A Parameter Investigation of the ϵ-greedy Exploration Ratio Adaptation Method in Multi-agent Reinforcement Learning

Takuya Okano okano565656@gmail.com
Masaki Onishi onishi@ni.aist.go.jp
Department of Policy and Planning Sciences
University of Tsukuba
Tsukuba, Ibaraki 305-0577, Japan

Artificial Intelligence Research Center
National Institute of Advanced Industrial Science and Technology
Tsukuba, Ibaraki 305-8560, Japan

Itsuki Noda i.noda@aist.go.jp
Artificial Intelligence Research Center
National Institute of Advanced Industrial Science and Technology
Tsukuba, Ibaraki 305-8560, Japan

Abstract

Balancing the amounts of exploration and exploitation is one of the most important issues in reinforcement learning, which is also important while attempting to achieve favorable performances in multi-agent and non-stationary environments. In a previous study, we proposed the ϵ-greedy exploration ratio ϵ adaptation method, which is referred to as “Win or Update Exploration ratio”(WoUE). This method is capable of adapting ϵ in multi-agent and non-stationary environments. However, while using this method, we have only one attempt to set the parameter that determines ϵ's update amount. In this study, we investigate the manner in which this parameter affects the WoUE's performance. Thus, we confirm that there is an optimal value for the WoUE's performance, and we further evaluate WoUE using the related adaptation method. We confirm that WoUE method improves the performance as compared with that obtained using the related methods regardless of the WoUE's parameter in multi-agent and non-stationary environments.

Keywords: Reinforcement Learning, Multi-agent Reinforcement Learning, Exploration Ratio, ϵ-greedy

1. Introduction

Exploration ratio adaptation in ϵ-greedy is an important issue in reinforcement learning (Sutton and Barto, 1998). Several studies have proposed various adaptation methods, which only target single agent reinforcement learning and/or stationary environments. However, several difficulties arise while attempting to adapt these methods to scenarios with multi-agent and non-stationary environments. Therefore, we proposed exploration ratio adaptation method WoUE which adapts the exploration ratio in multi-agent and non-stationary environments. To function, this method requires a parameter that determines the exploration ratio update amount at one time. In this study, we investigate the influence that is exhibited by various parameters on the system performance, and we evaluate WoUE and the related adaptation method of exploration ratio using each parameter.

2. Background

2.1 Domain

To evaluate the exploration ratio adaptation methods, we use the non-stationary "Repeated Resource Sharing Problem" (RRSP)(Noda and Ohta, 2008).

RRSP is a congestion game where agents share their resources. Mathematically, this RRSP can be represented by

$$\begin{aligned} RRSP &= < G, C, A, R >, \\ G &= \{g_1, g_2, \ldots, g_M\}, \\ C &= \{c_{g_1}, c_{g_2}, \ldots, c_{g_M}\}, \\ A &= \{a_1, a_2, \ldots, a_N\}, \end{aligned}$$

where G represents the set of resources, C is the set of resource capacities, A is the set of agents, M is the number of resources, and N is the number of agents. In addition, n_j, the number of agents that share the same resource ($j \in G$), determines the reward function, $R(n_j, c_j)$, which is a monotonically decreasing function w.r.t. n_j/c_j. Each agent selects a resource j according to ϵ-greedy and gets a reward $R(n_j, c_j)$ by the selecting resource, and learns by reinforcement learning throughout the process, i.e., updates the expected reward $Q(j)$ by $Q(j) \leftarrow (1-\alpha)Q(j) + \alpha R(n_j, c_j)$, where α is a learning ratio. This process is repeated during every step. The global optimal condition in the game is defined as the average of the distribution of all agent rewards is high value and the variance is minimum value. In other words, the global optimal condition is the capacity composition ratio of resource $j \in G$ and the number of agents, who select the resource j, is equal to a specific condition($\frac{c_j}{\sum_{k \in G} c_k} = \frac{n_j}{N} : \forall j \in G$). Therefore, we define the system performance, i.e., $Error$, using the following equation:

$$Error \triangleq \sum_{j \in G} (\frac{c_j}{\sum_{k \in G} c_k} - \frac{n_j}{N})^2. \tag{1}$$

Here, the objective in the game is to minimize the $Error$ by multi-agent reinforcement learning as well as to evaluate the system performance using the $Error$.

We also introduce non-stationary environments to RRSP to reflect real-world situations. The capacity of each resource may change over time by a certain probability. We simulate numerous real-world problems using the non-stationary RRSP. On applying the congestion problem to RRSP, the road represents the resource and the car is an agent. We assume that the change in the resource capacity is an impassable or blocked road, consequently, RRSP creates a new road. Therefore, we use a non-stationary RRSP that resource's capacity change over time by a certain probability to evaluate the multi-agent reinforcement learning performance. Figure 1 shows the image diagram of the non-stationary RRSP.

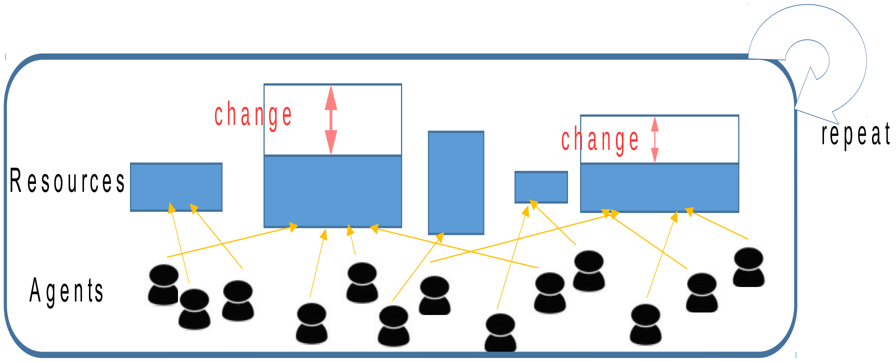

Figure 1: Diagram of the non-stationary RRSP.

2.2 Exploration Ratio Adaptation in Multi-agent and Non-stationary Environments

The exploration ratio ϵ in ϵ-greedy is a parameter used to determine the balance between exploration and exploitation(Sutton and Barto, 1998). The ratio is a parameter that affects the learning performance as well as multi-agent reinforcement learning performance.

It is well known that this ratio has a positive value but should eventually converge to zero in a stationary environment. In non-stationary environments, however, this ratio should not converge to zero. Because the environment changes continuously over time, this ratio should be kept positive to follow the changes of the environment.

Moreover, in multi-agent environments, the agents influence each other. Therefore, the exploration ratio is also considered as a parameter that indicates the level of influence of an agent on the other agents. As a result, obtaining a suitable value for this parameter is important in order to improve performance of multi-agent reinforcement learning, which is a difficult yet essential issue.

Several previous studies have explored and proposed various methods to adjust the exploration ratio. However, these studies only targeted single agent reinforcement learning and/or stationary environments (Tokic, 2010; Tokic and Palm, 2011; Auer et al., 2002; Hester et al., 2013; Singh et al., 2000).

3. Win or Update Exploration Ratio

In our previous study, we proposed "Win or Update Exploration ratio"(WoUE) method (Okano and Noda, 2017), which is an ϵ-greedy exploration ratio ϵ adaptation method in multi-agent and non-stationary environments.

In WoUE, agent i who receives a lower reward compared with the average reward of all agents(total average) updates its ϵ_i, with the objective of reaching equilibrium. The ratios are updated to minimize the variance of the distribution of all agents' rewards during a certain period according to the following equation:

$$\epsilon_i \leftarrow \begin{cases} \epsilon_i + \eta(\mu_R - R_i)(R_{\epsilon_i} - R_{(1-\epsilon_i)}) & (\mu_R > R_i) \\ \epsilon_i & (\mu_R \leq R_i), \end{cases} \tag{2}$$

where μ_R is the average reward for all agents respectively, ϵ_i is the exploration ratio of the agent i for a certain period, R_i is the average reward that the agent i receives associated with the exploration ratio ϵ_i for a certain period. $R_{(1-\epsilon_i)}$ is the average reward derived from exploitation, and R_{ϵ_i} is the average reward derived from exploration. The update amount is given by the following equation:

$$(\mu_R - R_i)(R_{\epsilon_i} - R_{(1-\epsilon_i)}) \propto -\frac{\partial \sigma_R^2}{\partial \epsilon_i}. \tag{3}$$

Therefore, we observe that agent i updates its exploration ratio, ϵ_i, in an attempt to minimize the variance of the distribution of all agents' rewards σ_R^2. In other words, the agent i updates ϵ_i to reach equilibrium. Because only the agents whose rewards are lower than total average, it can be expected to avoid falling into a sub-optimal equilibrium where the average reward is low. The derivation of Equation (3) is shown in Appendix A. Since each agent only requires μ_R, this method is capable of updating ϵ in such a manner that minimizes σ_R^2 by only broadcasting μ_R to all agents. This method is easily adapted if there are many agents in the environment.

In this study, we use a different version of WoUE (Okano and Noda, 2017), which updates ϵ at each step in order to increase the opportunity to update ϵ. To update at each step, the parameters μ_R, R_i, $R_{(1-\epsilon_i)}$, and R_{ϵ_i} are modified with Exponential Moving Average. WoUE algorithm is shown in Appendix B.

4. Investigation of WoUE's Parameter η

WoUE initializes the parameter η that modifies ϵ from Equation (2). We investigated the effects of η on the *Error*(i.e., the system performance) via experimentation. The results indicate that there is an optimal η. When η is high, WoUE can achieve a good performance.

We investigate changes of the *Error* for each η in the non-stationary RRSP. In this experiment, we compare the *Error* with each η. Figure 2 shows the *Error* of WoUE for each η. Each line represents a different step number (the update time of the exploration ratio ϵ). The results indicate that WoUE *Error* curves over each η have convex downward shapes, which confirms that there is an optimal η parameter which minimizes the *Error*. When both η and the step number are small, the system performance is poor. Figure 3 shows the changes in ϵ with respect to WoUE with a small(left) and a large(right) value of η. Figure 3 confirms that updating ϵ is quite slow. Poor performance due to the insufficient

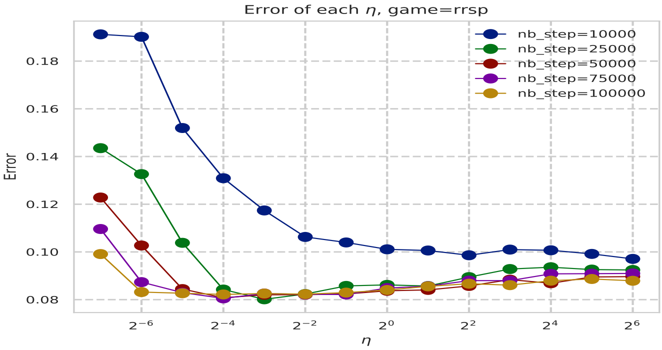

Figure 2: *Error*(i.e., the system performance) of WoUE associated with each η.

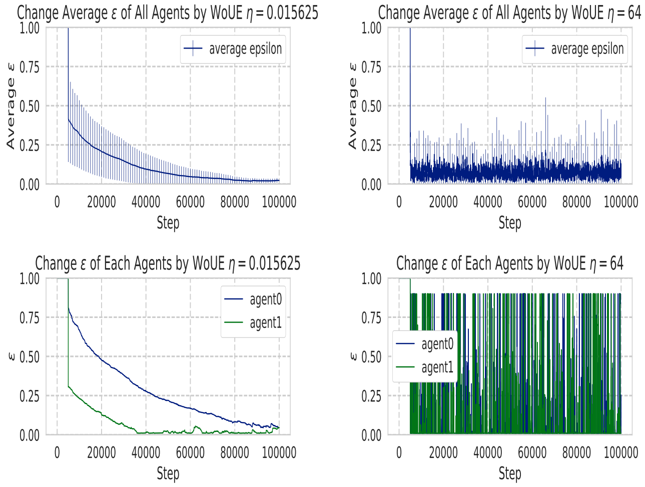

Figure 3: Change in ϵ by WoUE with $\eta = 2^{-6}$ (left) and $\eta = 2^{6}$ (right).The upper graphs are the average ϵ for all agents and the lower graphs are ϵ for the two sampled agents.

amount of time taken in this experiment to update ϵ. If there is a sufficient amount of the update time, WoUE performs well. Although this method produces small $Error$ quickly, when η is large, the $Error$ is bigger than the minimum value. Figure 3 indicates that ϵ changes rapidly, due to the large value of η, (ϵ transforms from a minimum ϵ to a maximum ϵ, or from a maximum to a minimum). In addition, both the exploration noise and the $Error$ increase.

Thus, if there is a sufficient amount of time to adapt ϵ, η should be set to a small value. When there is insufficient amount of time, η should be set to a large value.

5. Comparison with Related Work

We compared WoUE with other methods in a non-stationary RRSP scenario. Based on these comparisons, we conformed that WoUE achieve a better performance than the other methods regardless of the various η values in non-stationary environments.

5.1 Comparative Method

As comparative methods, we chose a normal ϵ-greedy and "Value-Difference Based Exploration"(VDBE) method (Tokic, 2010). VDBE modifies ϵ in ϵ-greedy method based on temporal-difference errors. Even though VDBE is used as an adaptation method for ϵ in single agent environments, this method is easily adapted to multi-agent environments. VDBE has two parameters σ and δ. Based on the results from (Tokic, 2010), we set $\delta = \frac{1}{M}$. In this experiment, we use a variety of σ values and evaluate the normal ϵ-greedy method without modifying ϵ.

5.2 Experiment

We evaluate these methods using a non-stationary RRSP. This experiment comprises several agents ($N = 100$), many resources ($M = 5$), the reward function ($R(n_j, c_j) = \frac{c_j}{1+n_j}$), and the original resource capacities ($C = \{1, 2, 4, 8, 16\}$). Each capacity of resource doubles when a capacity is original and it restores to the original value when a capacity is already doubled by probability of 0.001 within each step. The parameters of each method are described in Appendix C. The learning ratio α for all agents is $\alpha = 0.3$. In these experiments, all agents act randomly for 5% of the steps such that all agents learn the environment.

Method and Parameter	Average $Error$(SD)	Average ϵ(SD)
$\epsilon = 0.01$	0.104(0.035)	0.010(0.000)
$\epsilon = 0.1$	0.114(0.019)	0.100(0.000)
VDBE $\sigma = 0.001$	0.569(0.069)	0.788(0.032)
VDBE $\sigma = 0.01$	0.160(0.024)	0.246(0.058)
VDBE $\sigma = 0.1$	0.156(0.058)	0.001(0.003)
VDBE $\sigma = 1$	0.233(0.079)	0.000(0.000)
WoUE $\eta = 0.015625$	0.109(0.021)	0.090(0.030)
WoUE $\eta = 1$	**0.095(0.023)**	0.029(0.009)
WoUE $\eta = 64$	0.096(0.023)	0.046(0.028)

Table 1: Average $Error$ and average ϵ of all agents for each method in the non-stationary RRSP. SD is the standard deviation for $Error$ and ϵ.

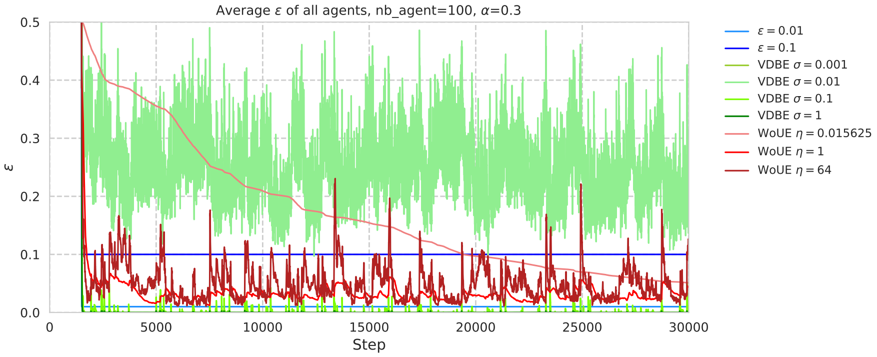

Figure 4: Changes in average ϵ of all agents for each method in non-stationary RRSP. VDBE with $\sigma = 0.001$ is out of this figure because the ϵ is very high.

Table 1 shows the $Error$(i.e., the system performance) and the average ϵ for all agents for each adaptation method in the latter half of all steps. The VDBE's $Error$ varies for each σ value and the $Error$ is worse compared with the other methods, i.e., the VDBE's lowest $Error$ is higher than the WoUE's highest $Error$. On the other hand, the results indicate that WoUE has better performance than the other methods. Even though the VDBE's performance varies for each parameter, the WoUE's performance is favorable regardless of the parameters.

Figure 4 shows the change in the average ϵ for all agents. Here, the X-axis is the step number, and the Y-axis is the average ϵ for all agents. ϵ adapted by VDBE varies according to a specific parameter. This parameter affects the system performance sensitivity. Even though the SD of ϵ adapted by WoUE varies, the average ratio is similar. In another parameters, we confirmed the same result in Appendix D.

If WoUE's η is small, the system performance is lacking due to a lowly changing ϵ. However if we continually update ϵ, system performance will improve.

6. Conclusion and Future work

In this study, we investigated the parameter η in WoUE which is adaptation method of ϵ in ϵ-greedy and compared the results with recent studies in multi-agent and non-stationary environments. The WoUE's parameter η determines ϵ's the update amount. We conducted several experiments to investigate the effects of this parameter. Our results indicate that there are optimal η values. We confirmed that when η is small, WoUE achieves a better performance despite long simulation periods. When η is high, even though WoUE achieves a better performance quicker, the system performance is worse than the best performance, due to the exploration noise. We evaluated WoUE with recent studies on VDBE with various parameters. Results indicate that WoUE performs more favorably than VDBE in multi-agent and non-stationary environments regardless of the parameters used.

The evaluation of WoUE in real world applications and its theoretical analysis are topics of future investigations.

Acknowledgments

This presentation is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and JSPS KAKENHI Grant Number 18H03301.

Appendix A.

We further derive Equation(3). First, we express the average reward of the agent i R_i as

$$R_i \approx (1-\epsilon_i)R_{(1-\epsilon_i)} + \epsilon_i R_{\epsilon_i}. \tag{4}$$

Equation(4) can be derived from the following set of equations:

$$\begin{aligned}
R_i &= \frac{1}{T}\sum^{T} r_t && (5)\\
&= \frac{1}{T}(\sum^{T_{(1-\epsilon_i)}} r_{(1-\epsilon_i)} + \sum^{T_{\epsilon_i}} r_{\epsilon_i}) && (6)\\
&= \frac{1}{T}(\frac{T_{(1-\epsilon_i)}}{T_{(1-\epsilon_i)}}\sum^{T_{(1-\epsilon_i)}} r_{(1-\epsilon_i)} + \frac{T_{\epsilon_i}}{T_{\epsilon_i}}\sum^{T_{\epsilon_i}} r_{\epsilon_i}) && (7)\\
&= \frac{T_{(1-\epsilon_i)}}{T}\frac{1}{T_{(1-\epsilon_i)}}\sum^{T_{(1-\epsilon_i)}} r_{(1-\epsilon_i)} + \frac{T_{\epsilon_i}}{T}\frac{1}{T_{\epsilon_i}}\sum^{T_{\epsilon_i}} r_{\epsilon_i} && (8)\\
&\approx (1-\epsilon_i)\frac{1}{T_{(1-\epsilon_i)}}\sum^{T_{(1-\epsilon_i)}} r_{(1-\epsilon_i)} + \epsilon_i\frac{1}{T_{\epsilon_i}}\sum^{T_{\epsilon_i}} r_{\epsilon_i} && (9)\\
&= (1-\epsilon_i)R_{(1-\epsilon_i)} + \epsilon_i R_{\epsilon_i}, && (10)
\end{aligned}$$

$T_{1-\epsilon_i}$ represents the exploitation number and T_{ϵ_i} is the exploration number by agent i. $r_{(1-\epsilon_i)}$ is the reward from exploitation and r_{ϵ_i} is the reward from exploration.
Next we derive Equation(3). The variance σ_R^2 is calculated from:

$$\sigma_R^2 = \frac{1}{N}\sum_{i=0}^{N}(\mu_R - R_i)^2. \tag{11}$$

We differentiate σ_R^2 with respect to ϵ_i

$$\begin{aligned}
\frac{\partial\sigma_R^2}{\partial\epsilon_i} &\approx \frac{2}{N}(\mu_R - R_i)\frac{\partial(\mu_R - R_i)}{\partial\epsilon_i} && (12)\\
&\propto \frac{2}{N}(\mu_R - R_i)(R_{(1-\epsilon_i)} - R_{\epsilon_i}) && (13)\\
&\propto (\mu_R - R_i)(R_{(1-\epsilon_i)} - R_{\epsilon_i}) && (14)\\
-\frac{\partial\sigma_R^2}{\partial\epsilon_i} &\propto (\mu_R - R_i)(R_{\epsilon_i} - R_{(1-\epsilon_i)}). && (15)
\end{aligned}$$

Therefore, we arrive at Equation(3):

$$(\mu_R - R_i)(R_{\epsilon_i} - R_{(1-\epsilon_i)}) \quad\propto\quad -\frac{\partial\sigma_R^2}{\partial\epsilon_i}. \tag{16}$$

Appendix B.

Algorithm 1 Multi-agent Reinforcement Learning with WoUE

```
Initialze N agents
for step $t = 0$ to T do
    for agent $i = 0$ to N do
        Select action, using $\epsilon$-greedy
    end for
    for agent $i = 0$ to N do
        Get reward $r_i$
        Update average reward $R_i \leftarrow \alpha r_i + (1 - \alpha) R_i$
        Update $Q_i(a)$
    end for
    Compute average reward of all agents $\mu_{R,t} = \frac{1}{N} \sum_{i=0}^{N} r_i$
    Update average reward of all agents $\mu_R \leftarrow \alpha \mu_{R,t} + (1 - \alpha) \mu_R$
    Broadcast $\mu_R$ to all agents
    for agent $i = 0$ to N do
        if agent $i$ exploited then
            Update exploited reward $R_{(1-\epsilon_i)} \leftarrow \epsilon_i r_i + (1 - \epsilon_i) R_{(1-\epsilon_i)}$
        else if agent $i$ explored then
            Update explored reward $R_{\epsilon_i} \leftarrow (1 - \epsilon_i) r_i + \epsilon_i R_{\epsilon_i}$
        end if
        if $R_i < \mu_R$ then
            $\epsilon_i \leftarrow \epsilon_i + \eta(\mu_R - R_i)(R_{\epsilon_i} - R_{(1-\epsilon_i)})$
            $\epsilon_i \leftarrow$ checkMinMax$(\epsilon_i)$
        end if
    end for
end for

function checkMinMax$(\epsilon)$
    if $\epsilon < minimum\_\epsilon$ then
        $\epsilon \leftarrow minimum\_\epsilon$
    else if $maximum\_\epsilon < \epsilon$ then
        $\epsilon \leftarrow maximum\_\epsilon$
    end if
    return $\epsilon$
end function
```

Appendix C.

Name	**Parameter details**
WoUE	$\eta = \{0.0125, 1, 245\}$, $minimum\_\epsilon = 0.01$, $maximum\_\epsilon = 0.9$
VDBE	σ=$\{$0.001, 0.01, 0.1, 1 $\}$, $\delta = \frac{1}{M}$

Table 2: Adaptation Methods for ϵ and their parameters.

Appendix D.

Herein, we evaluated on the adaptation methods with other settings. This experiment comprises a number of agents ($N = 300$), a number of resources ($M = 8$), and the original resource capacities ($C = \{3, 6, 9, 12, 15, 18, 21, 24\}$). The learning ratio for all agents is $\alpha = 0.1$. Results are shown in Table 3 and Figure 5 . We confirmed that these results are similar to the result in Section 5.

Method and Parameter	Average $Error$(SD)	Average ϵ(SD)
$\epsilon = 0.01$	0.112(0.057)	0.010(0.000)
$\epsilon = 0.1$	0.091(0.034)	0.100(0.000)
VDBE $\sigma = 0.001$	0.126(0.025)	0.280(0.056)
VDBE $\sigma = 0.01$	0.170(0.067)	0.004(0.006)
VDBE $\sigma = 0.1$	0.247(0.041)	0.000(0.000)
VDBE $\sigma = 1$	0.300(0.051)	0.000(0.000)
WoUE $\eta = 0.015625$	0.104(0.024)	0.207(0.021)
WoUE $\eta = 1$	0.075(0.030)	0.063(0.016)
WoUE $\eta = 64$	**0.071**(0.032)	0.077(0.037)

Table 3: Average $Error$ and average ϵ of all agents for each method in the non-stationary RRSP. SD is the standard deviation for $Error$ and ϵ.

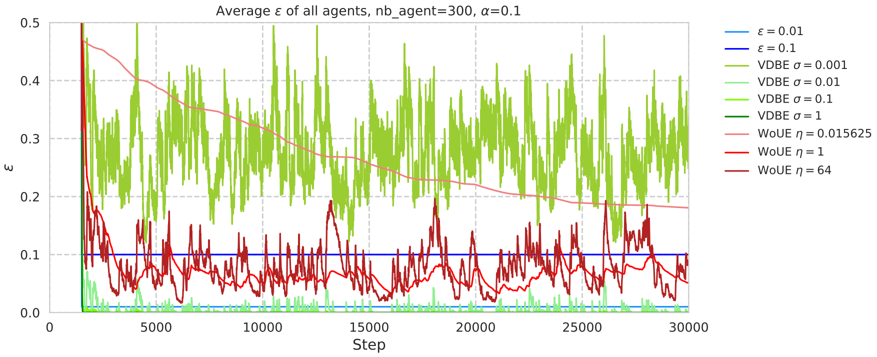

Figure 5: Changes in average ϵ of all agents for each method in non-stationary RRSP. VDBE with $\sigma = 0.001$ is out of this graph.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Todd Hester, Manuel Lopes, and Peter Stone. Learning exploration strategies in model-based reinforcement learning. In *The Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2013.

Itsuki Noda and Masayuki Ohta. Meta-level control of multiagent learning in dynamic repeated resource sharing problems. In Tu-Bao Ho and Zhi-Hua Zhou, editors, *Proc. of PRICAI 2008*, pages 296–308. Springer, Dec. 2008.

Takuya Okano and Itsuki Noda. Adaptation method of the exploration ratio based on the orientation of equilibrium in multi-agent reinforcement learning under non-stationary environments. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 21(5):939–947, Sep. 2017.

Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38 (3):287–308, Mar 2000.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning : An Introduction.* MIT Press, 1998.

Michel Tokic. Adaptive ϵ-greedy exploration in reinforcement learning based on value differences. In *Proceedings of the 33rd Annual German Conference on Advances in Artificial Intelligence*, KI'10, pages 203–210, Berlin, Heidelberg, 2010.

Michel Tokic and Gnther Palm. Value-difference based exploration: Adaptive control between epsilon-greedy and softmax. In *KI 2011: Advances in Artificial Intelligence*, volume 7006 of *Lecture Notes in Computer Science*, pages 335–346. Springer, 2011.