Automatica 104 (2019) 90–101

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

On convergence rates of game theoretic reinforcement learning algorithms[☆]

Zhisheng Hu[a], Minghui Zhu[a,*], Ping Chen[b], Peng Liu[c]

[a] School of Electrical Engineering and Computer Science, Pennsylvania State University, 201 Old Main, University Park, PA, 16802, USA
[b] JD.com, No. 18 Kechuang 11 Street, BDA, Beijing, 10111, China
[c] College of Information Sciences and Technology, Pennsylvania State University, 201 Old Main, University Park, PA, 16802, USA

ARTICLE INFO

Article history:
Received 11 December 2016
Received in revised form 23 August 2018
Accepted 5 February 2019
Available online 12 March 2019

Keywords:
Distributed control
Game theory
Learning algorithms

ABSTRACT

This paper investigates a class of multi-player discrete games where each player aims to maximize its own utility function. Each player does not know the other players' action sets, their deployed actions or the structures of its own or the others' utility functions. Instead, each player only knows its own deployed actions and its received utility values in recent history. We propose a reinforcement learning algorithm which converges to the set of action profiles which have maximal stochastic potential with probability one. Furthermore, an upper bound on the convergence rate is derived and is minimized when the exploration rates are restricted to **p**-series. The algorithm performance is verified using a case study in the smart grid.

1. Introduction

Game theory provides a mathematically rigorous framework for multiple players to reason about each other. In recent years, game theoretic learning has been increasingly used to control large-scale networked systems due to its inherent distributed nature. In particular, the network-wide objective of interest is encoded as a game whose Nash equilibria correspond to desired network-wide configurations. Numerical algorithms are then synthesized for the players to identify Nash equilibria via repeated interactions. Multi-player games can be categorized into discrete games and continuous games. In a discrete (resp. continuous) game, each player has a finite (resp. an infinite) number of action candidates. As for discrete games, learning algorithms include best-response dynamics, better-response dynamics, fictitious play, regret matching, logit-based dynamics and replicator dynamics. Please refer to Basar and Olsder (1999), Fudenberg and Levine (1998), Sandholm (2010) and Young (2001) for detailed discussion. As an important class of continuous games, generalized Nash games were first formulated in Arrow and Debreu (1954), and see survey paper (Facchinei & Kanzow, 2007) for a comprehensive exposition. A number of algorithms have been proposed to compute generalized Nash equilibria, including, to name a few, ODE-based methods (Rosen, 1965), nonlinear Gauss–Seidel-type approaches (Pang, Scutari, Facchinei, & Wang, 2008), iterative primal–dual Tikhonov schemes (Yin, Shanbhag, & Mehta, 2011), and best-response dynamics (Palomar & Eldar, 2010). Game theory and its learning have found many applications; e.g., traffic routing in Internet (Altman, Basar, & Srikant, 2002), urban transportation (Roumboutsos & Kapros, 2008), mobile robot coordination (Arslan, Marden, & Shamma, 2007; Hatanaka, Wasa, Funada, Charalambides, & Fujita, 2016) and power markets (Wang, Shanbhag, & Meyn, 2012; Zhu, 2014).

In many applications, players can only access limited information about the game of interest. For example, each player may not know the structure of its own utility function. Additionally, during repeated interactions, each player may not be aware of the actions of other players. These informational constraints motivate recent study on payoff-based or reinforcement learning algorithms where the players adjust their actions only based on their own previous actions and utility measurements. The papers (Hatanaka et al., 2016; Marden, Young, Arslan, & Shamma, 2009; Zhu & Martínez, 2013) study discrete games, and their approaches are based on stochastic stability (Foster & Young, 1990). As mentioned in Remark 3.2 of Zhu and Martínez (2013), the paper (Marden et al., 2009) proposes an algorithm to find

[☆] Z. Hu and M. Zhu were partially supported by ARO, United States W911NF-13-1-0421 (MURI), NSA, United States H98230-15-1-0289 and National Science Foundation (NSF), United States ECCS-1710859. P. Chen and P. Liu were partially supported by ARO, United States W911NF-13-1-0421 (MURI) and National Science Foundation (NSF), United States CNS-1422594. The material in this paper was partially presented at the First ACM Workshop on Moving Target Defense, in conjunction with 2014 ACM Conference on Computer and Communications Security, November 3, 2014, Scottsdale, Arizona, USA. This paper was recommended for publication in revised form by Associate Editor Gurdal Arslan under the direction of Editor Ian R. Petersen.

[*] Corresponding author.
E-mail addresses: zxh128@psu.edu (Z. Hu), muz16@psu.edu (M. Zhu), chenping19851@hotmail.com (P. Chen), pliu@ist.psu.edu (P. Liu).

https://doi.org/10.1016/j.automatica.2019.02.032
0005-1098/© 2019 Elsevier Ltd. All rights reserved.

Nash equilibrium of weakly acyclic games with an arbitrarily high probability by choosing an arbitrarily small and fixed exploration rate in advance. The analysis in Marden et al. (2009) is based on homogeneous Markov chains and more specifically the theory of resistance trees (Young, 1993). Zhu and Martínez (2013) extend the results in Marden et al. (2009) by adopting diminishing exploration rates and ensure convergence to Nash equilibrium and global optima with probability one. The analysis of Zhu and Martínez (2013) is based on strong ergodicity of inhomogeneous Markov chains. As for continuous games, the papers (Frihauf, Krstic, & Basar, 2012; Liu & Krstic, 2011; Stankovic, Johansson, & Stipanovic, 2012) employ extremum seeking and the paper (Zhu & Frazzoli, 2016) uses finite-difference approximations to estimate unknown partial (sub)gradients. Notice that all the aforementioned papers focus on asymptotic convergence and none of them quantifies convergence rates.

Contribution: In this paper, we study a class of multi-player discrete games where each player is unaware of the other players' action sets, their deployed actions or the structures of its own or the others' utility functions. We propose a reinforcement learning algorithm where, at each iteration, each player, on the one hand, exploits successful actions in recent history via comparing received utility values, and on the other hand, randomly explores any feasible action with a certain exploration rate. The algorithm is proven to be convergent to the set of action profiles with maximum stochastic potential with probability one. Furthermore, an upper bound on the convergence rate is derived and is minimized when the exploration rates are restricted to **p**-series. When the interactions of the players consist of a weakly acyclic game, the convergence to the set of pure Nash equilibria is guaranteed. The algorithm performance is verified using a case study in the smart grid. A preliminary version of this paper was published in Zhu, Hu, and Liu (2014) where convergence rates are not discussed. Further, Zhu et al. (2014) focus on the application of cyber security, and this paper focuses on the theory of learning in games. The analysis of these two papers is significantly different.

2. Problem formulation and learning algorithm

In this section, we introduce a class of multi-player games where the information each player accesses is limited. Then, we present a learning algorithm under which the action profiles of the players converge to the set of action profiles which have maximum stochastic potential.

2.1. Game formulation

The interactions of N players are characterized as a non-cooperative game. Each component of the game will be discussed in the following paragraphs.

Players. We consider N players $\mathcal{V} \triangleq \{1, \dots, N\}$ and each player has a finite set of actions. Let $\mathcal{A}_i$ denote the action set of player i and $a^i \in \mathcal{A}_i$ denote an action of player i. Denote $\mathcal{S} \triangleq \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ as the Cartesian product of the action sets, where $s \triangleq (a^1, \dots, a^N) \in \mathcal{S}$ is denoted as an action profile of the players.

Utility. Under the influence of an action profile, the system generates a utility value for each player. The utility function for player $i \in \mathcal{V}$ is defined as $u_i : \mathcal{S} \to \mathbb{R}$. At the end of iteration t, the utility value $u_i(t) = u_i(s(t))$ is measured and sent to player i.

Informational constraint. Each player does not know the other players' action sets or their deployed actions. Besides, each player is unaware of the structure of its own or the others' utility functions. At iteration t, each player only knows its deployed actions and its received utility values in the past; i.e., $a^i(0), \dots, a^i(t-1), u_i(0), \dots, u_i(t-1)$.

The above informational constraint has been studied in several recent papers. For example, the authors in Hatanaka et al. (2016), Stankovic et al. (2012) and Zhu and Martínez (2013) investigate coverage optimization problems for mobile sensor networks where mobile sensors are unaware of environmental distribution functions. The authors in Marden, Ruben, and Pao (2013) study the problem of optimizing energy production in wind farms where each turbine knows neither the functional form of the power generated by the wind farm nor the choices of other turbines. The authors in Frihauf et al. (2012) and Zhu and Frazzoli (2016) consider convex games where each player cannot access its game components.

2.2. Problem statement

Under the above informational constraint, we aim to synthesize a learning algorithm under which the action profiles of the players converge to the set of action profiles with maximum stochastic potential. We will quantify the convergence rate of the proposed algorithm in contrast to asymptotic convergence in existing work.

2.3. Learning algorithm

Inspired by Zhu and Martínez (2013), we propose a learning algorithm called the RL algorithm, where each player updates its actions only based on its previous actions and its received utility values. On the one hand, each player chooses the most successful action in recent history. It represents the exploitation phase. However, the exploitation is not sufficient to guarantee that the player can choose the best action given others'. So on the other hand, the player uniformly chooses one action from its action set. It represents the exploration phase. The specific update rule is stated in the RL algorithm. At iterations $t = 0$ and $t = 1$, each player uniformly chooses one action from its action set (Line 3). Starting from iteration $t = 2$, with probability $1 - \tilde{\epsilon}_i(t)$, player i chooses the action which generates a higher utility value in last two iterations as current action (Lines 8–13). This represents the exploitation where player i reinforces its previous successful actions. With probability $\tilde{\epsilon}_i(t)$, player i uniformly selects an action from its action set $\mathcal{A}_i$ (Line 14). This represents the exploration and makes sure that each action profile is selected infinitely often. Note that sample($\mathcal{A}_i$) in Line 14 represents uniformly choosing one element from set $\mathcal{A}_i$.

Algorithm 1. Reinforcement learning (RL) algorithm

```
1:  while $0 \le t \le 1$ do
2:    for $i \in \mathcal{V}$ do
3:      $a^i(t) \leftarrow$ sample($\mathcal{A}_i$);
4:    end for
5:  end while
6:  while $t \ge 2$ do
7:    for $i \in \mathcal{V}$ do
8:      With prob. $(1 - \tilde{\epsilon}_i(t))$,
9:      if $u_i(t-1) \ge u_i(t-2)$ then
10:       $a^i(t) = a^i(t-1)$;
11:     else
12:       $a^i(t) = a^i(t-2)$;
13:     end if
14:     With prob. $\tilde{\epsilon}_i(t)$, $a^i(t) \leftarrow$ sample($\mathcal{A}_i$);
15:   end for
16: end while
```

3. Analysis

In this section, we will present the analytical results of the RL algorithm.

3.1. Notations and assumptions

We first introduce the notations and assumptions used throughout the paper. Denote by $|\mathcal{V}|$ the cardinality of player set, $|\mathcal{A}_i|$ the cardinality of action set of player i and $|\mathcal{A}|_\infty \triangleq \max_{i\in\mathcal{V}} |\mathcal{A}_i|$ the maximum cardinality among all action sets. The exploration rate for player i at iteration t is decomposed into two parts; i.e., $\tilde{\epsilon}_i(t) \triangleq \epsilon_i(t) + e_i(t) \in (0, 1]$, where $\epsilon_i(t) = \gamma_i \epsilon^c(t)$, $\gamma_i > 0$, $\epsilon^c(t)$ is common for all the players, γ_i represents the heterogeneity, and $e_i(t)$ represents the exploration deviation. Define $e(t) \triangleq (e_1(t), \ldots, e_N(t))^T$, $\tilde{\epsilon}(t) \triangleq (\tilde{\epsilon}_1(t), \ldots, \tilde{\epsilon}_N(t))^T$ and $\epsilon(t) \triangleq (\epsilon_1(t), \ldots, \epsilon_N(t))^T$. And we define $e_r(t) \triangleq \|e(t)\|_\infty^N / \prod_{i=1}^N \tilde{\epsilon}_i(t)$. Here we denote by $\|\cdot\|_\infty$ the infinity norm of a vector. In addition, we also use $\|\cdot\|$ to represent the L^1-norm of a vector, and $\|P\|$ to represent the 1-norm of a matrix.

Assumption 1. (1) For each $i \in \mathcal{V}$, $\epsilon_i(t) \in (0, 1]$ is non-negative, strictly decreasing, and $\lim_{t\to\infty} \epsilon_i(t) = 0$. (2) The sequences $\{\prod_{i=1}^N \epsilon_i(t)\}$ and $\{\prod_{i=1}^N \tilde{\epsilon}_i(t)\}$ are not summable. (3) $\lim_{t\to\infty} e_r(t) = 0$.

Assumption 1 indicates that the players can choose heterogeneous exploration rates. The exploration rates diminish slowly enough and their deviations decrease in faster rates than the common part. In the paper (Zhu & Martínez, 2013), it is assumed that exploration rates $\epsilon_i(t)$ are identical for all i, diminishing and not summable. Assumption 1 allows for heterogeneous exploration rates and includes homogeneous exploration rates in the paper (Zhu & Martínez, 2013) as a special case. Actually, papers (Koshal, Nedić, & Shanbhag, 2013; Yousefian, Nedić, & Shanbhag, 2013) adopt heterogeneous step-sizes for distributed optimization and game theory. They impose similar assumptions on the step-sizes.

Markov chain induced by the RL algorithm. Denote by $\mathcal{Z} \triangleq \mathcal{S} \times \mathcal{S}$ the state space, where each state $z(t) \triangleq (s(t), s(t+1))$ consists of the action profiles at iteration t and the next iteration. And denote by $diag(\mathcal{S} \times \mathcal{S}) \triangleq \{(s, s) | s \in \mathcal{S}\}$ the diagonal space of $\mathcal{Z}$. By the definition of $z(t)$, the sequence $\{z(t)\}_{t\geq 0}$ forms a time-inhomogeneous Markov chain which is denoted by $\mathcal{M}$. We define $P^{\tilde{\epsilon}(t)}$ as the transition matrix of Markov chain $\mathcal{M}$ at iteration t, where each entry $P^{\tilde{\epsilon}(t)}(z', z)$ represents the transition probability from state z' to z. Besides, denote by $\pi(t)$ the distribution on $\mathcal{Z}$ at iteration t.

***z*-tree of time-homogeneous Markov chain $\mathcal{M}^{\tilde{\epsilon}}$.** Given any two distinct states z' and z of Markov chain $\mathcal{M}^{\tilde{\epsilon}}$, consider all paths starting from z' and ending at z. Denote by $p_{z'z}$ the probability of the path from z' to z. We define graph $\mathcal{G}(\tilde{\epsilon})$ where each vertex of $\mathcal{G}(\tilde{\epsilon})$ is a state z of Markov chain $\mathcal{M}^{\tilde{\epsilon}}$ and the probability on edge (z', z) is $p_{z'z}$. A *z-tree* on $\mathcal{G}(\tilde{\epsilon})$ is a spanning tree rooted at z such that from every vertex $z' \neq z$, there is a unique path from z' to z. Denote by $G_{\tilde{\epsilon}}(z)$ the set of all *z-trees* on $\mathcal{G}(\tilde{\epsilon})$ rooted at z. The total probability of a *z-tree* is the product of the probabilities of its edges. The *stochastic potential* of the state z is the largest total probability among all *z-trees* in $G_{\tilde{\epsilon}}(z)$. Let $\Lambda(\tilde{\epsilon})$ be the states which have maximum stochastic potential for a particular $\tilde{\epsilon} \in (0, 1]^N$. Denote the limit set $\Lambda^* \triangleq \lim_{\tilde{\epsilon}\to 0} \Lambda(\tilde{\epsilon})$. And the elements in Λ^* are referred to as *stochastically stable states*.

Remark 1. The above notions are inspired by the resistance trees theory (Young, 1993). However, the above notions are defined for any $\tilde{\epsilon} \in (0, 1]$ instead of $\tilde{\epsilon} \to 0$ in the resistance trees theory. This allows us to characterize the transient performance of the RL algorithm. □

3.2. Main analytical result

The following theorem is the main analytical result of this paper. It shows that the state $z(t)$ converges to the set of stochastically stable states with probability one. Moreover, the convergence rate is quantified using the distance between $\pi(t)$ and the limiting distribution π^*; i.e., $D(t) \triangleq \|\pi(t) - \pi^*\|$. The formal proof of Theorem 1 will be given in Section 5.

Theorem 1. *If Assumption 1 holds, the following properties hold for the RL algorithm:*

(P1) $\lim_{t\to\infty} Pr\{z(t) \in \Lambda^*\} = 1$ and $\Lambda^* \subseteq diag(\mathcal{S} \times \mathcal{S})$;

(P2) there exist positive integer t_{min} and positive constant C such that for any $t^* > t_{min}$ and $t \geq t^* + 1$, the following is true:

$$D(t) \leq \min\{2, C(\|\epsilon(t^*)\|_\infty + \|\epsilon(t)\|_\infty + e_r(t^*) + \exp(-\sum_{\tau=t^*}^{t-1} \prod_{i=1}^{N} \epsilon_i(\tau)|\mathcal{A}_i|) + \exp(-\sum_{\tau=t^*}^{t-1} \prod_{i=1}^{N} \tilde{\epsilon}_i(\tau)|\mathcal{A}_i|))\}. \quad (1)$$

4. Discussion

4.1. Weakly acyclic games

In this section, we study the special case where the interactions of the players consist of a weakly acyclic game. A game is called to be weakly acyclic if from every action profile, there exists a finite best-response improvement path leading from the action profile to a pure Nash equilibrium. And it is known that any weakly acyclic game has at least one pure Nash equilibrium (Fabrikant, Jaggard, & Schapira, 2010; Milchtaich, 1996; Young, 1993).

Definition 1 (*Pure Nash equilibrium*)**.** An action profile $s_* \triangleq (a_*^1, \ldots, a_*^i, \ldots, a_*^N)$ is a pure Nash equilibrium if $\forall i \in \mathcal{V}, \forall a^i \in \mathcal{A}_i$, $u_i(s_*) \geq u_i(a^i, a_*^{-i})$.

Denote the set of pure Nash equilibria of the game Γ as $\mathcal{N}(\Gamma)$ and $diag(\mathcal{N}(\Gamma) \times \mathcal{N}(\Gamma)) \triangleq \{(s, s) | s \in \mathcal{N}(\Gamma)\}$. The following corollary implies that the action profiles converge to $\mathcal{N}(\Gamma)$ with probability one.

Corollary 1. *If Assumption 1 holds and Γ is a weakly acyclic game, then it holds that $\lim_{t\to\infty} Pr\{z(t) \in diag(\mathcal{N}(\Gamma) \times \mathcal{N}(\Gamma))\} = 1$ for the RL algorithm.*

From Theorem 1, we have $\lim_{t\to\infty} Pr\{z(t) \in \Lambda^*\} = 1$ and $\Lambda^* \subseteq diag(\mathcal{S} \times \mathcal{S})$. Then following the proofs of Lemma 4.2 and Claims 3–4 in Proposition 4.3 in Zhu and Martínez (2013), we can get that $\Lambda^* \subseteq diag(\mathcal{N}(\Gamma) \times \mathcal{N}(\Gamma))$ if Γ is weakly acyclic.

Remark 2. As shown in Marden et al. (2009) and Zhu and Martínez (2013), when games are weakly acyclic, stochastically stable states are contained in the set of pure Nash equilibrium. To our best knowledge, weakly acyclic games are the most general ones which have such property. When a game is not weakly acyclic, stochastically stable states can still be used to characterize where the algorithm converges. So, stochastically stable states are of broader applicability than pure Nash equilibrium. □

4.2. Estimate of constant C in inequality (1)

The following corollary estimates constant C in inequality (1). For presentation simplicity, denote $|\gamma|_{min} \triangleq \min_{i\in\mathcal{V}} \gamma_i$, $C_{min} \triangleq \min\{(|\gamma|_{min}/|\mathcal{A}|_\infty)^{N|\mathcal{S}|^2}, 1\}$, $C_{max} \triangleq \max\{1, \|\gamma\|_\infty^{N|\mathcal{S}|^2}\}$.

Corollary 2. *If Assumption 1 holds and the exploration rates satisfy that* $\|\tilde{\epsilon}(t)\|_\infty \leq \min\{1/N(N+1)^{|\mathcal{S}|^2}, C_{min}/2(N|\mathcal{S}|^{|\mathcal{S}|^2+4}(N+1)^{|\mathcal{S}|^2} 2^{(N+1)|\mathcal{S}|^2/2}C_{max})\}$ *for all t, then the constant C in inequality (1) can be estimated as* $C = \max\{4^N, 64|\mathcal{S}|^{|\mathcal{S}|^2+4}2^{(N+1)|\mathcal{S}|^2/2}\frac{C_{max}}{C_{min}}\}$.

The proof of Corollary 2 will be given in Section 5.4. Clearly, the constant C increases as N and $|\mathcal{S}|$ increase. In addition, as $|\gamma|_{min}$ decreases, C_{min} decreases and C increases. Similarly, as $\|\gamma\|_\infty$ increases, C_{max} increases and C increases. This indicates that the heterogeneity of the exploration rates could slow down the algorithm.

4.3. Optimal exploration rates

An interesting question is how to choose the exploration rates to minimize the upper bound in inequality (1). This is an infinite-dimension and non-convex optimization problem and hard to solve in general. For analytical tractability, we restrict the exploration rates to be $\mathbf{p}$-series which have been widely used in stochastic approximation and convex optimization (Bertsekas, 2015; Hasminskii & Silver, 1972; Kushner & Yin, 2003). In particular, let $e_i(t) = 0$ and $\epsilon_i(t) = 1/(t^{\mathbf{p}/N}|\mathcal{A}_i|), \mathbf{p} \in (0, 1], \forall i \in \mathcal{V}$. This choice satisfies Assumption 1. We aim to choose $\mathbf{p} \in (0, 1]$ to minimize the upper bound of $D(t)$. With such restriction, inequality (1) becomes (here we ignore the trivial term 2 on the right-hand side of (1)):

$$D(t) \leq C(2\exp(-\sum_{\tau=t^*}^{t-1}\frac{1}{\tau^{\mathbf{p}}}) + \frac{1}{t^{*\mathbf{p}/N}} + \frac{1}{t^{\mathbf{p}/N}})$$
$$= C(2\exp(\sum_{\tau=1}^{t^*-1}\frac{1}{\tau^{\mathbf{p}}} - \sum_{\tau=1}^{t-1}\frac{1}{\tau^{\mathbf{p}}}) + \frac{1}{t^{*\mathbf{p}/N}} + \frac{1}{t^{\mathbf{p}/N}})$$
$$\leq C(2\exp(1 + \int_1^{t^*-1}\frac{1}{x^{\mathbf{p}}}dx - \int_1^{t}\frac{1}{x^{\mathbf{p}}}dx) + \frac{1}{t^{*\mathbf{p}/N}} + \frac{1}{t^{\mathbf{p}/N}}). \quad (2)$$

The second inequality of (2) is a result of inequality (2) of Chlebus (2009). When $\mathbf{p} \in (0, 1)$, inequality (2) becomes $D(t) \leq C(2\exp(1 + \frac{(t^*-1)^{1-\mathbf{p}}}{1-\mathbf{p}})\exp(\frac{-t^{1-\mathbf{p}}}{1-\mathbf{p}}) + 1/(t^{*\mathbf{p}/N}) + 1/(t^{\mathbf{p}/N}))$. Since $\lim_{t\to\infty}(\frac{t^{1-\mathbf{p}}}{1-\mathbf{p}})/(\frac{\mathbf{p}}{N}\ln t) = \infty$, we have $\lim_{t\to\infty}\exp(\frac{-t^{1-\mathbf{p}}}{1-\mathbf{p}})/\exp(-\frac{\mathbf{p}}{N}\ln t) = 0$. So the term $\frac{1}{t^{\mathbf{p}/N}}$ dominates the term $2\exp(1+\frac{(t^*-1)^{1-\mathbf{p}}}{1-\mathbf{p}})\exp(\frac{-t^{1-\mathbf{p}}}{1-\mathbf{p}})$ as t increases. When $\mathbf{p} = 1$, inequality (2) becomes $D(t) \leq C(2\exp(1 + \ln(t^* - 1))\frac{1}{t} + \frac{1}{t^{*1/N}} + \frac{1}{t^{1/N}})$. Analogously, we have $\lim_{t\to\infty}(\frac{1}{t})/(\frac{1}{t^{1/N}}) = 0$. So the term $1/(t^{1/N})$ dominates the term $1/t$ as t increases. In both cases, $1/(t^{\mathbf{p}/N})$ dominates the upper bound in (2). When $\mathbf{p} = 1$, $1/(t^{\mathbf{p}/N})$ decreases fastest among $\mathbf{p} \in (0, 1]$. Therefore, $\epsilon_i(t) = 1/(t^{1/N}|\mathcal{A}_i|)$ is optimal among $\mathbf{p}$-series.

4.4. Explicit convergence rate

If the exploration rates and exploration deviations are given, we can explicitly quantify how fast the algorithm will reach the set Λ^*. Assume the exploration rate for player i is $1/(t^{1/N}|\mathcal{A}_i|)$, and $e_i(t) = 0, \forall i \in \mathcal{V}$. Then we have:

$$D(t) \leq C(2\exp(-\sum_{\tau=t^*}^{t-1}\frac{1}{\tau}) + \frac{1}{t^{*1/N}} + \frac{1}{t^{1/N}})$$
$$\leq C(2\exp(1 - \int_{t^*-1}^{t}\frac{1}{x}dx) + \frac{1}{t^{*1/N}} + \frac{1}{t^{1/N}})$$
$$\leq C(\frac{2e(t^*-1)}{t} + \frac{2}{t^{*1/N}}). \quad (3)$$

The second inequality of (3) follows the same steps of (2) by replacing $\frac{1}{\tau^{\mathbf{p}}}$ with $\frac{1}{\tau}$. Given any $\delta > 0$, $D(t) \leq \delta$ for all $t \geq e(4C)^{N+1}/\delta^{N+1} - 4Ce/\delta$. Roughly speaking, it takes $O(1/\delta^{N+1})$ iterations to reach error δ.

4.5. Memory and communication

The RL algorithm only requires each player to remember its own utility values and actions in recent history. So the memory cost is low. In addition, the communication cost is case dependent. In Zhu and Martínez (2013), the utility function of each robot only depends on the actions of its own and nearby robots. The communication range of each robot is twice of its sensing range. So the communication graphs are time-varying and usually sparse. In Section 6, each customer can communicate with the system operator. So the communication graph is a fixed star graph.

5. Proofs

We will prove Theorem 1 and Corollary 2 in this section.

5.1. Analysis of the H-RL algorithm

For the sake of analysis, we introduce the H-RL algorithm, which has time-homogeneous exploration rates; i.e., $\tilde{\epsilon}(t) = \tilde{\epsilon} \in (0, 1]^N, \forall t \geq 0$ in the RL algorithm. Then $\{z(t)\}$ in the H-RL algorithm forms a time-homogeneous Markov chain $\mathcal{M}^{\tilde{\epsilon}}$ with the transition matrix $P^{\tilde{\epsilon}}$. The analysis of the H-RL algorithm provides preliminary results for that of the RL algorithm. The following lemma studies the properties of the feasible transitions in the Markov chain $\mathcal{M}^{\tilde{\epsilon}}$.

Lemma 1. *Given any* $\tilde{\epsilon} \in (0, 1]^N$, *each nonzero entry in transition matrix* $P^{\tilde{\epsilon}}$ *is a polynomial of the variables* $\{\tilde{\epsilon}_i, 1-\tilde{\epsilon}_i\}_{i\in\mathcal{V}}$. *In addition, the coefficients of the polynomials are independent of* $\tilde{\epsilon}$.

Proof. Consider any two states $x, y \in \mathcal{Z}$ that the transition from x to y is feasible within one step. In particular, $x = (s(0), s(1))$ and $y = (s(1), s(2))$, where $s(t) = (a^1(t), \ldots, a^N(t))$ for $t \in \{0, 1, 2\}$. And the transition probability is $P^{\tilde{\epsilon}}(x, y) = \prod_{i\in\mathcal{V}} Pr\{a^i(2)|a^i(0), a^i(1)\}$.

Given states x and y, the set of players can be partitioned into two sets: $\mathcal{V}_{er}(x, y) \triangleq \{i \in \mathcal{V}|a^i(2) \notin \{a^i(0), a^i(1)\}\}$, $\mathcal{V}_{ex}(x, y) \triangleq \{j \in \mathcal{V}|a^j(2) = a^j(\arg\max_{t\in\{0,1\}}\{u_j(s(t))\})\}$. For any $i \in \mathcal{V}_{er}(x, y)$, $Pr\{a^i(2)|a^i(0), a^i(1)\} = \tilde{\epsilon}_i/|\mathcal{A}_i|$. For any $j \in \mathcal{V}_{ex}(x, y)$, $a^j(2)$ can be achieved by exploitation or exploration, and then $Pr\{a^j(2)|a^j(0), a^j(1)\} = (1-\tilde{\epsilon}_j) + \tilde{\epsilon}_j/|\mathcal{A}_j|$. Then the transition probability can be written as:

$$P^{\tilde{\epsilon}}(x, y) = \prod_{i\in\mathcal{V}_{er}(x,y)}\frac{\tilde{\epsilon}_i}{|\mathcal{A}_i|}\prod_{j\in\mathcal{V}_{ex}(x,y)}((1-\tilde{\epsilon}_j) + \frac{\tilde{\epsilon}_j}{|\mathcal{A}_j|}). \quad (4)$$

It is clear that $P^{\tilde{\epsilon}}(x, y)$ is a polynomial of $\{\tilde{\epsilon}_i, 1-\tilde{\epsilon}_i\}_{i\in\mathcal{V}}$ with coefficients independent of $\tilde{\epsilon}$. □

Given any $\tilde{\epsilon} \in (0, 1]^N$, define stochastic vector $\pi^*(\tilde{\epsilon})$ as the stationary distribution of the Markov chain $\mathcal{M}^{\tilde{\epsilon}}$; i.e., $\pi^*(\tilde{\epsilon})^T P^{\tilde{\epsilon}} = \pi^*(\tilde{\epsilon})^T$. In the H-RL algorithm, when a player performs exploration, it can choose any element in its action set. One can see that, for any pair of states $x, y \in \mathcal{Z}$, y can be reached from x within finite steps, and Markov chain $\mathcal{M}^{\tilde{\epsilon}}$ is ergodic. By Lemma 3.1 in Chapter 6 of Freidlin, Szücs, and Wentzell (2012), $\pi^*(\tilde{\epsilon})$ can be written as follows:

$$\pi^*(\tilde{\epsilon}) \triangleq \left[\pi^*_{z_1}(\tilde{\epsilon}) \quad \cdots \quad \pi^*_{z_{|\mathcal{Z}|}}(\tilde{\epsilon})\right]^T, \quad (5)$$

where $\pi^*_z(\tilde{\epsilon}) = \frac{\sigma_z(\tilde{\epsilon})}{\sum_{z'\in\mathcal{Z}}\sigma_{z'}(\tilde{\epsilon})}$, $\sigma_z(\tilde{\epsilon}) = \sum_{T\in G_{\tilde{\epsilon}}(z)}\prod_{(z',z)\in E(T)}P^{\tilde{\epsilon}}(z', z)$ and $E(T)$ is the edge set of tree T.

5.2. Stationary distributions without exploration deviations

Now let us consider an auxiliary scenario where $e_i(t) = 0$ for all t and for all $i \in \mathcal{V}$. Then stochastic vector $\hat{\pi}^*(\epsilon(t))$ such that $\hat{\pi}^*(\epsilon(t))^T P^{\epsilon(t)} = \hat{\pi}^*(\epsilon(t))^T$ has the same form of (5) with exploration deviations being 0. For notational simplicity, we refer to $\hat{\pi}^*(\epsilon(t))$ as $\hat{\pi}^*(t)$. The following lemma shows that $\{\hat{\pi}^*(t)\}_{t\geq 0}$ converges to a limiting distribution with a certain rate, and the support of the limiting distribution is Λ^*.

Lemma 2. *If Assumption 1 holds and $e_i(t) = 0$ for all t and all i, then the sequence $\{\hat{\pi}^*(t)\}_{t\geq 0}$ converges to the limiting distribution π^* whose support is $\Lambda^* \subseteq diag(\mathcal{S}\times\mathcal{S})$. Moreover, the convergence rate could be quantified as: $\left\|\hat{\pi}^*(t) - \pi^*\right\| \leq C_\epsilon \|\epsilon(t)\|_\infty$ for some constant $C_\epsilon > 0$.*

Proof. The proof is divided into two claims.

Claim 1. *The limiting distribution $\pi^* \triangleq \lim_{t\to\infty} \hat{\pi}^*(t)$ exists, and its support is $\Lambda^* \subseteq diag(\mathcal{S}\times\mathcal{S})$.*

Proof. By Lemma 1, for any $\epsilon(t) \in (0,1]^N$, the non-zero entries of $P^{\epsilon(t)}$ are polynomials of $\{\epsilon_i(t), 1-\epsilon_i(t)\}$ since $e_i(t) = 0$ for all $i \in \mathcal{V}$ with time-homogeneous coefficients. Then $\sigma_z(\epsilon(t))$ and $\sum_{z'\in\mathcal{Z}} \sigma_{z'}(\epsilon(t))$ are polynomials of $\{\epsilon_i(t), 1-\epsilon_i(t)\}$ with time-homogeneous coefficients. Recall that $\epsilon_i(t) = \gamma_i \epsilon^c(t)$. For particular state $z \in \mathcal{Z}$, $\sigma_z(\epsilon(t))$ and $\sum_{z'\in\mathcal{Z}} \sigma_{z'}(\epsilon(t))$ are polynomials of $\epsilon^c(t)$, and $\hat{\pi}^*_z(\epsilon(t))$ is a ratio of two polynomials of $\epsilon^c(t)$:

$$\hat{\pi}^*_z(\epsilon(t)) = \frac{\alpha_z(\epsilon^c(t))}{\beta(\epsilon^c(t))}, \tag{6}$$

where $\alpha_z(\epsilon^c(t)) = b^z_k\epsilon^c(t)^k + b^z_{k+1}\epsilon^c(t)^{k+1} + \cdots + b^z_h\epsilon^c(t)^h$, $\beta(\epsilon^c(t)) = b_k\epsilon^c(t)^k + b_{k+1}\epsilon^c(t)^{k+1} + \cdots + b_h\epsilon^c(t)^h$ and $k \geq 0$. Without loss of generality, we assume that of b_k is non-zero. When $\epsilon^c(t)$ is sufficiently small, $b^z_k\epsilon^c(t)^k$ and $b_k\epsilon^c(t)^k$ dominate $\alpha_z(\epsilon^c(t))/\beta(\epsilon^c(t))$. Then the limit of the $\hat{\pi}^*_z(\epsilon(t))$ can be represented as $\pi^*_z = \lim_{t\to\infty}\hat{\pi}^*_z(\epsilon(t)) = b^z_k/b_k$. Note that b^z_k and b_k are also time-homogeneous.

By the definition of Λ^*, if b^z_k is non-zero, then $z \in \Lambda^*$; And if $b^z_k = 0$, then $z \notin \Lambda^*$. We know $\pi^*_z = b^z_k/b_k$, therefore the support of π^* is contained in Λ^*.

For any $\epsilon(t) \in (0,1]^N$ Assume a state $z = (s(0), s(1)) \in \Lambda(\epsilon(t))$ but $z \notin diag(\mathcal{S}\times\mathcal{S})$, i.e., $s(0) \neq s(1)$, where $s(0) = (a^1(0), \ldots, a^N(0)), s(1) = (a^1(1), \ldots, a^N(1))$. Since $z \in \Lambda(\epsilon(t))$, then there is a tree $T_{max}(\epsilon(t))$ rooted at z such that it has largest total probability. We construct a tree T' by adding the following path from z to $z' = (s(2), s(2))$ through $\hat{z} = (s(1), s(2))$: $z \xrightarrow{P^{\epsilon(t)}(z,\hat{z})} \hat{z} \xrightarrow{P^{\epsilon(t)}(\hat{z},z')} z'$, where $s(2) = (a^1(2), \ldots, a^N(2))$ and $a^i(2) = a^i(\arg\max_{\tau\in\{0,1\}}\{u_i(s(\tau))\})$, $\forall i$. By Lemma 1, $\mathcal{V}_{ex}(z,\hat{z}) = \mathcal{V}_{ex}(\hat{z},z') = \mathcal{V}, \mathcal{V}_{er}(z,\hat{z}) = \mathcal{V}_{er}(\hat{z},z') = \emptyset$. So $P^{\epsilon(t)}(z,\hat{z}) = \prod_{i\in\mathcal{V}}((1-\epsilon_i(t)) + \epsilon_i(t)/|\mathcal{A}_i|)$ and $P^{\epsilon(t)}(\hat{z},z') = \prod_{i\in\mathcal{V}}((1-\epsilon_i(t)) + \epsilon_i(t)/|\mathcal{A}_i|)$.

Let us consider the edge leaving z': $z' \xrightarrow{P^{\epsilon(t)}(z',z'')} \hat{z}'' = (s(2), s(3))$, where $s(3) = (a^1(3), \ldots, a^N(3))$ with at least one player i such that $a^i(3) \neq a^i(2)$. That is $|\mathcal{V}_{er}(z',z'')| \geq 1$. Then the transition probability of the leaving edge satisfies:

$$P^{\epsilon(t)}(z',z'') = \prod_{j\in\mathcal{V}_{er}(z',z'')} \frac{\epsilon_j(t)}{|\mathcal{A}_j|} \prod_{i\in\mathcal{V}_{ex}(z',z'')} ((1-\epsilon_i(t)) + \frac{\epsilon_i(t)}{|\mathcal{A}_i|}).$$

Then $P^{\epsilon(t)}(z,\hat{z})$, $P^{\epsilon(t)}(\hat{z},z')$ and $P^{\epsilon(t)}(z',z'')$ are dominated by their lowest degree terms when $\epsilon(t)$ is sufficiently small. In particular, the lowest degree terms of $P^{\epsilon(t)}(z,\hat{z})$ and $P^{\epsilon(t)}(\hat{z},z')$ are constant terms while the lowest degree term $P^{\epsilon(t)}(z',z'')$ is at least first-degree term. Then there exists some $\epsilon_M \in (0,1]$ such that $P^{\epsilon(t)}(z,\hat{z})P^{\epsilon(t)}(\hat{z},z') > P^{\epsilon(t)}(z',z''), \forall\epsilon(t) \in (0,\epsilon_M)$. That is, the total probability of T' is larger than that of $T_{max}(\epsilon(t))$. We reach a contradiction. Therefore, $\Lambda^* \subseteq diag(\mathcal{S}\times\mathcal{S})$ because $\Lambda(\epsilon(t)) \subseteq diag(\mathcal{S}\times\mathcal{S})$ holds for any sufficiently small $\epsilon(t)$. □

Claim 2. *$\left\|\hat{\pi}^*(t) - \pi^*\right\| \leq C_\epsilon\|\epsilon(t)\|_\infty$ for some constant $C_\epsilon > 0$.*

Proof. From Claim 1, we have $\lim_{t\to\infty}\hat{\pi}^*(t) = \pi^*$ and the support of π^* is Λ^*. Now study the convergence rate.

$$\begin{aligned}
&\left\|\hat{\pi}^*(t) - \pi^*\right\| = \sum_{z\in\mathcal{Z}} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| \\
&= \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| + \sum_{z\notin\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| \\
&= \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| + \sum_{z\notin\Lambda^*} \hat{\pi}^*_z(\epsilon(t)) \\
&= \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| + 1 - \sum_{z\in\Lambda^*} \hat{\pi}^*_z(\epsilon(t)) \\
&= \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| + \sum_{z\in\Lambda^*} \pi^*_z - \sum_{z\in\Lambda^*} \hat{\pi}^*_z(\epsilon(t)) \\
&\leq \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| + \sum_{z\in\Lambda^*} |\hat{\pi}^*_z(\epsilon(t)) - \pi^*_z| \\
&= 2\sum_{z\in\Lambda^*} \left| \frac{b^z_k\epsilon^c(t)^k + \cdots + b^z_h\epsilon^c(t)^h}{b_k\epsilon^c(t)^k + \cdots + b_h\epsilon^c(t)^h} - \frac{b^z_k}{b_k} \right| \\
&= 2\sum_{z\in\Lambda^*} \left| \frac{L_z(\epsilon^c(t)^{k+1}, \ldots, \epsilon^c(t)^h)}{L(\epsilon^c(t)^k, \ldots, \epsilon^c(t)^h)} \right| \\
&= 2\epsilon^c(t) \sum_{z\in\Lambda^*} \left| \frac{L_z(1, \epsilon^c(t), \ldots, \epsilon^c(t)^{h-k-1})}{L(1, \epsilon^c(t), \ldots, \epsilon^c(t)^{h-k})} \right|,
\end{aligned} \tag{7}$$

where L_z and L are linear functions, the constant term of L is non-zero. And by Assumption 1-(1), for any $z \in \Lambda^*$, $L_z(1, \epsilon^c(t), \epsilon_a(t)^{h-k})$ and $L(1, \epsilon^c(t), \epsilon^c(t)^{h-k-1})$ converge as $t \to \infty$ because $\lim_{t\to\infty}\epsilon^c(t) = 0$. Also because Λ^* contains finite elements, then $2\sum_{z\in\Lambda^*}\left|\frac{L_z(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k-1})}{L(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k})}\right|$ is uniformly bounded. And we can always choose a constant $C_\epsilon \geq 2\sum_{z\in\Lambda^*}\left|\frac{L_z(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k-1})}{L(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k})}\right|$ such that $\|\hat{\pi}^*(t) - \pi^*\| \leq C_\epsilon\|\epsilon(t)\|_\infty$. It completes the proof of Claim 2. □

The following lemma shows that the sequence $\{\|\hat{\pi}^*(t) - \hat{\pi}^*(t+1)\|\}_{t\geq 0}$ is summable and gives the explicit partial sums of the sequence when t is large.

Lemma 3. *If Assumption 1 holds and $e_i(t) = 0$ for all t and all i, then $\sum_{\tau=0}^{+\infty}\left\|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\right\| < +\infty$. Moreover, there exists t^{i_0} such that $\sum_{\tau=t^i}^{t}\|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\| \leq 2\|\hat{\pi}^*(t^i) - \pi^*\| + 2\|\hat{\pi}^*(t) - \pi^*\|$, $\forall t^i > t^{i_0}$ and $\forall t > t^i$.*

Proof. In this paper, for any vector, we choose the L^1-norm, then $\sum_{\tau=0}^{+\infty}\left\|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\right\| = \sum_{\tau=0}^{+\infty}\sum_{z\in\mathcal{Z}}|\hat{\pi}^*_z(\epsilon(\tau)) - \hat{\pi}^*_z(\epsilon(\tau+1))|$. Then adapting the proofs of Claim 6 in Zhu and Martínez (2013), we can get that $\sum_{\tau=0}^{+\infty}\sum_{z\in\mathcal{Z}}|\hat{\pi}^*_z(\epsilon(\tau)) - \hat{\pi}^*_z(\epsilon(\tau+1))| < +\infty$. And there exists t_{i_0} such that the partial sum $\sum_{\tau=t^i}^{t}\left\|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\right\|$ with $t^i > t^{i_0}$ and $t > t^i$ satisfies $\sum_{\tau=t^i}^{t}\|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\| \leq 2\left\|\hat{\pi}^*(t^i) - \pi^*\right\| + 2\left\|\hat{\pi}^*(t) - \pi^*\right\|$. □

5.3. Proof of Theorem 1

Lemma 2 shows that $\hat{\pi}^*(t) \to \pi^*$ whose support is Λ^*. Now we proceed to finish the proofs of Theorem 1 by showing $\pi(t) \to \pi^*$ and quantifying its convergence rate.

Proof. For any $t \geq 2$, it holds that,

$$\left\|\pi(t) - \pi^*\right\| \leq \|\pi(t) - \hat{\pi}(t)\| + \|\hat{\pi}(t) - \pi^*\|. \tag{8}$$

We want to prove that the two terms on the right-hand side of (8) converge to 0 with certain rates.

Claim 3. $\lim_{t\to\infty} \left\|\hat{\pi}(t) - \pi^*\right\| = 0$ *and there exists some* $t^\vee$ *such that for any* $t_3^* > t^\vee$ *and* $t > t_3^* + 1$, $\left\|\hat{\pi}(t) - \pi^*\right\| \leq C_q \exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) + 4C_\epsilon \|\epsilon(t_3^*)\|_\infty + C_\epsilon \|\epsilon(t)\|_\infty$ *for some constants* $C_\epsilon, C_q > 0$.

Proof. For any $t \geq 2$, it holds that,

$$\left\|\hat{\pi}(t) - \pi^*\right\| \leq \left\|\hat{\pi}(t) - \hat{\pi}^*(t)\right\| + \left\|\hat{\pi}^*(t) - \pi^*\right\|, \tag{9}$$

where $\hat{\pi}(t)$ is the distribution on $\mathcal{Z}$ at t when the exploration deviations $e_i(t) = 0$ for all $i \in \mathcal{V}$. Let $x(t) \triangleq \left\|\hat{\pi}(t) - \hat{\pi}^*(t)\right\|$ and $y(t) \triangleq \left\|\hat{\pi}^*(t) - \pi^*\right\|$.

Let us first consider $x(t)$. Note that $\hat{\pi}^*(t)^T P^{\epsilon(t)} = \hat{\pi}^*(t)^T$. Then we have:

$$\begin{aligned} x(t) &= \left\|\hat{\pi}(t) - \hat{\pi}^*(t)\right\| \\ &= \left\|\hat{\pi}(t) - \hat{\pi}^*(t-1) + \hat{\pi}^*(t-1) - \hat{\pi}^*(t)\right\| \\ &\leq \|\{P^{\epsilon(t-1)}\}^T \hat{\pi}(t-1) - \{P^{\epsilon(t-1)}\}^T \hat{\pi}^*(t-1)\| \\ &+ \left\|\hat{\pi}^*(t-1) - \hat{\pi}^*(t)\right\|. \end{aligned} \tag{10}$$

By (4) in the proof of Lemma 1, the nonzero entries in $\{P^{\epsilon(t-1)}\}^T$ can be represented as polynomials of $\{\epsilon_i(t-1), 1 - \epsilon_i(t-1)\}$. Taking the nonzero entry $\prod_{i=1}^N \epsilon_i(t-1)/|\mathcal{A}_i|$, we can decompose $\{P^{\epsilon(t-1)}\}^T$ into $\{P^{\epsilon(t-1)}\}^T = (\prod_{i=1}^N \epsilon_i(t-1)/|\mathcal{A}_i|)Q + R(t-1)$, where Q is a $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix with all entries that are 1. Because $P^{\epsilon(t-1)}$ is a transition matrix, then $\{P^{\epsilon(t-1)}\}^T$ is a column stochastic matrix where the sum of each column is the same and is equal to 1. It follows that the sum of each column in $(\prod_{i=1}^N \epsilon_i(t-1)/|\mathcal{A}_i|)Q$ equals $(\prod_{i=1}^N \epsilon_i(t-1)/|\mathcal{A}_i|)|\mathcal{Z}| = \prod_{i=1}^N \epsilon_i(t-1)|\mathcal{A}_i|$ since $|\mathcal{Z}| = (\prod_{i=1}^N |\mathcal{A}_i|)^2$, and the sum of each column in $R(t-1)$ equals $c(t-1) = 1 - \prod_{i=1}^N \epsilon_i(t-1)|\mathcal{A}_i|$.

By (1) in Assumption 1, $\prod_{i=1}^N \epsilon_i(t-1)|\mathcal{A}_i|$ strictly decreases to 0. Then there exists a $t^{|\mathcal{A}|}$ such that $\prod_{i=1}^N \epsilon_i(t-1)|\mathcal{A}_i| < 1$ for all $t \geq t^{|\mathcal{A}|}$, which implies $0 < c(t-1) < 1$ for all $t \geq t^{|\mathcal{A}|}$ and the column sums of $c(t-1)^{-1}R(t-1)$ equal 1. Let $v(t-1) \triangleq \left\|\hat{\pi}^*(t-1) - \hat{\pi}^*(t)\right\|$. And consider $t \geq t^{|\mathcal{A}|}$, then inequality (10) becomes:

$$\begin{aligned} x(t) &\leq \|\{P^{\epsilon(t-1)}\}^T(\hat{\pi}(t-1) - \hat{\pi}^*(t-1))\| + v(t-1) \\ &= \| \prod_{i=1}^N \frac{\epsilon_i(t-1)}{|\mathcal{A}_i|} Q(\hat{\pi}(t-1) - \hat{\pi}^*(t-1)) \\ &+ R(t-1)(\hat{\pi}(t-1) - \hat{\pi}^*(t-1)) \| + v(t-1), \end{aligned} \tag{11}$$

where $\hat{\pi}(t-1)$ and $\hat{\pi}^*(t-1)$ are both stochastic vectors whose sum of elements is equal to one. And by the construction of Q, we have $Q(\hat{\pi}(t-1) - \hat{\pi}^*(t-1)) = 0$. Then inequality (11) becomes:

$$\begin{aligned} x(t) &\leq c(t-1) \parallel c(t-1)^{-1}R(t-1) \\ &\times (\hat{\pi}(t-1) - \hat{\pi}^*(t-1)) \parallel + v(t-1) \\ &\leq c(t-1)\|c(t-1)^{-1}R(t-1)\| \\ &\times \|\hat{\pi}(t-1) - \hat{\pi}^*(t-1)\| + v(t-1) \\ &= c(t-1)x(t-1) + v(t-1), \end{aligned}$$

where $c(t-1) \in (0,1)$ for all $t \geq t^{|\mathcal{A}|}$. With inequality $\log(1-x) < -x$, $\forall x \in (0,1)$, for any $t > t^{|\mathcal{A}|}$ and $t_3^* \geq t^{|\mathcal{A}|}$, we have :

$$\begin{aligned} x(t) &\leq c(t-1)x(t-1) + v(t-1) \\ &\leq \prod_{\tau=t_3^*}^{t-1} c(\tau)x(t_3^*) + v(t-1) + \sum_{\tau=t_3^*}^{t-2}(\prod_{i=\tau+1}^{t-1} c(i)v(\tau)) \\ &\leq x(t_3^*) \prod_{\tau=t_3^*}^{t-1} \exp(-(1-c(\tau))) + v(t-1) \\ &+ \sum_{\tau=t_3^*}^{t-2}(\prod_{i=\tau+1}^{t-1} \exp\left(-(1-c(i))\right) v(\tau)) \\ &\leq x(t_3^*) \exp(-\sum_{\tau=t_3^*}^{t-1}(1-c(\tau))) + \sum_{\tau=t_3^*}^{t-1} v(\tau). \end{aligned} \tag{12}$$

Note that inequality (12) holds for any $t_3^* \geq t^{|\mathcal{A}|}$. And by Lemma 3, $v(\tau)$ is summable, we first take the limit of t and then take the limit of t_3^*, by the summability of $v(\tau)$, we can have $\lim_{t_3^*\to\infty} \lim_{t\to\infty} \sum_{\tau=t_3^*}^{t-1} v(\tau) = 0$. And for any $t^i > t^{i_0}$ and $t > t^i$, $\sum_{\tau=t^i}^{t} \|\hat{\pi}^*(\tau) - \hat{\pi}^*(\tau+1)\| \leq 2||\hat{\pi}^*(t^i+1) - \pi^*\| + 2||\hat{\pi}^*(t) - \pi^*\|$.

Let $t^\vee \triangleq \max\{t^{i_0}, t^{|\mathcal{A}|}\}$. Since Inequality (12) holds for any $t_3^* \geq t^{|\mathcal{A}|}$, we can take $t_3^* > t^\vee$. And $1 - c(\tau) = \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|$, for any $t > t_3^*$, inequality (12) becomes:

$$\begin{aligned} x(t) &\leq x(t_3^*) \exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) \\ &+ 2||\hat{\pi}^*(t_3^*+1) - \pi^*\| + 2||\hat{\pi}^*(t-1) - \pi^*\| \\ &= x(t_3^*) \exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) + 2y(t_3^*+1) + 2y(t-1). \end{aligned} \tag{13}$$

Combining inequalities (9) and (13), for any $t_3^* > t^\vee$:

$$\begin{aligned} \left\|\hat{\pi}(t) - \pi^*\right\| &\leq x(t_3^*) \exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) \\ &+ 2y(t_3^*+1) + 2y(t-1) + y(t). \end{aligned} \tag{14}$$

By (2) in Assumption 1, $\prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|$ is not summable. Therefore, for any $t_3^* > t^\vee$, we have $\lim_{t\to\infty} x(t_3^*)$ $\exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) = 0$. By Lemma 2, we have:

$$\begin{aligned} &2y(t_3^*+1) + 2y(t-1) + y(t) \\ &\leq 2C_\epsilon \|\epsilon(t_3^*+1)\|_\infty + 2C_\epsilon|\epsilon(t-1)_\infty + C_\epsilon \|\epsilon(t)\|_\infty \\ &\leq 4C_\epsilon \|\epsilon(t_3^*)\|_\infty + C_\epsilon \|\epsilon(t)\|_\infty, \end{aligned}$$

where $2C_\epsilon \|\epsilon(t_3^*)+1\|_\infty + 2C_\epsilon \|\epsilon(t-1)\|_\infty \leq 4C_\epsilon \|\epsilon(t_3^*)\|_\infty$ since $\epsilon_i(t)$ is strictly decreasing to 0. And there exists a positive constant C_q such that $x(t_3^*) \leq C_q$. Therefore, for any $t_3^* > t^\vee$ and $t > t_3^* + 1$, (14) becomes $\left\|\hat{\pi}(t) - \pi^*\right\| \leq C_q \exp(-\sum_{\tau=t_3^*}^{t-1} \prod_{i=1}^N \epsilon_i(\tau)|\mathcal{A}_i|) + 4C_\epsilon \|\epsilon(t_3^*)\|_\infty + C_\epsilon \|\epsilon(t)\|_\infty$. Therefore we reach Claim 3. □

Claim 4. $\lim_{t\to\infty} \|\pi(t) - \hat{\pi}(t)\| = 0$ *and there exists some* t^c *such that for any* $t_4^* \geq t^c + 1$ *and* $t \geq t_4^*$, $\|\pi(t) - \hat{\pi}(t)\| \leq C_c \exp(-\sum_{\tau=t_4^*}^{t-1} \prod_{i=1}^N \tilde{\epsilon}_i(\tau)|\mathcal{A}_i|) + 4^N e_r(t_4^*)$ *for some constant* $C_c > 0$.

Proof. Based on triangle inequality, for all $t \geq 2$,

$$\begin{aligned}&\left\|\pi(t)-\hat{\pi}(t)\right\| \leq \|\pi(t)-\{P^{\tilde{\epsilon}(t-1)}\}^T\hat{\pi}(t-1)\|\\ &+\|\{P^{\tilde{\epsilon}(t-1)}\}^T\hat{\pi}(t-1)-\hat{\pi}(t)\|. \qquad (15)\end{aligned}$$

With $\pi(t)=\{P^{\tilde{\epsilon}(t-1)}\}^T\pi(t-1)$ and $\hat{\pi}(t)=\{P^{\epsilon(t-1)}\}^T\hat{\pi}(t-1)$, (15) becomes:

$$\begin{aligned}&\left\|\pi(t)-\hat{\pi}(t)\right\| \leq \|\{P^{\tilde{\epsilon}(t-1)}\}^T\left(\pi(t-1)-\hat{\pi}(t-1)\right)\|\\ &+\|(\{P^{\tilde{\epsilon}(t-1)}\}^T-\{P^{\epsilon(t-1)}\}^T)\hat{\pi}(t-1)\|. \qquad (16)\end{aligned}$$

Based on Lemma 1, the nonzero entries in $\{P^{\tilde{\epsilon}(t-1)}\}^T$ can be represented as polynomials of $\{\tilde{\epsilon}_i(t-1), 1-\tilde{\epsilon}_i(t-1)\}$. Taking the nonzero entry $\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i|$, we can decompose $\{P^{\tilde{\epsilon}(t-1)}\}^T$ into the following: $\{P^{\tilde{\epsilon}(t-1)}\}^T = (\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i|)Q + R'(t-1)$, where Q is a $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix with all entries that are 1. Because $P^{\tilde{\epsilon}(t-1)}$ is a transition matrix, then the $\{P^{\tilde{\epsilon}(t-1)}\}^T$ is a column stochastic matrix where each column sum is equal to 1. It follows that the column sums of $(\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i|)Q$ equal $(\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i|)|\mathcal{Z}| = \prod_{i=1}^N \tilde{\epsilon}_i(t-1)|\mathcal{A}_i|$, and the column sums of $R'(t-1)$ equal $c(t-1) = 1-\prod_{i=1}^N \tilde{\epsilon}_i(t-1)|\mathcal{A}_i|$.

Let $x(t) = \left\|\pi(t)-\hat{\pi}(t)\right\|$. And by the structure of Q and the fact that $\pi(t-1)$ and $\hat{\pi}(t-1)$ are both stochastic vectors whose sum of elements is equal to 1, we have $Q(\pi(t-1)-\hat{\pi}(t-1))=0$. Then (16) becomes:

$$\begin{aligned}&x(t) \leq \|R'(t-1)(\pi(t-1)-\hat{\pi}(t-1))\|\\ &+\|(\{P^{\tilde{\epsilon}(t-1)}\}^T-\{P^{\epsilon(t-1)}\}^T)\hat{\pi}(t-1)\|\\ &\leq c(t-1)x(t-1)+\|\{P^{\tilde{\epsilon}(t-1)}\}^T-\{P^{\epsilon(t-1)}\}^T\|. \qquad (17)\end{aligned}$$

By (4) in the proof of Lemma 1, any entry in $P^{\tilde{\epsilon}(t-1)}$ can be represented as a summation of at most 2^N polynomials. And each polynomial is a product of N monomials; e.g., $\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i|$. And the entries in $P^{\epsilon(t-1)}$ have the same form with $\tilde{\epsilon}_i(t-1)=\epsilon_i(t-1)$. Then the difference of any pair of entries $(P^{\tilde{\epsilon}(t-1)}(x,y), P^{\epsilon(t-1)}(x,y))$ has at most 4^N terms (for example, $\prod_{i=1}^N \tilde{\epsilon}_i(t-1)/|\mathcal{A}_i| - \prod_{i=1}^N \epsilon_i(t-1)/|\mathcal{A}_i|$ has 2^N-1 terms). And each term is less than $\prod_{i=1}^N \|e(t-1)\|_\infty/|\mathcal{A}_i|$, where $\|e(t)\|_\infty = \max\{|e_c^1(t)|,\ldots,|e_c^N(t)|\}$. Then any pair of entries $(P^{\tilde{\epsilon}(t-1)}(x,y), P^{\epsilon(t-1)}(x,y))$ satisfy that:

$$|P^{\tilde{\epsilon}(t-1)}(x,y)-P^{\epsilon(t-1)}(x,y)| \leq 4^N\prod_{i=1}^N \|e(t-1)\|_\infty/|\mathcal{A}_i|.$$

Then (17) becomes:

$$x(t) \leq c(t-1)x(t-1)+4^N\|e(t-1)\|_\infty^N\prod_{i=1}^N|\mathcal{A}_i|. \qquad (18)$$

Let $\chi(t) = x(t)-4^Ne_r(t)$, where $e_r(t) = (\|e(t)\|_\infty^N\prod_{i=1}^N|\mathcal{A}_i|)/(1-c(t))$. Then from (18), we can get:

$$\chi(t) \leq c(t-1)\chi(t-1)+4^Ne_r(t-1)-4^Ne_r(t). \qquad (19)$$

Inequality (19) holds for any $t \geq 2$. Recall that $\tilde{\epsilon}_i(t) \in (0,1]^N, \forall i \in \mathcal{V}, \forall t$, then $c(t-1) \leq 1$. By simple algebraic operations, we have $c(t-1) \geq 1-\prod_{i=1}^N(\epsilon_i(t-1)+\|e(t-1)\|_\infty)|\mathcal{A}_i|$. And by (1) and (3) in Assumption 1, $\prod_{i=1}^N(\epsilon_i(t-1)+\|e(t-1)\|_\infty)$ converges to 0. Therefore, there exists a t^c such that $\prod_{i=1}^N(\epsilon_i(t-1)+\|e(t-1)\|_\infty)|\mathcal{A}_i| < 1, \forall t \geq t^c+1$. Then $c(t-1) \in (0,1), \forall t \geq t^c+1$. By manipulating inequality $\log(1-x) < -x, \forall x \in (0,1)$ and $\exp(-x) < 1, \forall x > 0$, (19) can be rewritten for all $t_4^* \geq t^c+1$ and $t \geq t_4^*$ as follows:

$$\begin{aligned}\chi(t) &\leq \prod_{\tau=t_4^*}^{t-1}c(\tau)\chi(t_4^*)+4^Ne_r(t-1)-4^Ne_r(t)\\ &+\sum_{\tau=t_4^*}^{t-2}(\prod_{j=\tau+1}^{t-1}c(j))4^N(e_r(\tau)-e_r(\tau+1))\\ &\leq \chi(t_4^*)\exp(-\sum_{\tau=t_4^*}^{t-1}(1-c(\tau)))+\sum_{\tau=t_4^*}^{t-1}4^N(e_r(\tau)-e_r(\tau+1))\\ &\leq \chi(t_4^*)\exp(-\sum_{\tau=t_4^*}^{t-1}(1-c(\tau)))+4^N(e_r(t_4^*)-e_r(t)).\end{aligned}$$

Plug $\chi(t) = x(t)-4^Ne_r(t)$ and $c(t) = 1-\prod_{i=1}^N\tilde{\epsilon}_i(t)|\mathcal{A}_i|$ in the above inequality, and we have:

$$x(t) \leq \chi(t_4^*)\exp(-\sum_{\tau=t_4^*}^{t-1}\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|)+4^Ne_r(t_4^*).$$

By Assumption 1-(2), $\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|$ is not summable. Therefore, $\lim_{t\to\infty}\chi(t_4^*)\exp(-\sum_{\tau=t_4^*}^{t-1}\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|) = 0$. And by Assumption 1-(3), $\lim_{t_4^*\to\infty}4^Ne_r(t_4^*) = 0$. We first take the limit of t and then take the limit of t_4^*, we can have $\lim_{t_4^*\to\infty}\lim_{t\to\infty}\chi(t_4^*)\exp(-\sum_{\tau=t_4^*}^{t-1}\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|)+4^Ne_r(t_4^*) = 0$. And there exists a positive constant C_c such that $\chi(t_4^*) \leq C_c$. Then for $t_4^* \geq t^c+1$ and $t \geq t_4^*$:

$$\|\pi(t)-\hat{\pi}(t)\| \leq C_c\exp(-\sum_{\tau=t_4^*}^{t-1}\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|)+4^Ne_r(t_4^*).$$

Therefore we reach Claim 4. □

Combining Claims 3 and 4, we get that for Markov chain $\mathcal{M}$, its state distribution $\{\pi(t)\}$ converges to limiting distribution π^*. Moreover, by triangle inequality, Claims 3 and 4, there exists some $t_{min} = \max\{t^\vee, t^c\}$ and $C \geq \max\{C_q, 4C_\epsilon, C_c, 4^N\}$, such that for any $t^* > t_{min}$ and $t > t^*+1$, $D(t) \leq C(\|\epsilon(t^*)\|_\infty + \|\epsilon(t)\|_\infty + e_r(t^*) + \exp(-\sum_{\tau=t^*}^{t-1}\prod_{i=1}^N\epsilon_i(\tau)|\mathcal{A}_i|) + \exp(-\sum_{\tau=t^*}^{t-1}\prod_{i=1}^N\tilde{\epsilon}_i(\tau)|\mathcal{A}_i|))$. And it is easy to get that $\|\pi(t)-\pi^*\| = \sum_{z\in\mathcal{Z}}|\pi_z(\epsilon(t))-\pi_z^*| \leq \sum_{z\in\mathcal{Z}}|\pi_z(\epsilon(t))| + \sum_{z\in\mathcal{Z}}|\pi_z^*| \leq 2$. It completes the proof of Theorem 1. □

Remark 3. The paper (Mitra, Romeo, & Sangiovanni-Vincentelli, 1986) provides convergence rate analysis of strongly ergodic Markov chains. Our analysis is different, and it leads to a tighter upper bound and allows for a larger class of exploration rates. □

5.4. *Proof of Corollary 2*

Proof. From the last two paragraphs of Section 5.3, the constant C in inequality (1) can be estimated as $C \geq \max\{C_q, 4C_\epsilon, C_c, 4^N\}$. Now we will prove that there exists a set of feasible constants $C_q, 4C_\epsilon, C_c$ such that $\max\{4^N, 64|\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|^2 2^{(N+1)|\mathcal{Z}|/2}\frac{C_{max}}{C_{min}}\}$ $\geq \max\{C_q, 4C_\epsilon, C_c, 4^N\}$.

From Claim 3, C_q can be any constant that satisfies $C_q \geq \left\|\hat{\pi}(t_3^*)-\hat{\pi}^*(t_3^*)\right\| = \sum_{z\in\mathcal{Z}}|\hat{\pi}_z(\epsilon(t_3^*))-\hat{\pi}_z^*(\epsilon(t_3^*))|$. Since $\hat{\pi}_z(\epsilon(t_3^*)) \in [0,1]$ and $\hat{\pi}_z^*(\epsilon(t_3^*)) \in [0,1]$, we have $\sum_{z\in\mathcal{Z}}|\hat{\pi}_z(\epsilon(t_3^*))-\hat{\pi}_z^*(\epsilon(t_3^*))| \leq \sum_{z\in\mathcal{Z}}|\hat{\pi}_z(\epsilon(t_3^*))| + \sum_{z\in\mathcal{Z}}|\hat{\pi}_z^*(\epsilon(t_3^*))| = \sum_{z\in\mathcal{Z}}\hat{\pi}_z(\epsilon(t_3^*)) + \sum_{z\in\mathcal{Z}}\hat{\pi}_z^*(\epsilon(t_3^*)) = 2$. Then C_q can be $C_q = 2 \geq \left\|\hat{\pi}(t_3^*)-\hat{\pi}^*(t_3^*)\right\|$. And from Claim 4, C_c can be any constant such that $C_c \geq \chi(t_4^*) = \left\|\pi(t_4^*)-\hat{\pi}(t_4^*)\right\| - 4^Ne_r(t_4^*)$. Here, $\left\|\pi(t_4^*)-\hat{\pi}(t_4^*)\right\| - 4^Ne_r(t_4^*) \leq \left\|\pi(t_4^*)-\hat{\pi}(t_4^*)\right\| \leq \sum_{z\in\mathcal{Z}}|\pi_z(\epsilon(t_4^*)) - \hat{\pi}_z(\epsilon(t_4^*))| \leq 2$. Then C_c can be $C_c = 2$.

From Claim 2, C_ϵ can be any constant that satisfies $C_\epsilon \geq 2\sum_{z\in\Lambda^*}|\frac{L_z(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k-1})}{L(1,\epsilon^c(t),\ldots,\epsilon^c(t)^{h-k})}| = \underline{C_\epsilon}$. And by (7), we have:

$$\underline{C_\epsilon} = 2\sum_{z\in\Lambda^*}\frac{1}{\epsilon^c(t)}\left|\frac{b_k^z\epsilon^c(t)^k+\cdots+b_h^z\epsilon^c(t)^h}{b_k\epsilon^c(t)^k+\cdots+b_h\epsilon^c(t)^h}-\frac{b_k^z}{b_k}\right|$$
$$= 2\sum_{z\in\Lambda^*}\left|\frac{b_k(b_{k+1}^z\epsilon^c(t)^k+\cdots+b_h^z\epsilon^c(t)^{h-1})}{b_k(b_k\epsilon^c(t)^k+\cdots+b_h\epsilon^c(t)^h)}\right.$$
$$\left.-\frac{b_k^z(b_{k+1}\epsilon^c(t)^k+\cdots+b_h\epsilon^c(t)^{h-1})}{b_k(b_k\epsilon^c(t)^k+\cdots+b_h\epsilon^c(t)^h)}\right|$$
$$\leq 2\sum_{z\in\mathcal{Z}}(\left|\frac{|b_{k+1}^z|+\cdots+|b_h^z|\epsilon^c(t)^{h-k-1}}{b_k+\cdots+b_h\epsilon^c(t)^{h-k}}\right|$$
$$+\left|\frac{|b_{k+1}|+\cdots+|b_h|\epsilon^c(t)^{h-k-1}}{b_k+\cdots+b_h\epsilon^c(t)^{h-k}}\right|), \tag{20}$$

where we use $\frac{b_k^z}{b_k} \leq 1$ in the inequality. Based on equations (5) and (6), we have the following relation:

$$b_k^z\epsilon^c(t)^k+\cdots+b_h^z\epsilon^c(t)^h = \sigma_z(\epsilon(t))$$
$$= \sum_{T\in G_{\epsilon(t)}(z)}\prod_{(x,y)\in E(T)}P^{\epsilon(t)}(x,y). \tag{21}$$

By Lemma 1, we have

$$P^{\epsilon(t)}(x,y) = \prod_{i\in\mathcal{V}_{er}(x,y)}\frac{\gamma_i\epsilon^c(t)}{|\mathcal{A}_i|}\times\prod_{j\in\mathcal{V}_{ex}(x,y)}(1+(-\gamma_j+\frac{\gamma_j}{|\mathcal{A}_j|})\epsilon^c(t)). \tag{22}$$

So $P^{\epsilon(t)}(x,y)$ is a polynomial of $\epsilon^c(t)$ and can be written as $P^{\epsilon(t)}(x,y) = d_0(x,y)+d_1(x,y)\epsilon^c(t)+\cdots+d_N(x,y)\epsilon^c(t)^N$. Note that some coefficient $d_m(x,y)$ could be 0. Denote by $\hat{d}(x,y)$ the coefficient of the least degree term in $P^{\epsilon(t)}(x,y)$. In fact, coefficients $b_k^z,\ldots,b_h^z, b_k,\ldots,b_h$ consist of $d_m(x,y)$, where $m\in\{0,\ldots,N\}$. In Claim 5, we will first find an upper bound of $|d_m(x,y)|$ and a lower bound of $|\hat{d}(x,y)|$. In Claims 6 and 7, we will estimate the last sum in inequality (20) by finding an upper bound of $|b_{k+1}^z|+\cdots+|b_h^z|$ and $|b_{k+1}|+\cdots+|b_h|$ and a lower bound of $|b_k+\cdots+b_h\epsilon^c(t)^{h-k}|$.

Claim 5. *For any $x,y \in \mathcal{Z}$ and $m \in \{0,\ldots,N\}$, $|d_m(x,y)| \leq 2^{N/2}\max\{1,\|\gamma\|_\infty^N\}$. And for any $x,y\in\mathcal{Z}$, $\hat{d}(x,y) \geq \min\{(|\gamma|_{min}/|\mathcal{A}|_\infty)^N, 1\}$.*

Proof. Expanding the right-hand side of (22) yields the sum of at most 2^N monomials where each monomial is a product of $\prod_{i\in\mathcal{V}_{er}(x,y)}(\gamma_i\epsilon^c(t))/|\mathcal{A}_i|$, 1 and $(-\gamma_j+\gamma_j/|\mathcal{A}_j|)\epsilon^c(t)$. Then $d_m(x,y) = \sum_{\{\mathcal{V}'(x,y)\subseteq\mathcal{V}_{ex}(x,y)||\mathcal{V}_{er}(x,y)|+|\mathcal{V}'(x,y)|=m\}}(\prod_{i\in\mathcal{V}_{er}(x,y)}\gamma_i/|\mathcal{A}_i|\prod_{j\in\mathcal{V}'(x,y)}(-\gamma_j+\gamma_j/|\mathcal{A}_j|))$. And there are $\binom{|\mathcal{V}_{ex}(x,y)|}{m-|\mathcal{V}_{er}(x,y)|}$ choices of $\mathcal{V}'(x,y)$. Since $|\mathcal{V}_{ex}(x,y)|\leq N$ here we use the upper bound $\binom{|\mathcal{V}_{ex}(x,y)|}{m-|\mathcal{V}_{er}(x,y)|} \leq \binom{N}{\lfloor N/2\rfloor} = 2^N(\prod_{i=1}^N\frac{i}{2})/(\lfloor N/2\rfloor!\lfloor N/2\rfloor!) \leq 2^{N/2}$ for presentation simplicity, where $\lfloor\cdot\rfloor$ is the floor function. Note that $\|\gamma\|_\infty = \max_{i\in\mathcal{V}}\gamma_i$, then $\gamma_i/|\mathcal{A}_i| < \|\gamma\|_\infty$ and $|-\gamma_j+(\gamma_j/|\mathcal{A}_j|)| \leq \|\gamma\|_\infty$. Then for all $m \in \{0,\ldots,N\}$, it holds that $|d_m(x,y)| \leq 2^{N/2}(\max\{1,\|\gamma\|_\infty\})^m \leq 2^{N/2}\max\{1,\|\gamma\|_\infty^N\}$. By equation (22), $\hat{d}(x,y) = \prod_{i\in\mathcal{V}_{er}(x,y)}(\gamma_i/|\mathcal{A}_i|)$ when $\mathcal{V}_{er}(x,y)\neq\emptyset$ or $\hat{d}(x,y) = 1$ when $\mathcal{V}_{er}(x,y) = \emptyset$. Note that, $\hat{d}(x,y) > 0$ for any $x,y\in\mathcal{Z}$. With $\gamma_i/|\mathcal{A}_i| \geq |\gamma|_{min}/|\mathcal{A}|_\infty$, $\hat{d}(x,y) \geq \min\{(|\gamma|_{min}/|\mathcal{A}|_\infty)^N, 1\}$. □

Claim 6. $|b_{k+1}^z|+\cdots+|b_h^z|\epsilon^c(t)^{h-k-1}$ *and* $|b_{k+1}|+\cdots+|b_h|\epsilon^c(t)^{h-k-1}$ *are both upper bounded by* $2|\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|^2 2^{(N+1)|\mathcal{Z}|/2}C_{max}$ *when* $\|\tilde{\epsilon}(t)\|_\infty \leq 1/N(N+1)^{|\mathcal{Z}|}$.

Proof. Notice that

$$\prod_{(x,y)\in E(T)}P^{\epsilon(t)}(x,y) = \prod_{(x,y)\in E(T)}(d_0(x,y)+\cdots+d_N(x,y)\epsilon^c(t)^N)$$
$$= f_0^T+f_1^T\epsilon^c(t)+\cdots+f_{N|T|}^T\epsilon^c(t)^{N|T|}, \tag{23}$$

where $|T|$ is the number of edges of tree T. For the analytical simplicity, denote the enumeration of edges in T as $E(T) = \{g_1,g_2,\ldots,g_{|T|}\}$ and denote by $d_{l_g}(g)$ the coefficient of the l_g-th degree term in the polynomial $P^{\epsilon(t)}(x,y)$, where $l_g \in \{0,\ldots,N\}$. Then $\prod_{(x,y)\in E(T)}(d_0(x,y)+\cdots+d_N(x,y)\epsilon^c(t)^N) = \prod_{g=g_1}^{g_{|T|}}(d_0(g)+\cdots+d_N(g)\epsilon^c(t)^N)$, which can be expanded as the sum of $(N+1)^{|T|}$ monomials where each monomial is in the form of $\prod_{g=g_1}^{g_{|T|}}d_{l_g}(g)\epsilon^c(t)^{l_g}$. Then $f_m^T = \sum_{\{l_{g_1},\ldots,l_{g_{|T|}}|l_{g_1}+\cdots+l_{g_{|T|}}=m\}}\prod_{g=g_1}^{g_{|T|}}d_{l_g}(g)$, where $m\in\{0,\ldots,N|T|\}$. Finding combinations of $(l_{g_1},\ldots,l_{g_{|T|}})$ such that $l_{g_1}+\cdots+l_{g_{|T|}} = m$ can be cast to the problem of obtaining m points on $|T|$ $(N+1)$-sided dice (pages 23–24 in Uspensky (1937)). The number of all possible combinations equals the coefficient of $\epsilon^c(t)^m$ in the polynomial $(\sum_{i=0}^N\epsilon^c(t)^i)^{|T|}$. By generalizing the solution of problem 13 in Uspensky (1937), we can get the coefficient of $\epsilon^c(t)^m$ is $\sum_{i=0}^{\lfloor\frac{m}{N+1}\rfloor}(-1)^i\binom{|T|}{i}\binom{|T|+m-(N+1)i-1}{m-(N+1)i}$. And by multinomial theorem (Section 24.1.2 in Abramowitz and Stegun (1964), the summation of all coefficients in $(\sum_{i=0}^N\epsilon^c(t)^i)^{|T|}$ equals $(N+1)^{|T|}$. Combining Claim 5, we have $|f_m^T| \leq (N+1)^{|T|}(2^{N/2}\max\{1,\|\gamma\|_\infty^N\})^{|T|}$. Note that any tree $T\in G_{\epsilon(t)}(z)$ is a spanning tree of the graph $\mathcal{G}(\epsilon(t))$ where each vertex is a state $z\in\mathcal{Z}$. Then $|T| = |\mathcal{Z}|-1 \leq |\mathcal{Z}|$. Therefore, $|f_m^T| \leq (N+1)^{|\mathcal{Z}|}2^{N|\mathcal{Z}|/2}C_{max}$.

Now we consider the number of spanning trees of the directed graph $\mathcal{G}(\tilde{\epsilon})$ rooted at z; i.e., $|G_{\epsilon(t)}(z)|$. First let us introduce the Laplacian matrix (Biggs, 1993) $\mathcal{L}(\mathcal{G})$ of $\mathcal{G}(\tilde{\epsilon})$. Formally, $\mathcal{L}(\mathcal{G}) \triangleq \mathfrak{D}-\mathfrak{A}$, where $\mathfrak{D} = diag(d_{z_1},\ldots,d_{z_{|\mathcal{Z}|}})$ such that d_z is the out degree of vertex z; i.e., $d_z = |\{z'\in\mathcal{Z}|(z,z')\in E(\mathcal{G})\}|$. Similar to $E(T)$, $E(\mathcal{G})$ is the edges of graph $\mathcal{G}(\tilde{\epsilon})$. And $\mathfrak{A}$ is the adjacency matrix (Biggs, 1993) (a (0,1)-matrix with ones at places corresponding to entries where the vertices are adjacent and zeros otherwise) of $\mathcal{G}(\tilde{\epsilon})$. Then we define $\mathcal{L}(\mathcal{G})_z$ as the matrix by removing the zth row and column from $\mathcal{L}(\mathcal{G})_z$. By Tutte's Matrix-Tree Theorem (Tutte & Nash-Williams, 2001), $|G_{\epsilon(t)}(z)| = det(\mathcal{L}(\mathcal{G})_z)$. For the sake of analysis, we use l_i represent the ith column of $\mathcal{L}(\mathcal{G})_z$. And by Hadamard's inequality (Bjelica, 1995), $det(\mathcal{L}(\mathcal{G})_z) \leq \prod_{i=1}^{|\mathcal{Z}|-1}\|l_i\|_2$. Recall that each vertex in the graph is a state and each edge in the graph is a transition from one state to another. Then by the RL algorithm, for any state $z = (s,s')$, it has $|\mathcal{S}|$ out degree; i.e., $d_z = |\mathcal{S}|, \forall z\in\mathcal{Z}$. And the number of ones in one column is at most $|\mathcal{S}|$. Then, $\|l_i\|_2 \leq \sqrt{2|\mathcal{S}|^2}$. And $|G_{\epsilon(t)}(z)| \leq (\sqrt{2}|\mathcal{S}|)^{|\mathcal{Z}|-1} \leq (\sqrt{2}|\mathcal{S}|)^{|\mathcal{Z}|}$. For all $m\in\{k,\ldots,h\}$, it holds that $|b_m^z| \leq \sum_{T\in G_{\epsilon(t)}(z)}|f_m^T| \leq |\mathcal{S}|^{|\mathcal{Z}|}(N+1)^{|\mathcal{Z}|}2^{(N+1)|\mathcal{Z}|/2}C_{max}$.

Denote by $\tilde{d}(x,y)$ the coefficient of the second least degree term in $P^{\epsilon(t)}(x,y)$, then $f_{k+1}^T = \sum_{\{\hat{E}(T)\subseteq E(T)||\hat{E}(T)|=|T|-1\}}\prod_{g\in\hat{E}(T)}\hat{d}(g)\prod_{g'\in E(T)\setminus\hat{E}(T)}\tilde{d}(g')$. And there are $|T|$ choices of $\hat{E}(T)$. So $|f_{k+1}^T| \leq |T|2^{N|\mathcal{Z}|/2}C_{max}$ and $|b_{k+1}^z| \leq |\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|2^{(N+1)|\mathcal{Z}|/2}C_{max}$.

Based on equations (5), (6) and (21), we have:

$$b_k\epsilon^c(t)^k+\cdots+b_h\epsilon^c(t)^h = \sum_{z\in\mathcal{Z}}\sigma_z(\epsilon(t))$$
$$= \sum_{z\in\mathcal{Z}}b_k^z\epsilon^c(t)^k+\cdots+\sum_{z\in\mathcal{Z}}b_h^z\epsilon^c(t)^h. \tag{24}$$

Therefore, $|b_m| \leq \sum_{z\in\mathcal{Z}}|b_m^z| \leq |\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|(N+1)^{|\mathcal{Z}|}2^{(N+1)|\mathcal{Z}|/2}C_{max}$, $\forall m\in\{k,\ldots,h\}$ and $|b_{k+1}| \leq \sum_{z\in\mathcal{Z}}|b_{k+1}^z| \leq |\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|^2 2^{(N+1)|\mathcal{Z}|/2}$

C_{max}. By equation (23) and $|T| \leq |\mathcal{Z}|$, the largest degree of $\prod_{(x,y)\in E(T)} P^{\epsilon(t)}(x,y)$ is less than $N|\mathcal{Z}|$, hence $h \leq N|\mathcal{Z}|$. Then $|b^z_{k+2}|\epsilon^c(t)+\cdots+|b^z_h|\epsilon^c(t)^{h-k-1}$ and $|b_{k+2}|\epsilon^c(t)+\cdots+|b_h|\epsilon^c(t)^{h-k-1}$ have less than $N|\mathcal{Z}|$ terms. When $\|\tilde{\epsilon}(t)\|_\infty \leq 1/N(N+1)^{|\mathcal{Z}|}$, $|b^z_{k+2}|\epsilon^c(t)+\cdots+|b^z_h|\epsilon^c(t)^{h-k-1} \leq |\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|^2 2^{(N+1)|\mathcal{Z}|/2}C_{max}$. Similarly, $|b_{k+2}|\epsilon^c(t)+\cdots+|b_h|\epsilon^c(t)^{h-k-1} \leq |\mathcal{S}|^{|\mathcal{Z}|}|\mathcal{Z}|^2 2^{(N+1)|\mathcal{Z}|/2}C_{max}$. Then we reach Claim 6. □

Claim 7. $|b_k + b_{k+1}\epsilon^c(t)+\cdots+b_h\epsilon^c(t)^{h-k}| \geq \frac{1}{2}C_{min}$ *when* $\|\tilde{\epsilon}(t)\|_\infty \leq C_{min}/2(N|\mathcal{Z}|^2|\mathcal{S}|^{|\mathcal{Z}|}(N+1)^{|\mathcal{Z}|}2^{(N+1)|\mathcal{Z}|/2}C_{max})$.

Proof. Now let us consider the magnitude of the coefficient of the least degree term in $\sigma_z(\epsilon(t))$; i.e., $|b^z_k|$. Denote by $\hat{f}^T$ the coefficient of the least degree term in $\prod_{(x,y)\in E(T)} P^{\epsilon(t)}(x,y)$. From equation (21), $\hat{f}^T = \prod_{(x,y)\in E(T)} \hat{d}(x,y)$. By Claim 5, for any $T \in G_{\epsilon(t)}(z)$, $\hat{f}^T \geq \min\{\prod_{(x,y)\in E(T)}(|\gamma|_{min}/|\mathcal{A}|_\infty)^N, 1\} \geq C_{min}$ because $|T| \leq |\mathcal{Z}|$. Then $b^z_k \geq \hat{f}^T \geq C_{min}$.

From equation (24), we have $b_k = \sum_{z\in\mathcal{Z}} b^z_k$. Since b^z_k is positive, $b_k \geq b^z_k$. When $\|\tilde{\epsilon}(t)\|_\infty \leq C_{min}/2(N|\mathcal{Z}|^2|\mathcal{S}|^{|\mathcal{Z}|}(N+1)^{|\mathcal{Z}|}2^{(N+1)|\mathcal{Z}|/2}$
$C_{max})$, $|b_{k+1}\epsilon^c(t)+\cdots+b_h\epsilon^c(t)^{h-k}| \leq C_{min}/2$ and $|b_k+\cdots+b_h\epsilon^c(t)^{h-k}| \geq C_{min}/2$. □

Combining Claims 6 and 7, (20) becomes $\underline{C_\epsilon} \leq C_\epsilon$. And with $|\mathcal{Z}| = |\mathcal{S}|^2$, we can reach Corollary 2. □

6. Case study

We will evaluate the RL algorithm using the application of demand allocation market in Zhu (2014).

6.1. System components

Customers. We consider N customers $\mathcal{V} = \{1,\ldots,N\}$ and each customer $i \in \mathcal{V}$ has power demands $x_i \geq 0$ and wants to allocate its demands in one time slot within $\mathcal{A}_i = \{1,2,\ldots,|\mathcal{A}_i|\}$. The action $a^i \in \mathcal{A}_i$ is the time slot chosen by customer *i*. Each customer wants to satisfy its demands as soon as possible so it punishes late allocation. The cost function $c_i : \mathcal{A}_i \to \mathbb{R}$ is not decreasing; i.e., $c_i(a^i) \leq c_i(\hat{a}^i)$ if $\hat{a}^i > a^i$.

System operator. The system operator charges each customer some price based on demand distributions. In particular, given an action profile $s = (a^1,\ldots,a^N)$, the total demand allocated in time slot a^i is $\Xi_{a^i}(s) \triangleq \sum_{j\in\mathcal{V}} \mathbf{1}_{\{a^j=a^i\}}x_j$, where $\mathbf{1}_{\{\Pi\}}$ is an indicator function: $\mathbf{1}_{\{\Pi\}} = 1$ if Π is true and $\mathbf{1}_{\{\Pi\}} = 0$ if Π is false. The system operator charges customer i the price $p_a(\Xi_{a^i}(s))$.

Utility. The utility of customer i is the negative of the cost and price: $u_i(s) = -c_i(a^i) - p_a(\Xi_{a^i}(s))$.

Informational constraint. Each customer is unwilling to share its cost function c_i and private action a^i with other customers and the system operator. And the system operator does not want to disclose the pricing policy to the customers and only agrees to publicize the price value $p_a(s)$ given s. Therefore, each customer only knows its own utility values instead of the structure of the utility function.

6.2. Evaluation

Evaluation setup. In this section, we use Matlab simulations to evaluate the performance of the RL algorithm. Similar to the setup in Zhu (2014), we consider 100 customers and they have identical action sets consisting of 10 time slots. The demands of all customers are 1; i.e., $x_i = 1$ for any $i \in \mathcal{V}$. The cost function for customer i is set as $c_i(a^i) = 3^{a^i}$. And the pricing mechanism is $p_a(\Xi_{a^i}(s)) = \Xi_{a^i}(s)$.

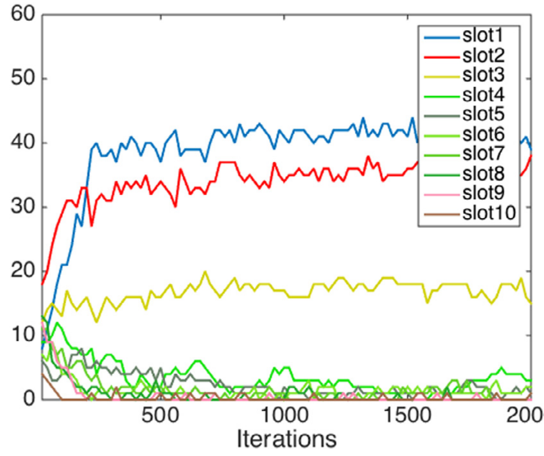


Fig. 1. Temporal aggregate demands allocated at ten time slots with diminishing exploration rate $\epsilon_i(t) = \frac{1}{10}t^{-\frac{1}{100}}$.

Nash equilibrium. By Lemma 2.1 in Zhu (2014), we know that the demand allocation game under the above setup is a potential game, and then a weakly acyclic game (Monderer & Shapley, 1996). Therefore the existence of pure Nash equilibrium is guaranteed.

Simulation results with diminishing exploration rates. Based on the evaluation setup, we simulate the interactions of the customers and system operator in Matlab. The exploration rates are chosen as $\epsilon_i(t) = \frac{1}{10}t^{-\frac{1}{100}}$ and the exploration deviations are chosen as $e_i(t) = 9/(10t^2)$ for all $i \in \mathcal{V}$. The duration of the simulation is 2000 iterations. From the above simulation, we observe that the action profiles converge to s_* in which there are 43 customers selecting slot 1, 37 customers selecting slot 2 and 20 customers selecting slot 3. The induced utility value is −46 for those choosing slot 1 and slot 2. Also, the induced utility value is −47 for those choosing slot 3. Moreover, no customer can benefit by unilateral deviations from s_*. From Definition 1, s_* is a pure Nash equilibrium. The simulation results in Fig. 1 confirm the convergence of the action profiles in Theorem 1.

We now proceed to use simulations to verify the optimal exploration rates. As discussed in Section 4.3, we restrict the exploration rates to be **p**-series. In particular, Fig. 2 compares the convergence of the RL algorithm for three cases $\epsilon_i(t) = \frac{1}{10}t^{-\frac{0.25}{100}}$, $\epsilon_i(t) = \frac{1}{10}t^{-\frac{0.5}{100}}$ and $\epsilon_i(t) = \frac{1}{10}t^{-\frac{1}{100}}$ (the optimal one), respectively. For ease of comparison, we only compare the temporal aggregate demands allocated at time slot 1 and only focus on the first 750 iterations. When the exploration rates are $\epsilon_i(t) = \frac{1}{10}t^{-\frac{1}{100}}$, the convergence is fastest. It is consistent with the discussion in Section 4.3.

Simulation results with measurement noises. In this part, we assume the utility values received by player i are subject to measurement noises; i.e., $\tilde{u}_i(t) \triangleq u_i(t) + w_i(t)$, where $w_i(t)$ is the measurement noise. The exploration rates are chosen as $\epsilon_i(t) = \frac{1}{10}t^{-\frac{1}{100}}$ and the exploration deviations are chosen as $e_i(t) = 9/(10t^2)$. The measurement noises are chosen as uniformly distributed over two different intervals $[-10, 10]$ and $[-20, 20]$, respectively. Compared with Fig. 1, Figs. 3–4 show that the action profiles oscillate and the convergence slows down as the noise magnitude increases. In addition, we also evaluate the

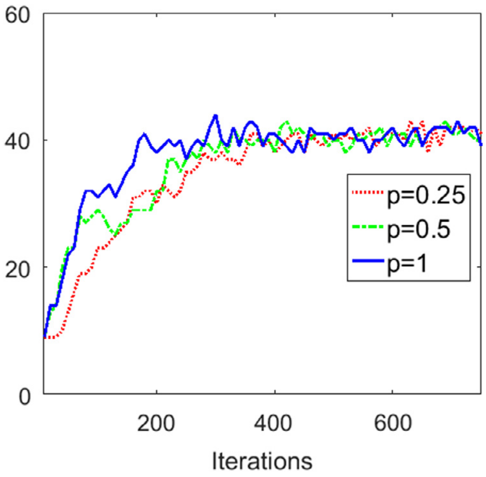


Fig. 2. Comparison of temporal aggregate demands allocated at time slot 1 with diminishing different exploration rates $\epsilon_i(t) = \frac{1}{10} t^{-\frac{p}{100}}$.

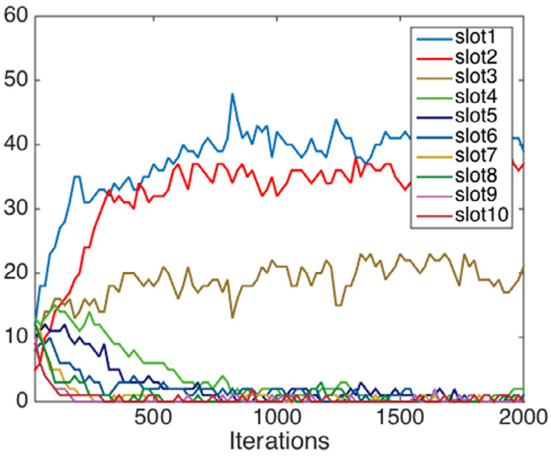


Fig. 3. Temporal aggregate demands allocated at ten time slots with uniformly distributed measurement noises in the interval $[-10, 10]$.

performance of the RL algorithm where the measurement noises are time-dependent. In particular, $w_i(t)$ is uniformly distributed over the interval $[-10\ln(t), 10\ln(t)]$. The result shown in Fig. 5 implies that the action profiles do not converge anymore.

Matlab simulation results with fixed exploration rates. Fig. 6 shows the evaluation of the RL algorithm with fixed exploration rates $\epsilon_i(t) = \frac{1}{10}^{-\frac{1}{100}}$. The exploration deviations are chosen as $e_i(t) = 9/(10t^2)$. The comparison of Figs. 1 and 6 shows that fixed exploration rates cause larger oscillations in steady state.

7. Conclusion

This paper investigates a class of multi-player discrete games where each player aims to maximize its own utility function with limited information about the game of interest. We propose the RL algorithm which converges to the set of action profiles which have maximal stochastic potential with probability one. The convergence rate of the proposed algorithm is analytically quantified. Moreover, the performance of the algorithm is verified by a case study in the smart grid. A future work is to study the scenario that the measurements of utility values are subject to non-stationary noises. In addition, the derived upper bound (1) could be conservative. Especially, the constant C could be large when N and $|\mathcal{S}|$ are large. Another future work is to find a tighter upper bound of the RL algorithm or improve the convergence rate by modifying the algorithm.

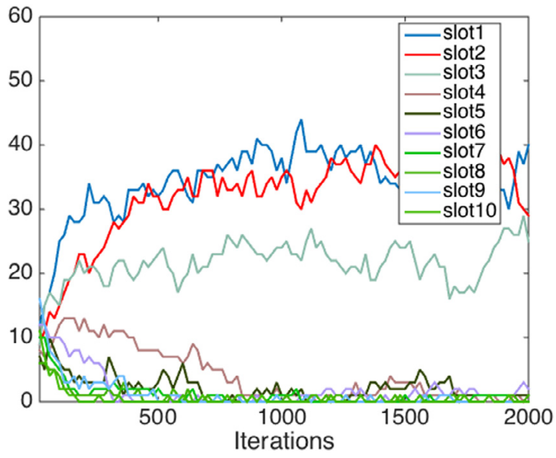


Fig. 4. Temporal aggregate demands allocated at ten time slots with uniformly distributed measurement noises in the interval $[-20, 20]$.

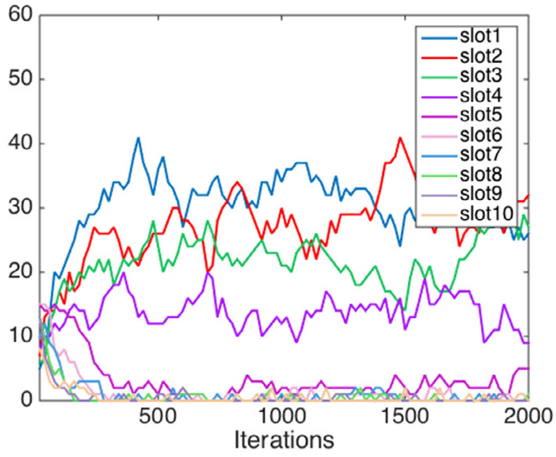


Fig. 5. Temporal aggregate demands allocated at ten time slots with uniformly distributed measurement noises in the interval $[-10\ln(t), 10\ln(t)]$.

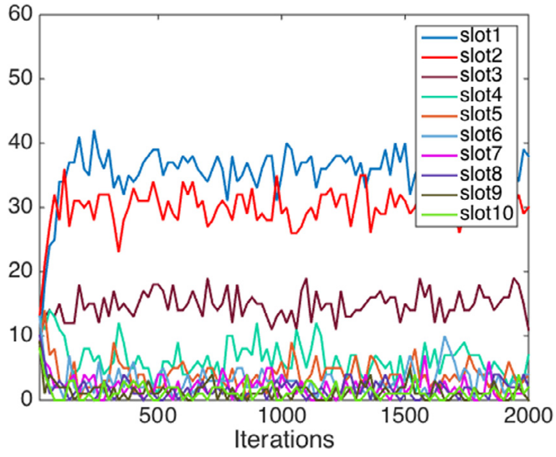


Fig. 6. Temporal aggregate demands allocated at ten time slots with fixed exploration rates $\epsilon_i(t) = \frac{1}{10}^{-\frac{1}{100}}$.

References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions: With formulas, graphs, and mathematical tables, vol. 55*. Courier Corporation.

Altman, E., Basar, T., & Srikant, R. (2002). Nash equilibria for combined flow control and routing in networks: Asymptotic behavior for a large number of users. *IEEE Transactions on Automatic Control*, *47*(6), 917–930.

Arrow, K., & Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*, *22*, 265–290.

Arslan, G., Marden, J. R., & Shamma, J. S. (2007). Autonomous vehicle-target assignment: A game-theoretical formulation. *ASME Journal on Dynamic Systems, Measurement, and Control*, *129*(5), 584–596.

Basar, T., & Olsder, G. (1999). *Dynamic noncooperative game theory*. SIAM Classics in Applied Mathematics.

Bertsekas, D. P. (2015). Convex optimization algorithms. Athena Scientific.

Biggs, N. L. (1993). *Algebraic graph theory*. Cambridge University Press.

Bjelica, M. (1995). Hadamard's inequality and fixed-point method. *Filomat*, *9*(3), 599–602.

Chlebus, E. (2009). An approximate formula for a partial sum of the divergent p-series. *Applied Mathematics Letters*, *22*(5), 732–737.

Fabrikant, A., Jaggard, A. D., & Schapira, M. (2010). On the structure of weakly acyclic games. In *International Symposium on Algorithmic Game Theory* (pp. 126–137). Athens, Greece.

Facchinei, F., & Kanzow, C. (2007). Generalized Nash equilibrium problems. *4OR*, *5*(3), 173–210.

Foster, D., & Young, P. (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology*, *38*(2), 219–232.

Freidlin, M. I., Szücs, J., & Wentzell, A. D. (2012). *Random perturbations of dynamical systems, vol. 260*. Springer Science & Business Media.

Frihauf, P., Krstic, M., & Basar, T. (2012). Nash equilibrium seeking in non-cooperative games. *IEEE Transactions on Automatic Control*, *57*(5), 1192–1207.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games, vol. 2*. MIT Press.

Hasminskii, R. Z., & Silver, B. (1972). *Stochastic approximation and recursive estimation, vol. 47*. American Mathematical Soc..

Hatanaka, T., Wasa, Y., Funada, R., Charalambides, A. G., & Fujita, M. (2016). A payoff-based learning approach to cooperative environmental monitoring for PTZ visual sensor networks. *IEEE Transactions on Automatic Control*, *61*(3), 709–724.

Koshal, J., Nedić, A., & Shanbhag, U. V. (2013). Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, *58*(3), 594–609.

Kushner, H., & Yin, G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.

Liu, S., & Krstic, M. (2011). Stochastic Nash equilibrium seeking for games with general nonlinear payoffs. *SIAM Journal on Control and Optimization*, *49*(4), 1659–1679.

Marden, J. R., Ruben, S. D., & Pao, L. Y. (2013). A model-free approach to wind farm control using game theoretic methods. *IEEE Transactions on Control Systems Technology*, *21*(4), 1207–1214.

Marden, J. R., Young, H. P., Arslan, G., & Shamma, J. S. (2009). Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization*, *48*(1), 373–396.

Milchtaich, I. (1996). Congestion games with player-specific payoff functions. *Games and Economic Behavior*, *13*(1), 111–124.

Mitra, D., Romeo, F., & Sangiovanni-Vincentelli, A. (1986). Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, *18*(3), 747–771.

Monderer, D., & Shapley, L. (1996). Potential games. *Games and Economic Behavior*, *14*(1), 124–143.

Palomar, D., & Eldar, Y. (2010). *Convex optimization in signal processing and communications*. Cambridge University Press.

Pang, J. S., Scutari, G., Facchinei, F., & Wang, C. (2008). Distributed power allocation with rate constraints in Gaussian parallel interference channels. *IEEE Transactions on Information Theory*, *54*(8), 3471–3489.

Rosen, J. (1965). Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, *33*(3), 520–534.

Roumboutsos, A., & Kapros, S. (2008). A game theory approach to urban public transport integration policy. *Transport Policy*, *15*(4), 209–215.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

Stankovic, M., Johansson, K., & Stipanovic, D. (2012). Distributed seeking of Nash equilibria with applications to mobile sensor networks. *IEEE Transactions on Automatic Control*, *57*(4), 904–919.

Tutte, W., & Nash-Williams, C. (2001). *Graph theory*. Cambridge University Press.

Uspensky, J. V. (1937). *Introduction to mathematical probability*. McGraw-Hill.

Wang, G., Shanbhag, U. V., & Meyn, S. P. (2012). On nash equilibria in duopolistic power markets subject to make-whole uplift. In *2012 IEEE 51st IEEE conference on decision and control* (pp. 472–477). Maui, Hawaii, USA.

Yin, H., Shanbhag, U., & Mehta, P. (2011). Nash equilibrium problems with scaled congestion costs and shared constraints. *IEEE Transactions on Automatic Control*, *56*(7), 1702–1708.

Young, H. P. (1993). The evolution of conventions. *Econometrica: Journal of the Econometric Society*, *61*, 57–84.

Young, H. P. (2001). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Yousefian, F., Nedić, A., & Shanbhag, U. V. (2013). A distributed adaptive steplength stochastic approximation method for monotone stochastic Nash games. In *2013 American Control Conference* (pp. 4765–4770).

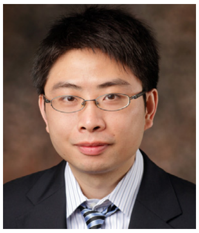
Zhu, M. (2014). Distributed demand response algorithms against semi-honest adversaries. In *IEEE Power and Energy Society General Meeting. 943*. National Harbor, MD.

Zhu, M., & Frazzoli, E. (2016). Distributed robust adaptive equilibrium computation for generalized convex games. *Automatica*, *63*(1), 82–91.

Zhu, M., Hu, Z., & Liu, P. (2014). Reinforcement learning algorithms for adaptive cyber defense against heartbleed. In *First ACM Workshop on Moving Target Defense in Association with 2014 ACM Conference on Computer and Communications Security* (pp. 51–58). Scottsdale, Arizona, USA.

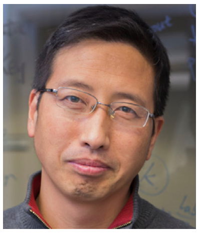
Zhu, M., & Martínez, S. (2013). Distributed coverage games for energy-aware mobile sensor networks. *SIAM Journal on Control and Optimization*, *51*(1), 1–27.

Zhisheng Hu is a Ph.D. student in the School of Electrical Engineering and Computer Science at the Pennsylvania State University. Prior to that, he worked as a research assistant in the School of Information Science and Technology at Sun Yat-sen University. He received the B.E. degree in electrical engineering from Sun Yat-sen University, Guangdong, China, in 2013. His research interests mainly focus on machine learning in network and system security.



Minghui Zhu is an assistant professor in the School of Electrical Engineering and Computer Science at the Pennsylvania State University. Prior to that, he was a postdoctoral associate in the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology. He received Ph.D. in Engineering Science (Mechanical Engineering) from the University of California, San Diego in 2011. His research interests lie in the control and decision-making of multi-agent networks with applications in robotic networks, security and the smart grid. He is the co-author of the book "Distributed optimization-based control of multi-agent networks in complex environments" (Springer, 2015). He is an associate editor of the Conference Editorial Board of the IEEE Control Systems Society. He received the award of Outstanding Graduate Student of Mechanical and Aerospace Engineering at the University of California, San Diego in 2011, and the Dorothy Quiggle Career Development Professorship in Engineering at the Pennsylvania State University in 2013. He was selected as an outstanding reviewer of Automatica in 2013 and 2014.

Ping Chen received his Ph.D. from Nanjing University in 2012. Ping Chen received Microsoft Fellowship in 2011 and worked as Post-doc scholar at Penn State University in 2014. His research interests are in software and system security. He has published 30+ papers in top conferences and journals, including CCS, USENIX Security, and DSN. He served as Program committee of Third ACM Workshop on Moving Target Defense (MTD 2016), reviewers for multiple conferences and journals, including TDSC, IET, ESORICS, and CCS.



Peng Liu received his BS and MS degrees from the University of Science and Technology of China, and his Ph.D. from George Mason University in 1999. Dr. Liu is the Raymond G. Tronzo, M.D. Professor of Cybersecurity, founding Director of the Center for Cyber-Security, Information Privacy, and Trust, and founding Director of the Cyber Security Lab at Penn State University. His research interests are in all areas of computer security. He has published numerous papers in top conferences and journals. His research has been sponsored by NSF, ARO, AFOSR, DARPA, DHS, DOE, AFRL, NSA, TTC, CISCO, and HP. He has served as a program (co-)chair or general (co-)chair for over 10

international conferences (e.g., Asia CCS 2010) and workshops (e.g., MTD 2016). He chaired the Steering Committee of SECURECOMM during 2008–14. He has served in over 100 program committees and reviewed papers for numerous journals. He is an associate editor for IEEE TDSC. He is a recipient of the DOE Early Career Principle Investigator Award. He has co-led the effort to make Penn State a NSA-certified National Center of Excellence in Information Assurance Education and Research. He has advised or co-advised over 35 Ph.D. dissertations to completion.