Reinforcement Learning in Multi-Agent Systems with Partial History Sharing

Jalal Arabneydi
Department of Electrical and Computer Engineering
McGill University
Montreal, QC H3A 0E9
jalal.arabneydi@mail.mcgill.ca

Aditya Mahajan
Department of Electrical and Computer Engineering
McGill University
Montreal, QC H3A 0E9
aditya.mahajan@mcgill.ca

Abstract

In this paper, we are interested in systems with multiple agents that wish to cooperate in order to accomplish a common task while a) agents have different information (decentralized information) and b) agents do not know the complete model of the system i.e., they may only know the partial model or may not know the model at all. The agents must learn the optimal strategies by interacting with their environment i.e., by multi-agent Reinforcement Learning (RL). The presence of multiple agents with different information makes multi-agent (decentralized) reinforcement learning conceptually more difficult than single-agent (centralized) reinforcement learning.

We propose a novel multi-agent reinforcement learning algorithm that learns ϵ-team-optimal solution for systems with partial history sharing information structure, which encompasses a large class of multi-agent systems including delayed sharing, control sharing, mean field sharing, etc. Our approach consists of two main steps as follows: 1) the multi-agent (decentralized) system is converted to an equivalent single-agent (centralized) POMDP (Partial Observable Markov Decision Process) using the common information approach of Nayyar et al, TAC 2013, and 2) based on the obtained POMDP, an approximate RL algorithm is constructed using a novel methodology. Particularly, in the second step, since the POMDP obtained in the first step requires the complete-knowledge of system model, we introduce a new concept that we call "Incrementally Expanding Representation (IER)". The main feature of IER is to remove the dependency of the POMDP from complete-knowledge of the model. Then, based on an appropriately defined IER, we follow three sub-steps: 2a) convert the POMDP to a countable-state MDP Δ, 2b) approximate Δ with a sequence of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$, and 2c) use a RL algorithm to learn optimal strategy of MDP Δ_N.

We show that the performance of the RL strategy converges to the optimal performance exponentially fast. We illustrate the proposed approach and verify it numerically by obtaining a multi-agent Q-learning algorithm for two-user Multi Access Broadcast Channel (MABC) which is a benchmark example for multi-agent systems.

Keywords: Reinforcement Learning, Multi-agent Systems, Stochastic Systems.

Acknowledgements

We are deeply indebted to iCORE for their generous support of RLDM2015.

1 Introduction

In this paper, we propose a multi-agent Reinforcement Learning (RL) algorithm that guarantees team-optimal solution. Existing approaches for multi-agent learning may be categorized as follows: exact methods and heuristics. The exact methods rely on the assumption that the information structure is such that all controllers can consistently update the Q-function. These include approaches that rely on social convention and rules to restrict the decisions made by the controllers [5]; approaches that use communication to convey the decisions to all controllers [6]; and approaches that assume that the Q-function decomposes into a sum of terms, each of which is independently updated by a controller [7]. Heuristic approaches include joint action learners heuristic [8] where each controller learns the empirical model of the system in order to estimate the control action of other controllers; frequency maximum Q-value heuristic [9] where controllers keep track of the frequency with which each action leads to a "good" outcome; heuristic Q-learning [10] which assigns a rate of punishment for each controller; and distributed Q-learning [11] which uses predator-prey models to assign heuristic sub-goals to individual controllers. To best of our knowledge, there is no RL approach that guarantees team optimal solution. In this paper, we present an approach that guarantees team-optimal solution.

2 System Model

Let $X_t \in \mathcal{X}$ denote the state of a dynamical system controlled by n agents. At time t, agent i observes $Y_t^i \in \mathcal{Y}^i$ and chooses $U_t^i \in \mathcal{U}^i$. For ease of notation, we denote the joint action and the joint observation by $\mathbf{U}_t = (U_t^1, \ldots, U_t^n)$ and $\mathbf{Y}_t = (Y_t^1, \ldots, Y_t^n)$, respectively. The dynamics of the system are given by

$$X_{t+1} = f(X_t, \mathbf{U}_t, W_t^s), \tag{1}$$

and the observations are given by

$$\mathbf{Y}_t = h(X_t, \mathbf{U}_{t-1}, W_t^o). \tag{2}$$

In this paper, all system variables are considered finite valued. Let $I_t^i \subseteq \{\mathbf{Y}_{1:t}, \mathbf{U}_{1:t-1}\}$ be information available at agent i at time t. The collection $(\{I_t^i\}_{t=1}^{\infty}, i = 1, \ldots, n)$ is called the *information structure*. In this paper, we restrict attention to an information structure called *partial history sharing* (PHS) [1], which will be defined later.

At time t, agent i chooses action U_t^i according to *control law* g_t^i as follows

$$U_t^i = g_t^i(I_t^i). \tag{3}$$

We denote $\mathbf{g}^i = (g_1^i, g_2^i, \ldots)$ as *strategy* of agent i and $\mathbf{g} = (\mathbf{g}^1, \ldots, \mathbf{g}^n)$ as joint strategy of agents. The performance of strategy $\mathbf{g}$ is measured by the following infinite-horizon discounted cost

$$J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(\mathbf{X}_t, \mathbf{U}_t)\right], \tag{4}$$

where discount factor $\beta \in (0,1)$. We are interested in the following problem.

Problem 1 *Given the information structure, action spaces $\{\mathcal{U}^i\}_{i=1}^n$, observation spaces $\{\mathcal{Y}^i\}_{i=1}^n$, discount factor β, and any $\epsilon > 0$, develop a (model-based or model-free) reinforcement learning algorithm that guarantees an ϵ-optimal strategy $\mathbf{g}^*$.*

3 Preliminaries on Partial History Sharing

Herein, we present a simplified version of partial history sharing information structure, originally presented in [1].

Definition 1 ([1], Partial History Sharing (PHS)) *Consider a decentralized control system with n agents. Let I_t^i denote the information available to agent i at time t. Assume $I_t^i \subseteq I_{t+1}^i$. Then, split the information at each agent into two parts: common information $C_t = \bigcap_{i=1}^n I_t^i$ i.e. the information shared between all agents and local information $M_t^i = I_t^i \backslash C_t$ that is the local information of agent i. Define $Z_t := C_{t+1} \backslash C_t$ as common observation, then $C_{t+1} = Z_{1:t}$. An information structure is called partial history sharing when the following conditions are satisfied:*

a) The update of local information $M_{t+1}^i \subseteq \{M_t^i, U_t^i, Y_{t+1}^i\} \backslash Z_t, \quad i \in \{1, \ldots, n\}$.

b) For every agent i, the size of local information M_t^i and the size of common observation Z_t are uniformly bounded in time t.

These conditions are fairly mild and are satisfied by a large class of models.

Remark 1 *Note that conditions (a) and (b) are valid even if there is no common information between agents i.e., $C_t = \emptyset$.*

4 Approach

In this section, we derive our results for systems that have partial history sharing information structure defined above. Our approach consists of two steps. In the first step, we consider the setup of the complete-knowledge of the model and use the *common information approach* of [1] to convert the multi-agent system with PHS information structure to an equivalent single-agent POMDP. In the second step, based on the obtained POMDP, we develop an approximate RL algorithm for the setup of incomplete-knowledge of the model. We show that the error associated with the approximate RL converges to zero exponentially fast.

4.1 Step 1: An Equivalent single-agent POMDP

In this section, we present common information approach of [1] and its main results for the setup of complete-knowledge of the model described in Section 2. Let $\Gamma_t^i : \mathcal{M}^i \mapsto \mathcal{U}^i$ be the mapping from the local information of subsystem i to action of subsystem i at time t i.e. $U_t^i = \Gamma_t^i(M_t^i)$.

Consider a virtual *coordinator* (single agent) that observes the common information shared between all subsystems by time t i.e. C_t. Based on C_t, the coordinator prescribes functions $\mathbf{\Gamma}_t = (\Gamma_t^1, \dots, \Gamma_t^n) \in \mathcal{G}$ to subsystems, where $\mathcal{G} = \prod_{i=1}^n \mathcal{G}^i$ denotes the space of joint mappings $\mathbf{\Gamma}_t$ and $\mathcal{G}^i$ denotes the space of mappings Γ_t^i. Hence, $\Gamma_t^i = \psi_t^i(C_t), \quad \forall i \in \{1, \dots, n\}$ where $\boldsymbol{\psi}_t = \{\psi_t^1, \dots, \psi_t^n\}$ is called the *coordination law* and $\mathbf{\Gamma}_t = (\Gamma_t^1, \dots, \Gamma_t^n)$ is called the *prescription*. In the sequel, for ease of notation, we will use the following compact form for the coordinator's law, $\mathbf{\Gamma}_t = \boldsymbol{\psi}_t(C_t)$. We call $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots\}$ as the *coordination strategy*. In the *coordinated system*, dynamics and cost function are as same as those in the original problem in Section 2. In particular, the infintie-horizon discounted cost in the coordinated system is as follows:

$$J(\boldsymbol{\psi}) = \mathbb{E}^{\boldsymbol{\psi}}\left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(\mathbf{X}_t, \mathbf{\Gamma}_t^1(M_t^1), \dots, \Gamma_t^n(M_t^n))\right]. \tag{5}$$

Lemma 1 ([1], Proposition 3) *The original system described in Section 2 with PHS information structure is equivalent to the coordinated system.*

According to [1], $\Pi_t = \mathbb{P}(\mathbf{X}_t, \mathbf{M}_t | Z_{1:t-1}, \mathbf{\Gamma}_{1:t-1})$ is an information state for the coordinated system with initial state $\Pi_1 = P_X$. It is shown in [1] that

1. There exists a function ϕ such that $\Pi_{t+1} = \phi(\Pi_t, \mathbf{\Gamma}_t, Z_t)$.
2. The observation Z_t only depends on (Π_t, Γ_t) i.e. $\mathbb{P}(Z_t{=}z_t|\Pi_{1:t}{=}\pi_{1:t}, \mathbf{\Gamma}_{1:t}{=}\boldsymbol{\gamma}_{1:t}){=}\mathbb{P}(Z_t{=}z_t|\Pi_t{=}\pi_t, \mathbf{\Gamma}_t{=}\boldsymbol{\gamma}_t)$.
3. There exists a function $\hat{\ell}$ such that $\hat{\ell}(\pi_t,\boldsymbol{\gamma}_t){=}\mathbb{E}[\ell(\mathbf{X}_t, \mathbf{U}_t | Z_{1:t-1}{=}z_{1:t-1}, \mathbf{\Gamma}_{1:t}{=}\boldsymbol{\gamma}_{1:t})]$.

Assume that the initial state π_1 is fixed. Let $\mathcal{R}$ denote the reachable set of above centralized POMDP that contains all the realizations of π_t generated by $\pi_{t+1} = \phi(\pi_t, \boldsymbol{\gamma}, z), \forall \boldsymbol{\gamma} \in \mathcal{G}, \forall z \in \mathcal{Z}, \forall t \in \mathbb{N}$, with initial information state π_1. Note that since all the variables are finite valued, then $\mathcal{G}$ (set of all prescriptions $\boldsymbol{\gamma}$) and $\mathcal{Z}$ (set of all observations of the coordinator) are finite sets. Hence, $\mathcal{R}$ is at most a countable set. In the next step, we develop an approximate RL algorithm based on the obtained POMDP for the setup of incomplete-knowledge of the model.

4.2 Step 2: An Approximate RL algorithm for POMDP

In the previous step, we identified a single-agent POMDP that is equivalent to the multi-agent system with PHS information structure. However, the obtained POMDP requires the complete knowledge of the model. To circumvent this requirement, we introduce a new concept that we call *Incrementally Expanding Representation* (IER). The main feature of IER is to remove the dependency of the POMDP from the complete knowledge of the model. Using the IER, we follow three sub-steps: 2a) convert the POMDP to a countable-state MDP Δ, 2b) construct a sequence of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$ of MDP Δ, and 2c) use a generic RL algorithm to learn an optimal strategy of Δ_N.

Definition 2 (Incrementally Expanding Representation (IER)) *Let $\{\mathcal{S}_k\}_{k=1}^{\infty}$ be a sequence of finite sets such that $\mathcal{S}_1 \subsetneq \mathcal{S}_2 \subsetneq \dots \subsetneq \mathcal{S}_k \subsetneq \dots$, and $\mathcal{S}_1$ is a singleton, say $\mathcal{S}_1 = \{s^*\}$. Let $\mathcal{S} = \lim_{k\to\infty} \mathcal{S}_k$ be the countable union of above finite sets, $B : \mathcal{S} \to \mathcal{R}$ be a sujrjective function that maps $\mathcal{S}$ to the reachable set $\mathcal{R}$, and $\tilde{f} : \mathcal{S} \times \mathcal{G} \times \mathcal{Z} \to \mathcal{S}$. The tuple $\langle \{\mathcal{S}_k\}_{k=1}^{\infty}, B, \tilde{f}\rangle$ is called an* incrementally expanding representation *(IER), if it satisfies the following properties:*

(P1) Incremental Expansion: For any $\boldsymbol{\gamma} \in \mathcal{G}, z \in \mathcal{Z}$, and $s \in \mathcal{S}_k$, we have that

$$\tilde{f}(s, \boldsymbol{\gamma}, z) \in \mathcal{S}_{k+1}. \tag{6}$$

(P2) Consistency: For any $(\boldsymbol{\gamma}_{1:t-1}, z_{1:t-1})$, let Π_t be the information state of the obtained POMDP and S_t be the state obtained by recursive application of (6) *starting from $S_1 = s^*$. Then, $\Pi_t = B(S_t)$.*

Lemma 2 *Every multi-agent system with PHS information structure has at least one IER.*

4.2.1 Countable-state MDP Δ

Let the tuple $\langle \{\mathcal{S}_k\}_{k=1}^{\infty}, B, \tilde{f} \rangle$ be an IER of the POMDP obtained in the first step. Then, define MDP Δ with countable state space $\mathcal{S}$, finite action space $\mathcal{G}$, and dynamics $\tilde{f}$ such that:

(F1) $\mathcal{S} = \lim_{k \to \infty} S_k$ is the (countable) state space and $\mathcal{G}$ is the finite action space of MDP Δ. The initial state is singleton s^*. The state $S_t \in \mathcal{S}_k$, $k \leq t$, evolves as follows:

$$S_{t+1} = \tilde{f}(S_t, \mathbf{\Gamma}_t, Z_t), \quad S_{t+1} \in \mathcal{S}_{k+1}, \mathbf{\Gamma}_t \in \mathcal{G}, Z_t \in \mathcal{Z},$$

where observation Z_t only depends on $(S_t, \mathbf{\Gamma}_t)$. At time t, there is a cost depending on the current state $S_t \in \mathcal{S}$ and action $\mathbf{\Gamma}_t \in \mathcal{G}$ given by $\tilde{\ell}(S_t, \mathbf{\Gamma}_t) = \hat{\ell}(B(S_t), \mathbf{\Gamma}_t) = \hat{\ell}(\Pi_t, \mathbf{\Gamma}_t)$.

(F2) State space $\mathcal{S}$, action space $\mathcal{G}$, and dynamics $\tilde{f}$ do not depend on the unknowns.

The performance of a stationary strategy $\tilde{\boldsymbol{\psi}} : \mathcal{S} \mapsto \mathcal{G}$ is quantified by $\tilde{J}(\tilde{\boldsymbol{\psi}}) = \mathbb{E}^{\tilde{\boldsymbol{\psi}}} \left[\sum_{t=1}^{\infty} \beta^{t-1} \tilde{\ell}(S_t, \mathbf{\Gamma}_t) \right]$.

Lemma 3 *There exists at least one Δ that satisfies F1 and F2. Also, let $\tilde{\boldsymbol{\psi}}^*$ be an optimal strategy of MDP Δ. Construct a strategy $\boldsymbol{\psi}^*$ for the coordinated system as follows: $\tilde{\boldsymbol{\psi}}^*(s) =: \boldsymbol{\psi}^*(B(s)), \forall s \in \mathcal{S}$. Then, $\tilde{J}(\tilde{\boldsymbol{\psi}}^*) = J(\boldsymbol{\psi}^*)$ and $\boldsymbol{\psi}^*$ is an optimal strategy for the coordinated system, and therefore can be used to generate an optimal strategy for the original multi-agent system.*

4.2.2 Finite-state incrementally expanding MDP Δ_N

In this part, we construct a series of finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$, that approximate the countable-state MDP Δ as follows. Let Δ_N be a finite-state MDP with state space $\mathcal{S}_N$ and action space $\mathcal{G}$. The transition probability of Δ_N is constructed as follows. Pick any arbitrary set $D^* \in S_N$. Remap every transition in Δ that takes the state $s \in \mathcal{S}_N$ to $s' \in \mathcal{S}_{N+1} \backslash \mathcal{S}_N$ to a transition from $s \in \mathcal{S}_N$ to any (not necessarily unique) state in D^*. In addition, the per-step cost function of Δ_N is simply a restriction of $\tilde{\ell}$ to $\mathcal{S}_N \times \mathcal{G}$. Also, we assume that there exists an action or a sequence of actions that if taken, the system transmits to a known state (states) d^* in D^*. Then, dynamics of Δ_N is as follows.

$$S_{t+1} = \begin{cases} \tilde{f}(S_t, \mathbf{\Gamma}_t, Z_t) & \tilde{f}(S_t, \mathbf{\Gamma}_t, Z_t) \in \mathcal{S}_N \\ d^* & \tilde{f}(S_t, \mathbf{\Gamma}_t, Z_t) \in \mathcal{S}_{N+1} \backslash \mathcal{S}_N \end{cases} \tag{7}$$

Theorem 1 *Let $\tilde{\boldsymbol{\psi}}^*$ be an optimal strategy of MDP Δ and $\tilde{\boldsymbol{\psi}}^*_N$ be an optimal strategy of MDP Δ_N. Then, the difference in performance is bounded as follows: $|\tilde{J}(\tilde{\boldsymbol{\psi}}^*) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}^*_N)| \leq \frac{2\beta^{\tau_N}}{1-\beta} L_{max}$, where L_{max} denotes the maximum instantaneous cost and τ_N is a model dependent parameter that is $N \leq \tau_N$.*

4.2.3 RL algorithm for MDP Δ_N

Let $\mathcal{T}$ be a generic (model-based or model-free) RL algorithm designed for finite-state MDPs with infinite horizon discounted cost. By a generic RL algorithm, we mean any algorithm which fits to the following framework. At each iteration $k \in \mathbb{N}$, $\mathcal{T}$ knows the state of system, selects one action, and observes an instantaneous cost and the next state. The strategy learned by $\mathcal{T}$ converges to an optimal strategy as $k \to \infty$.

Let $\tilde{\boldsymbol{\psi}}^k_N : \mathcal{S}_N \to \mathcal{G}$ be the learned strategy associated with RL algorithm $\mathcal{T}$ operating on MDP Δ_N at iteration k such that

$$\lim_{k \to \infty} |\tilde{J}_N(\tilde{\boldsymbol{\psi}}^k_N) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}^*_N)| = 0. \tag{8}$$

Now, we convert (translate) the strategies in Δ_N to strategies in the original multi-agent system described in Section 2, where the actual learning happens. Hence, we define a strategy $\boldsymbol{g}^k_N := (g^{k,i}_N, \ldots, g^{k,n}_N)$, at iteration k, as follows:

$$g^{k,i}_N(s, m^i) := \tilde{\boldsymbol{\psi}}^{k,i}_N(s)(m^i), \forall s \in \mathcal{S}_N, \forall m^i \in \mathcal{M}^i, \forall i, \tag{9}$$

where $\tilde{\boldsymbol{\psi}}^{k,i}_N$ denotes the ith term of $\tilde{\boldsymbol{\psi}}^k_N$ and state s updates according to (7).

Theorem 2 *Let J^* be the optimal performance of the original multi-agent system given in* (4). *Then, the approximation error associated with using the learned strategy is bounded as follows:*

$$\lim_{k \to \infty} |J^* - J(\boldsymbol{g}^k_N)| = |\tilde{J}(\tilde{\boldsymbol{\psi}}^*) - \tilde{J}_N(\tilde{\boldsymbol{\psi}}^*_N)| \leq \epsilon_N, \tag{10}$$

where $\epsilon_N = \frac{2\beta^{\tau_N}}{1-\beta} L_{max} \leq \frac{2\beta^N}{1-\beta} L_{max}$. Note that the error goes to zero exponentially in N.

5 Example

Consider a 2-user multi access broadcast channel (MABC) system first defined in [4]. The system consists of 2 users that have a buffer of size 1 (thus, $X_t = (X_t^1, X_t^2) \in \{0,1\}^2)$. Packets arrive at each user i according to independent Bernoulli processes with rate p^i. Each user observes the state of its own queue i.e. $(Y_t^i = X_t^i)$ and transmits if it has a packet (i.e. $U_t^i \in \{0,1\}$ and $U_t^i \leq X_t^i$). If only one user transmits, then the transmission is successful and the packet is removed from the queue. If both users transmit, there is a "collision" and the packets remain in the queues. Users can sense whether the channel was used or if a collision took place. Thus, the information available at each user i is $I_t^i = \{X_t^i, \mathbf{U}_{1:t-1}\}$, where $\mathbf{U}_t = (U_t^1, U_t^2)$. The objective is to maximize the throughput. Hence, the instantaneous reward is defined as follows: $r(X_t, \mathbf{U}_t) = U_t^1 + U_t^2 - 2U_t^1 U_t^2$.

At time t, the common observation $Z_t = \mathbf{U}_t$ and the common information $C_t = \{\mathbf{U}_{1:t-1}\}$. For this specific model, the prescription γ^i is completely specified by $A_t^i := \gamma_t^i(1)$ (since $\gamma_t^i(0)$ is always 0). Hence, $U_t^i = \gamma_t^i(X_t^i) = A_t^i \cdot X_t^i$. Therefore, we may equivalently assume that the coordinator generates actions $\mathbf{A}_t = (A_t^1, A_t^2)$. Define $\mathbf{\Pi}_t = (\Pi_t^1, \Pi_t^2)$, $\Pi_t^i = \mathbb{P}(X_t^i = 1 \mid \mathbf{U}_{1:t-1}, \mathbf{A}_{1:t-1})$, as information state for the coordinated system with initial state $\mathbf{\Pi}_1 = (p^1, p^2)$. Thus, the reachable set $\mathcal{R}$ is given by $\mathcal{R} := \{(1,1), (1,p^1), (p^2,1), (p^1,p^2)\} \cup \{(p^1, T_2^n p^2) : n \in \mathbb{N}\} \cup \{(T_1^n p^1, p^2) : n \in \mathbb{N}\}$, where $T_i^n q = T_i(T_i^{n-1} q)$. Let b_1, b_2 be any arbitrary number in $(0,1)$. Define $\mathcal{S} = \{S_k\}_{k=1}^{\infty}$ as the countable state space of Δ, where $S_1 = \{(0,0)\}$ and $S_k = \{(0,0), (0,1), (1,0), (1,1), (0, 1-b_1^i), (1-b_2^i, 0)\}_{i=1}^{k-1}$, $k \geq 2$. The action space is $\mathcal{A} = \{(0,1), (1,0), (1,1)\}$ (note that the action $(0,0)$ is dominated, so it is removed without loss of optimality).

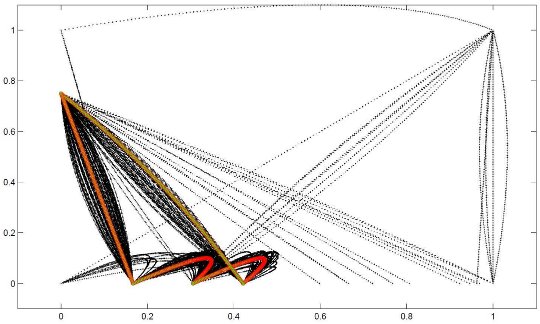

Figure 1: This figures shows the learning process of MDP Δ_N in a few snapshots. In this simulation, we use the following numerical values: $b_1 = 0.25, b_2 = 0.83, N = 20, \beta = 0.99, p^1 = 0.3, p^2 = 0.6$. In particular, the optimal strategy is a recurrent class consisting of states $(0, 1-b_1^1)$, $(1-b_2^1, 0)$, $(1-b_2^2, 0)$, and $(1-b_2^3, 0)$. The learning procedure is plotted in black and the optimal recurrent class is plotted in red. It is seen that the state of the system is eventually trapped in the optimal recurrent class. The optimal strategy says that user with rate of 0.6 must transmit 3 times more than the user with rate of 0.3.

References

[1] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis, Decentralized Stochastic Control with Partial History Sharing: A Common Information Approach, IEEE Transaction on Automatic Control, vol. 58, no. 7, 2013.

[2] Tilak, Omkar and Mukhopadhyay, Snehasis, Partially decentralized reinforcement learning in finite, multi-agent Markov decision processes, AI Communications, vol. 24, no. 4, pp 293–309, 2011.

[3] Busoniu, Lucian and Babuska, Robert and De Schutter, Bart, A comprehensive survey of multiagent reinforcement learning, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.38, no. 2, pp 156–172, 2008.

[4] M. G. Hluchyj and R. G. Gallager, Multiacces of a slotted channel by nitely many users, Proc. of Nat. Tel. Con., pp. 421-427, 1981.

[5] M. T. J. Spaan, N. Vlassis, and F. C. A. Groen, High level coordination of agents based on multiagent Markov decision processes with roles, in Workshop on Coop. Rob., IEEE/RSJ Int., pp. 66-73, 2002.

[6] N. Vlassis, A concise introduction to multiagent systems and distributed AI,Univ. of Amsterdam, Tech. Rep., 2003.

[7] J. R. Kok, M. T. J. Spaan, and N. Vlassis, Non-communicative multirobot coordination in dynamic environment, Robotics and Autonomous Systems, Vol. 50, No. 2-3, pp. 99-114, 2005.

[8] C. Claus and C. Boutilier, The dynamics of reinforcement learning in cooperative multiagent systems, 10th Conf. on Inn. Appl. of AI , pp. 746-752, 1998.

[9] S. Kapetanakis and D. Kudenko, Reinforcement learning of coordination in cooperative multi-agent systems, 14th conf. on Inn. Appl. of AI , pp. 326-331, 2002.

[10] Laetitia Matignon, Guillaume J. Laurent and Nadine Le Fort-Piat, Hysteretic Q-Learning : an algorithm for Decentralized Reinforcement Learning in Cooperative Multiagent Teams, Int. Rob. and Sys., 2007.

[11] Jing huang, Bo Yang, and Da-You Liu, A distributed Q-learning Algorithm for Multi-Agent Team coordination, IEEE 40th Int. conf. on Mach. Ler. and Cyb., Vol. 1, pp. 108-113, Aug., 2005.