)} \cdot \left(\frac{\bar{\psi}'}{\gamma_i}\right)\kappa_{ic}\lambda_c\left(\frac{\bar{\psi}'}{\gamma_j}\right)\kappa_{jc}\lambda_c\right) \\
&= \sum_c K\bar{\psi}'^2\lambda_c^2\frac{\kappa_{ic}\kappa_{jc}}{\gamma_i\gamma_j} \\
&= \sum_c K\bar{\psi}'^2\frac{1}{\bar{\psi}'^2 N H_c(\epsilon)}\frac{\kappa_{ic}\kappa_{jc}}{\gamma_i\gamma_j} \\
&= \frac{K}{N}\sum_c \frac{\kappa_{ic}\kappa_{jc}}{\gamma_i\gamma_j H_c(\epsilon)} \\
&= \frac{K}{N}R_{ij}.
\end{aligned}
$$

Derivation of Equation (17)

$$
\begin{aligned}
\tilde{g}_a &= \mathrm{tr}\left(\boldsymbol{G}\right) \\
&= \mathrm{tr}\left(\boldsymbol{I}^{-1}\right) \\
&\propto \mathrm{tr}\left(\left[\frac{K}{N}\boldsymbol{R}\right]^{-1}\right) \\
&= \frac{N}{K}\mathrm{tr}\left(\boldsymbol{R}^{-1}\right).
\end{aligned}
$$

Derivation of Equation (18)

$$
\begin{aligned}
\mathcal{L}(\epsilon) &\propto T\sigma^2 + \frac{K\tilde{g}_a}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon) \\
&= T\sigma^2 + \frac{KNQ}{\epsilon TK} + \epsilon N(2 - \frac{K+1}{K}\epsilon) \\
&= T\sigma^2 + \frac{NQ}{\epsilon T} + \epsilon N(2 - \frac{K+1}{K}\epsilon).
\end{aligned}
$$

Acknowledgments

Prof. Toshiharu Sugawara gave an important hint for this work. This work was supported by JSPS KAKENHI 24300064.

References

Catteeuw, D., and Manderick, B. 2011. Heterogeneous populations of learning agents in the minority game. In *Adaptive and Learning Agents, LNCS, 7113,*, 100–113. Springer.

Challet, D., and Zhang, Y.-C. 1998. On the minority game: Analytical and numerical studies. *Physica A: Statistical and Theoretical Physics* 256(3–4):514–532.

Galstyan, A., and andKristina Lerman, S. K. 2003. Resource allocation games with changing resource capacities. In *Proc. of the 2nd Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, 145–152. ACM.

Galstyan, A., and Lerman, K. 2002. Adaptive boolean networks and minority games with time-dependent capacities. *Physical Review E* 66(015103).

Gordon, G. J.; Greenwald, A.; and Marks, C. 2008. No-regret learning in convex games. In Cohen, W. W.; McCallum, A.; and Roweis, S. T., eds., *ICML*, volume 307 of *ACM International Conference Proceeding Series*, 360–367. ACM.

Greenwald, A. R., and Jafari, A. 2003. A general class of no-regret learning algorithms and game-theoretic equilibria. In *COLT'03*, 2–12.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik, J. W., ed., *ICML*, 242–250. Morgan Kaufmann.

Jafari, A.; Greenwald, A.; Gondek, D.; and Ercal, G. 2001. On no-regret learning, fictitious play, and nash equilibrium. In *In Proceedings of the Eighteenth International Conference on Machine Learning*, 226–233. Springer.

Martinez-Cantin, R.; de Freitas, N.; Brochu, E.; Castellanos, J. A.; and Doucet, A. 2009. A bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Auton. Robots* 93–103.

Noda, I. 2009a. Adaptation of stepsize parameter for non-stationary environments by recursive exponential moving average. In *Prof. of ECML 2009 LNIID Workshop*, 24–31. ECML.

Noda, I. 2009b. Recursive adaptation of stepsize parameter for unstable environments. In Taylor, M., and Tuyls, K., eds., *Proc. of ALA-2009*, Paper–14.

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

Reddy, P. P., and Veloso, M. M. 2011. Learned behaviors of multiple autonomous agents in smart grid markets. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAA).

Rejeb, L.; Guessoum, Z.; and M'Hallah, R. 2005. The exploration-exploitation dilemma for adaptive agents. In *Proceedings of the Fifth European Workshop on Adaptive Agents and Multi-Agent Systems*.

Tokic, M. 2010. 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.

Zhang, K., and Pan, W. 2006. The two facets of the exploration-exploitation dilemma. In *Proceedings of the IEEE/WIC/ACM international conference on Intelligent Agent Technology (IAT-06)*, 371–380. Washington, DC, USA: IEEE Computer Society.

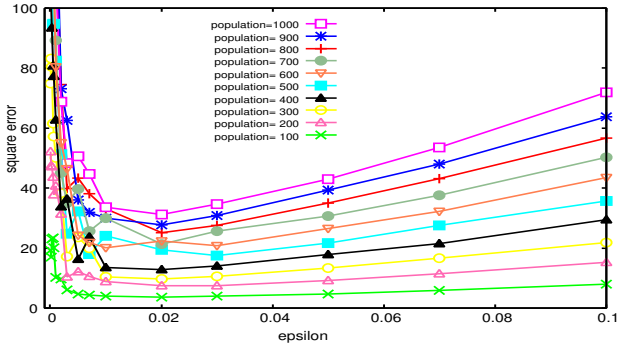

Figure 3: Changes of Average Learning Error in the case of reward function $r_c(d_c) = B - (d_c/\gamma_c)$

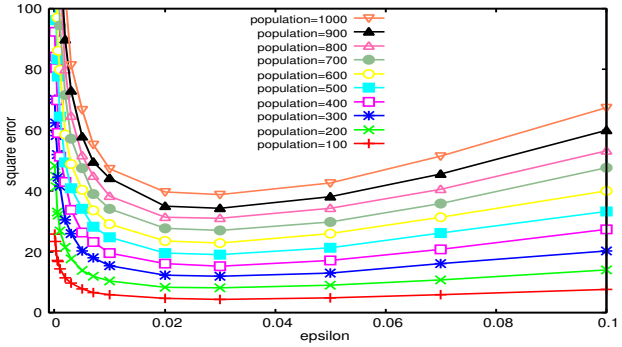

Figure 4: Changes of Average Learning Error in the case of reward function $r_c(d_c) = \gamma_c/d_c$

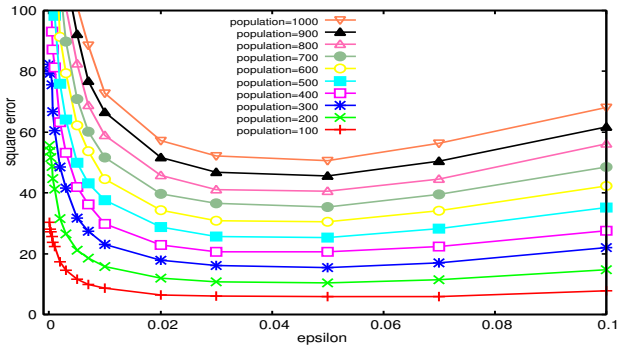

Figure 5: Changes of Average Learning Error in the case of reward function $r_c(d_c) = \sqrt{\gamma_c/d_c}$