Title

Distributed Q-Learning with State Tracking for Multi-agent Networked Control

Authors

Wang, Hang
Lin, Sen
Jafarkhani, Hamid
et al.

Publication Date

2020-12-22

Peer reviewed

DISTRIBUTED Q-LEARNING WITH STATE TRACKING FOR MULTI-AGENT NETWORKED CONTROL

Hang Wang
Arizona State University
Tempe, AZ 85287-7206
hwang442@asu.edu

Sen Lin
Arizona State University
Tempe, AZ 85287-7206
slin70@asu.edu

Hamid Jafarkhani
University of California, Irvine
Irvine, CA 92697-2625
hamidj@uci.edu

Junshan Zhang
Arizona State University
Tempe, AZ 85287-7206
Junshan.Zhang@asu.edu

December 24, 2020

ABSTRACT

This paper studies distributed Q-learning for Linear Quadratic Regulator (LQR) in a multi-agent network. The existing results often assume that agents can observe the global system state, which may be infeasible in large-scale systems due to privacy concerns or communication constraints. In this work, we consider a setting with unknown system models and no centralized coordinator. We devise a state tracking (ST) based Q-learning algorithm to design optimal controllers for agents. Specifically, we assume that agents maintain local estimates of the global state based on their local information and communications with neighbors. At each step, every agent updates its local global state estimation, based on which it solves an approximate Q-factor locally through policy iteration. Assuming decaying injected excitation noise during the policy evaluation, we prove that the local estimation converges to the true global state, and establish the convergence of the proposed distributed ST-based Q-learning algorithm. The experimental studies corroborate our theoretical results by showing that our proposed method achieves comparable performance with the centralized case.

1 Introduction

Distributed control of multi-agent systems (MASs) has garnered much attention in the past decade, due to its wide applicability in real-world problems, e.g., firefighting unmanned aerial vehicles maneuver, distributed resource allocation and robot swarms, etc. One main objective in this context is to learn local controllers for agents in a distributed manner so as to minimize the global cost [1]. For example, in the Linear Quadratic Regulator (LQR) control problem, the global objective is to minimize the sum of the local quadratic costs over all agents.

Nevertheless, the networked nature of MASs presents some unique challenges in designing distributed controllers. Observe that the agents are physically coupled with certain interconnections [2], e.g., the buses in a microgrid are interconnected through structural links such as the power transmission lines. Consequently, the controller synthesis at a bus has to account for the impact of other buses. To deal with the sophisticated coupling in MASs, the model-based distributed controller design has been studied in [3, 4, 5], where the interconnections among agents are modeled by a directed interaction graph to model the system dynamics. However, these studies assume that the underlying system model is known, which may be infeasible in large-scale systems.

To tackle the challenges in the MASs with unknown system models, data-driven approaches have emerged as a promising direction in learning local controllers. Notably, data-driven Q-learning [6], which is a model-free Reinforcement Learning (RL) approach [7], has been proposed to learn the optimal LQR controller online in the single agent case [8]. Motivated by this, some recent works (e.g., [9, 10]) apply the Q-learning in the multi-agent LQR control and show that

good performance can be achieved assuming that the knowledge of global state information is shared by a centralized coordinator [9, 10]. Nevertheless, such a centralized coordinator is often not available in many scenarios. It is therefore of great interest to study the MASs where each agent can only learn state information from neighbors with limited communication. Needless to say, this lacking of global state information inevitably makes the learning of the optimal controller more challenging, calling for distributed control based on *partial observations*.

In this work, we address the above problem by revisiting the distributed LQR control problem with only partial observations of the global state. Specifically, we consider a distributed MAS where each agent has a discrete Linear Time Invariant (LTI) system with unknown dynamics. We assume that each agent can only share information with its neighbors over a communication graph. In particular, we focus on a more practical setting where the physical interconnection is different from the communication interconnection. This is often the case in the emerging cyber-physical systems, e.g., microgrid systems with distributed generators (DGs), the DGs are physically interconnected by a microgrid electric power network, and communicate in the cyberlayer [11].

Without careful design, the performance of distributed Q-learning can significantly degrade with only partial observations. To tackle this challenge, we propose a distributed Q-learning approach with a novel *state tracking* strategy to facilitate the estimation of the global state through limited communication among neighboring agents. Intuitively, by exchanging state estimations with neighboring agents, an individual agent would be able to improve its global state estimator as the information continuously diffuses across the network. Based on such a global state estimation, each agent then solves an *approximate* Q-factor locally; and this learning process is carried out in parallel by all agents.

The main contributions of this work can be summarized as follows:

- Considering distributed LQR control in MASs with only partial observations, we propose a novel distributed Q-learning approach with state tracking (ST-Q), where each agent first constructs a global state estimator based on local communication with its neighbors, and then solves an approximate Q-learning problem accordingly.
- The convergence of distributed Q-learning algorithms in multi-agent LQR control has been underexplored. In this work, we fill this void and establish the convergence of the proposed distributed ST-Q algorithm.
- Compared with Q-learning under full observation [9] and Q-learning under partial observation only, our experimental results show that the proposed distributed ST-Q learning method achieves comparable performance with the full observation case.

2 Related Work

Distributed LQR Control. Distributed LQR control has recently garnered much attention. Notably, identical LTI system models across agents have been considered which are coupled either in a global cost function [12] or in the state space [13]. Since it is restrictive to assume identical systems for all agents, [5] explores distributed model predictive control for heterogeneous LTI systems. Regarding the communication structure among agents, [14] requires all-to-all communications for the optimal control, and [15] assumes that agents share information with all their physically coupled neighbors. In this work, we consider a more challenging setting where (i) the system model parameters are unknown, and (ii) the communication topology is different from the system interconnection topology (cf. [16]).

Multi-Agent Reinforcement Learning (MARL). Taking a distributed Q-learning approach, our focus is on MARL with partial observations (see, e.g., Dec-POMDP [17]). Information exchange is often utilized to facilitate the collaboration among agents. Notably, [18] proposes a fully decentralized MARL where each agent shares local value function estimates with its neighbors to achieve a network-wide consensus. In [19], an approach is devised to enable each agent to communicate its local estimation of global optimal policy parameters with its neighbors. Note that both methods assume full observation of the global state and control information to compute the gradient estimations. Assuming each agent solely has access to partial observations, a recent work [20] develops a policy gradient method where each agent shares an estimate of the global cost based on the local information only; further all agents act in parallel and have no control interaction with others. It is worth noting that the policy gradient method does not necessarily achieve the global optimum control policy [21].

Along a different line, an LQR control method using Q-learning is proposed [8], which provides the first convergence guarantee on Q-learning based optimal control for the single agent case. A recent work [22] presents a distributed Q-learning algorithm for coupled LTI systems with identical dynamics. Assuming that the global state information is available through a central coordinator, [9] establishes the convergence of distributed optimal controllers for coupled LTI systems. [15] studies the decentralized Q-learning, where the Q-function is calculated based on the local observations only and the system interconnection topology is assumed known. They use simulation studies to show that a near-optimal policy can be obtained. To fill the void, we propose a state tracking strategy to estimate the global state

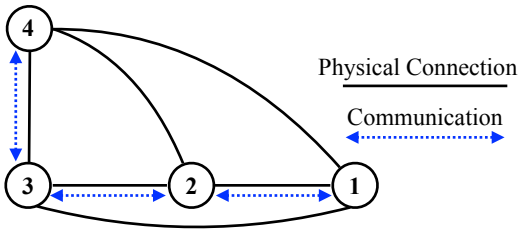


Figure 1: An example of the communication graph (dashed lines) and the interconnection graph (solid lines). All the agents are interconnected with physical connection, while the direct communication channel does not exist between neither Agent 4 and Agent 2 nor Agent 4 and Agent 1.

information at each agent based on its local information aggregated from neighbors. In such a way, the Q-factor can be solved more accurately at each agent by using the global state estimation.

3 Problem Formulation

3.1 Notation

Throughout the paper, the set $\{1, 2, \cdots, L\} \subseteq \mathbb{N}$ is denoted as $[L]$. The block-diagonal matrix B with blocks $\{B_i\}_{i\in[L]}$ is denoted as $B = \text{diag}(B_1, B_2, \cdots, B_L)$. A column vector which stacks subvectors $\{x_i\}_{i\in[L]}$ in a column is denoted as $X = \text{col}(x_1, x_2, \cdots, x_L)$. We use semicolon (;) to concatenate column vectors, hence $[x^\top, u^\top]^\top = [x; u]$. A graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges connecting agents. For a node $i \in \mathcal{V}$, we denote $\mathcal{N}_i = \{j \in \mathcal{V} | j \neq i, e_{ij} = (i, j) \in \mathcal{E}\} \cup \{i\}$ as the set of neighbors of node i in the graph $\mathcal{G}$. We use $0_{n\times n}$ to represent a $n \times n$ zero matrix.

3.2 Multi-Agent LTI System Model

Consider a multi-agent network consisting of L agents, where the LTI system dynamics at each agent $i \in [L]$ is given as follows:

$$x_i(t+1) = \sum_{j=1}^{L} A_{ij}x_j(t) + B_i u_i(t) \tag{1}$$

where $x_i(t) \in \mathbb{R}^n$ is Agent i's state vector and $u_i(t) \in \mathbb{R}^m$ is its control input at time t. $A_{ij} \in \mathbb{R}^{n\times n}$ and $B_i \in \mathbb{R}^{n\times m}$ are unknown system parameters. Putting the system models across all agents in a more compact form, we have the following global system model:

$$X(t+1) = AX(t) + BU(t) \tag{2}$$

where $X(t) = \text{col}(x_1(t), x_2(t), \cdots, x_L(t))$ is the global state vector, and $U(t) = \text{col}(u_1(t), u_2(t), \cdots, u_L(t))$ is the global control input vector. The global system matrix $A \in \mathbb{R}^{nL\times nL}$ is block-wise with entries A_{ij} for each $i, j \in [L]$ and $B = \text{diag}(B_1, B_2, \cdots, B_L)$.

We further define a graph $\mathcal{G}^d = ([L], \mathcal{E}^d)$ to model the interconnection topology, i.e., the state coupling, of the global system, where $[L]$ is the node (agent) set. Specifically, there exists an edge $e^d_{ij} \in \mathcal{E}^d$ between Agent i and Agent j if and only if they are interconnected, i.e., $A_{ij} \neq 0$. Let $\mathcal{N}^d_i$ denote as the set of neighbors of Agent i in the interconnection graph $\mathcal{G}^d$.

As is standard, we impose the following assumption on (A, B) in this study.

Assumption 1 (Stabilizability)**.** *The system parameters (A, B) in (2) are stabilizable.*

Assumption 1 indicates that there exists a control policy π with $U(t) = \pi(X(t))$, such that the closed loop system $X(t+1) = AX(t) + B\pi(X(t))$ is asymptotically stable.

3.3 Distributed LQR Control with Local Communication

Optimal distributed LQR control. For ease of expostion, we first present the optimal distributed LQR controller at each agent i assuming model parameters are known. For the subsystem (1) at each agent i, the stage cost incurred by

executing the control $u_i(t)$ in state $x_i(t)$ at time t is given by

$$g_i(x_i(t), u_i(t)) = x_i(t)^\top P_i x_i(t) + u_i(t)^\top R_i u_i(t)$$

where $P_i \in \mathbb{R}^{n\times n}$ and $R_i \in \mathbb{R}^{m\times m}$ are positive semi-definite matrices. Let $J_i(x_i(0)) = \sum_{\tau=0}^{\infty} g_i(x_i(\tau), u_i(\tau))$ denote the local cost function at Agent i. The primary goal of the distributed LQR control is to minimize the sum of the local costs of all agents:

$$\min_{\{u_i(\tau)\}} \sum_{i=1}^{L} J_i(x_i(0)), \text{ s.t. (1).} \tag{3}$$

Let $X(0) = \text{col}(x_1(0), x_2(0), \cdots, x_L(0))$, $P = \text{diag}(P_1, P_2, \cdots, P_L)$ and $R = \text{diag}(R_1, R_2, \cdots, R_L)$. It is clear that the distributed LQR control problem (3) is equivalent to the following LQR problem for the global system (2) with the initial state $X(0)$:

$$\min_{\{U(\tau)\}} J(X(0)), \text{ s.t. (2),} \tag{4}$$

with

$$J(X(0)) = \sum\nolimits_{\tau=0}^{\infty} X^\top(\tau) P X(\tau) + U^\top(\tau) Q U(\tau).$$

When the model parameters A and B are known, the optimal policy for (4) is given by linear feedback control, i.e.,

$$U(t) = K^* X(t),$$

where $K^* = -(Q + B^T S B)^{-1} B^T S A$ and S is the positive definite solution to the discrete Riccati equation:

$$S = A^T S A - A^T S B (Q + B^T S B)^{-1} B^T S A + P.$$

The optimal control policy for each agent thus can be obtained as

$$u_i(t) = K_i^* X(t), \tag{5}$$

where the optimal controller K_i^* is the i-th row of K^*. The LQR solution in this case can be efficiently computed via dynamic programming. In the case when the model parameters A and B are unknown but each agent has a full observation of the global state $X(t)$, problem (3) can be solved efficiently by using reinforcement learning approaches [23, 9].

Distributed LQR control with local communication. The primary focus of this paper is on distributed LQR control for a multi-agent network with unknown system parameters. Specifically, we define an undirected communication graph $\mathcal{G}^c = ([L], \mathcal{E}^c)$ to model the information exchange in the multi-agent network. There exists an edge $e_{ij}^c \in \mathcal{E}^c$ between Agent i and Agent j if and only if they can communicate. Let $\mathcal{N}_i^c$ denote the set of neighbors of Agent i in the communication graph $\mathcal{G}^c$. In particular, we consider a general setting where the interconnection graph $\mathcal{G}^d$ and the communication graph $\mathcal{G}^c$ can be distinct, as illustrated in Figure 1. Since each agent does not have the full observation of the global state, it can only make its control decisions based on the local information, giving rise to the distributed LQR control that only relies on the partial observation (POD-LQR).

Let $x_{\mathcal{N}_i}(t) \in \mathcal{X}_{\mathcal{N}_i}$ denote the state information available for Agent i at time t, which contains partial entries of the global state vectors. Agent i then selects the local control input $u_i(t) \in \mathcal{U}_i$, based on the information $x_{\mathcal{N}_i}(t)$ and a control policy $\tilde{\pi}_i$ with a linear feedback controller, i.e.,

$$\tilde{\pi}_i : \mathcal{X}_{\mathcal{N}_i} \mapsto \mathcal{U}_i. \tag{6}$$

We further assume that K_i is the feedback controller in the policy $\tilde{\pi}_i$. The goal of POD-LQR control is to find controllers that minimize the infinite horizon global cost function $J(X(0))$:

$$\min_{\{K_i\}} J(X(0)) = \sum\nolimits_{i=1}^{L} J_i(x_i(0)), \text{ s.t. (1), (6).} \tag{7}$$

In this work, we aim to achieve the optimal controller K_i^* for each agent i that is the same as in the case where the model parameters are known, by solving Problem (7) based on only a partial observation of the global state.

4 Distributed Q-learning with State Tracking

In this section, we propose a distributed Q-learning approach with state tracking to solve Problem (7), where each agent first constructs a global state estimator through communication with its neighbors, and then solves an approximate Q-learning problem locally using the state estimation. We start by presenting the preliminary on Q-learning with a full observation of the global state. Then, we present a state tracking scheme to facilitate the estimation of the global state through limited information exchange among neighboring agents. Finally, the proposed state tracking based policy iteration algorithm is presented in detail.

4.1 Global State based Q-Learning

As mentioned earlier, Q-learning can be utilized to solve the distributed LQR control problem (3) if each agent has a full observation of the global state $X(t)$. In what follows, we will briefly introduce the rationale behind Q-learning in the ideal case when the global state information is available at each agent.

Specifically, given the global state $X(t)$ and based on (5), we consider the local control policy $\pi_i:\ u_i(t) = K_i X(t)$ for some state feedback controller K_i. Then, the Q-factor for each agent i can be defined as follows:

$$Q_i(x_i(t), u_i(t)) = g_i(x_i(t), u_i(t)) + J_i(x_i(t+1)), \tag{8}$$

which gives the cumulative cost when agent i starts from the state-control pair $(x_i(t), u_i(t))$ and follows the policy π_i afterwards. Note that

$$J_i(x_i(t)) = Q_i(x_i(t), K_i X(t)).$$

The Bellman equation associated with the policy π_i for the Q-factor can be written as

$$\begin{aligned} &Q_i(x_i(t), K_i X(t)) \\ =&g_i(x_i(t), K_i X(t)) + Q_i(x_i(t+1), K_i X(t+1)), \end{aligned} \tag{9}$$

and the corresponding Bellman optimality equation is

$$\begin{aligned} &Q_i^*(x_i(t), K_i^* X(t)) \\ =&g_i(x_i(t), K_i^* X(t)) + Q_i^*(x_i(t+1), K_i^* X(t+1)). \end{aligned} \tag{10}$$

This implies that the optimal controller K_i^* can be achieved as:

$$K_i^* = \arg\min_{K_i} Q_i^*(x_i(t), K_i X(t)).$$

Therefore, to find the optimal controller K_i^*, it suffices to estimate the optimal Q-factor Q_i^*. And this can be achieved by using policy iteration where a sequence of monotonically improved policies and Q-factors can be obtained, by running a policy evaluation step and then a policy improvement step in a recursive manner. For a better understanding of the Q-learning approach for LQR control, we start with the policy improvement step.

Policy improvement. A key step in policy iteration is the policy improvement. Suppose we have determined the Q-factor Q_i for a controller K_i in the policy evaluation step. The policy improvement step aims to find a better controller:

$$K_i^{\text{new}} = \arg\min_{K_i}(Q_i(x_i(t), K_i X(t))). \tag{11}$$

Note that the cost function J_i is quadratic in the LQR control problem with a linear state feedback controller [7]. Then, it can be shown that

$$Q_i(x_i(t+1), K_i X(t+1)) = x_i(t+1)^\top S_i x_i(t+1).$$

Here, S_i is the cost matrix for the current controller K_i, which can be obtained by solving the discrete-time algebraic Riccati equation [24]. Thus, (8) can be rewritten as the following quadratic form:

$$Q_i(x_i(t), u_i(t)) = [X(t); u_i(t)]^\top H_i [X(t); u_i(t)], \tag{12}$$

where H_i is a symmetric block matrix defined as

$$H_i = \begin{bmatrix} H_{i,11} & H_{i,12} \\ H_{i,21} & H_{i,22} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_i{}^\top S_i \mathcal{A}_i + \tilde{P}_i & \mathcal{A}_i{}^\top S_i B_i \\ B_i{}^\top S_i \mathcal{A}_i & B_i{}^\top S_i B_i + R_i \end{bmatrix}.$$

Here, $\mathcal{A}_i$ is a row vector which stacks subvectors $\{A_{ij}\}_{j\in[L]}$ and $\tilde{P}_i = \text{diag}(0_{n\times n}, \cdots, P_i, \cdots, 0_{n\times n})$ is a diagonal block matrix with the (i,i)-th block to be P_i. Based on the samples of the state-control pair $[X(t); u_i(t)]$, (11) can be solved by using the first-order optimality condition:

$$K_i^{\text{new}} = -H_{i,22}^{-1} H_{i,21}. \tag{13}$$

Summarizing, to obtain an improved controller K_i^{new}, it suffices to determine the Q-factor, in particular, the matrix H_i, in the policy evaluation step.

Policy evaluation. To determine the matrix H_i in the policy evaluation step, along the same line as in [8], we reformulate the quadratic form of $Q_i(x_i(t), u_i(t))$ in (12) in a linear form parameterized by parameter θ_i:

$$Q_i(x_i(t), u_i(t)) = y_i(t)^\top \theta_i, \tag{14}$$

where $y_i(t) = [x_1^2(t), x_1(t)x_2(t), \cdots, x_L(t)u_i(t), u_i^2(t)]$ is a vector containing all of the quadratic basis over the elements in $[X(t); u_i(t)]$, and θ_i is a vector in $\mathbb{R}^{(Ln+m)(Ln+m+1)/2}$. Here, the parameter θ_i is obtained through some manipulation after removing the redundant elements of the symmetric matrix H_i, i.e., the elements in the lower triangle of H_i. It is clear that in order to determine H_i, it suffices to determine the parameter θ_i.

Based on the linear form (14), it is clear that the Bellman equation (9) is equivalent to the following:

$$g_i(x_i(t), u_i(t)) = (y_i(t) - y_i(t+1))^\top \theta_i \triangleq \phi_i(t)^\top \theta_i, \tag{15}$$

where $\phi_i(t) = y_i(t) - y_i(t+1)$. Note that $\phi_i(t)$ and the stage cost $g_i(x_i(t), u_i(t))$ can be known given the global state $X(t)$ and the control input $u_i(t)$. With sufficient samples of $(\phi_i(t), g_i(x_i(t), u_i(t)))$, θ_i can be obtained by solving a least square estimation problem.

4.2 State Tracking

It can be seen from (15) that the global state $X(t)$ is required to determine the parameter θ_i in the policy evaluation step, which however is not available in the POD-LQR control problem. To address this issue, we propose a state tracking scheme to facilitate the estimation of the global state $X(t)$ through the information exchange among agents over the communication graph $\mathcal{G}^c$.

More specifically, at time t each agent i maintains a local estimation $Z_i(t)$ of the global state $X(t)$:

$$Z_i(t) = \text{col}(\bar{x}_{i1}(t), \bar{x}_{i2}(t), \cdots, \bar{x}_{iL}(t)),$$

where $\bar{x}_{ij}(t)$ is the estimation of Agent j's state $x_j(t)$ at Agent i for time t. In particular, $\bar{x}_{ii}(t) = x_i(t)$. Next, each agent communicates with and aggregates information from its neighbors in the communication graph $\mathcal{G}^c$ to update the local estimation $Z_i(t+1)$ as follows.

Communication among neighboring agents. At time $t+1$, the communication among agents includes two steps. First, each agent i receives the state $x_j(t+1)$ from every neighbor $j \in \mathcal{N}_i^c$, and then updates the corresponding entries in its estimation $Z_i(t)$, i.e.,

$$\bar{x}_{ij}(t) \to x_j(t+1),\ \forall j \in \mathcal{N}_i^c.$$

Consequently, an updated estimation $\hat{Z}_i(t+1) = \text{col}(\hat{x}_{i1}(t+1), \hat{x}_{i2}(t+1), \cdots, \hat{x}_{iL}(t+1))$ can be obtained with

$$\hat{x}_{ij}(t+1) = \begin{cases} \bar{x}_{ij}(t) & \forall j \notin \mathcal{N}_i^c, \\ x_j(t+1) & \forall j \in \mathcal{N}_i^c. \end{cases}$$

Next, each agent i shares its updated global state estimation $\hat{Z}_i(t+1)$ with its neighbors in $\mathcal{G}^c$.

Update of global state estimation. After receiving the global state estimation $\hat{Z}_i(t+1)$ from the neighboring agents, Agent i reconstructs a new estimation $Z_i(t+1)$ by aggregating all available information. In particular, for $j \in \mathcal{N}_i^c$, Agent i has the accurate state information $x_j(t+1)$ of Agent j; for $j \notin \mathcal{N}_i^c$, Agent i computes the state estimation $\bar{x}_{ij}(t+1)$ by taking a weighted average of the corresponding estimations $\hat{x}_{kj}(t+1)$ from its neighbors $k \in \mathcal{N}_i^c$. To model this 'weighting' process, a doubly stochastic weight matrix, $W = [w_{ij}] \in \mathbb{R}^{L \times L}$, is used where $w_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}^c$. Otherwise, $w_{ij} = 0$. The specific update rule is shown as following

$$\bar{x}_{ij}(t+1) = \begin{cases} \sum_{k=1}^{L} w_{ik} \hat{x}_{kj}(t+1) & \forall j \notin \mathcal{N}_i^c, \\ x_j(t+1) & \forall j \in \mathcal{N}_i^c. \end{cases} \tag{16}$$

4.3 ST-based Policy Iteration for Q-learning

Based on the estimation $Z_i(t)$ of the global state $X(t)$ achieved by state tracking, each agent i now is able to carry out an approximate Q-learning locally by using policy iteration to solve the POD-LQR control problem (7). As mentioned earlier in Section 4.1, the policy iteration includes two main steps, i.e., the policy evaluation and the policy improvement step. We summarize the important steps below, and more details can be found in Algorithms (1a) and (1b).

Specifically, each agent $i \in [L]$ first starts with a stabilizing initial controller K_{i1}, and iteratively runs the policy evaluation and the policy improvement as shown in Fig. 2. At iteration q,

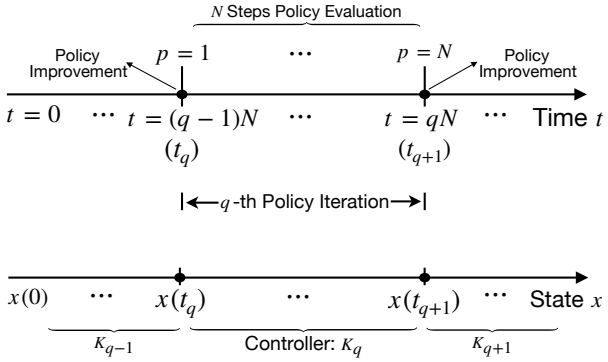


Figure 2: Illustration of the time scales in Algorithms (1a) and (1b): t denotes the time steps, and t_q denotes the time instance for the q-th policy iteration. Policy evaluation is carried out N times within each policy iteration.

Algorithm 1a ST based Policy Evaluation (ST-E)

Require: K_{iq}: evaluation controller, $\eta_i(t)$: excitation noise.
1: $p = 1$.
2: **for** $p = 1, \cdots, N$ **do**
3: Apply $u_i(t) = -K_{iq} Z_i(t) + \eta_i(t)$.
4: Measure $x_i(t+1)$ and receive $x_j(t+1)$, $j \in \mathcal{N}_i^c$.
5: Receive $\hat{Z}_j(t+1)$ from all $j \in \mathcal{N}_i^c$ and update $Z_i(t+1)$ following (16).
6: Obtain $u_i(t+1) = -K_{iq} Z_i(t+1)$.
7: Update $\hat{\theta}_{iq}(p)$ using (18).
8: Set $p = p + 1$ and $t = t + 1$.
9: **end for**
10: **Return** $\hat{\theta}_{iq}$.

Algorithm 1b ST based Q-learning (ST-Q)

Require: K_{i1}: initial stable controller, $\theta_{i1}(0) = 0$: initial estimation, $q = 1$: policy improvement index, $t = 0$: time index, ε_K: tolerance error, $x_i(0)$: initial state, $Z_i(0)$: initial global state estimation.
1: **repeat**
2: **for** Agent $i = 1, \cdots, L$ **do**
3: Estimate θ_{iq} through Algorithm (1a).
4: **end for**
5: **for** Agent $i = 1, \cdots, L$ **do**
6: Obtain H_{iq} from $\hat{\theta}_{iq}(N)$.
7: Update policy $K_{i(q+1)} = -H_{iq,22}^{-1} H_{iq,21}$.
8: Set $\hat{\theta}_{i(q+1)}(0) = \hat{\theta}_{iq}(N)$.
9: **end for**
10: Set $q = q + 1$.
11: **until** $\|\hat{\theta}_{i(q+1)} - \hat{\theta}_{iq}\| < \varepsilon_K, \forall i \in [L]$

- *Policy Evaluation.* In this step, each agent i aims to determine the Q-factor Q_{iq} for a given controller K_{iq}. As shown in (14), it suffices to determine the parameter θ_{iq}. To this end, we resort to least square estimation based on (15) with N samples per policy evaluation step:

$$\min_{\theta_{iq}} \sum_{t=t_q}^{t_q+N-1} \|\bar{\phi}_i(t)^T \theta_{iq} - g_i(x_i(t), u_i(t))\|^2, \tag{17}$$

where $\bar{\phi}_i(t) = \bar{y}_i(t) - \bar{y}_i(t+1)$ and $\bar{y}_i(t) = [\bar{x}_{i1}^2(t), \bar{x}_{i1}(t)\bar{x}_{i2}(t), \cdots, \bar{x}_{iL}(t)u_i(t), u_i^2(t)]$ is the vector containing all the quadratic basis over the elements in the estimated global information vector $[Z_i(t); u_i(t)]$. Here, $u_i(t) = -K_{iq} Z_i(t) + \eta_i(t)$ where $\eta_i(t)$ is the input noise to ensure that the system at each agent is

persistently excited. For ease of exposition, we denote $g_i(x_i(t), u_i(t))$ as $g_i(t)$, and consider $(\bar{\phi}_i(t), g_i(t))$ as a sample for the least square estimator.

To solve the least square estimation problem (17), an online gradient descent method is run for N iterations: At each iteration $p \in [1, N]$, each agent (i) constructs the global state estimation $Z_i(t)$ and $Z_i(t+1)$ so as to obtain a sample $(\bar{\phi}_i(t), g_i(t))$ as shown in Algorithm (1a), and (ii) updates the estimation of θ_{iq} by using gradient descent with a learning rate α:

$$\hat{\theta}_{iq}(p+1) = \hat{\theta}_{iq}(p) - \alpha\bar{\phi}_i(t)\big(\hat{\theta}_{iq}(p)^\top \bar{\phi}_i(t) - g_i(t)\big), \tag{18}$$

where $\hat{\theta}_{iq}(p)$ is the estimation of θ_{iq} at policy evaluation step p.

- *Policy Improvement.* Given $\hat{\theta}_{iq} = \hat{\theta}_{iq}(N)$ obtained in the policy evaluation step, each agent is able to reconstruct the matrix $\hat{H}_{iq}$, so that the controller can be updated as follows:

$$K_{i(q+1)} = -\hat{H}^{-1}_{iq,22}\hat{H}_{iq,21}.$$

This policy iteration procedure stops if the following condition is satisfied:

$$\|\hat{\theta}_{i(q+1)} - \hat{\theta}_{iq}\| < \varepsilon_K, \forall i \in [L]$$

where ε_K is a predefined threshold for the estimation error.

5 Convergence Analysis

In this section, we establish the convergence of the proposed ST-Q learning algorithm. To this end, we first make a few standard assumptions for multi-agent reinforcement learning [25, 26].

Assumption 2 (Communication Connectivity)**.** *The communication graph $\mathcal{G}^c$ is connected and static.*

Assumption 3 (Weight Matrix)**.** *There exists a positive constant η such that the weight matrix $W = [w_{ij}] \in \mathbb{R}^{L\times L}$ is doubly stochastic and $w_{ij} \geq \eta$ if $j \in \mathcal{N}_i^c$. In particular, $w_{ii} \geq \eta$, $\forall i \in [L]$.*

Assumption 4 (Decaying Excitation Noise)**.** *The input noise $\eta(t)$ is with the decaying factor $\upsilon(p)$,*

$$\upsilon(p) = c^p, \quad 0 < c < 1,$$
$$\eta(t) = \upsilon(p)\beta(t), \quad t_q \leq t < t_{q+1}.$$

The input noise for the global system is $E(t) = \Upsilon(p)\boldsymbol{\beta}(t)$. The system is further persistently excited with the input noise, i.e., $\forall i \in [L]$, $\forall q$:

$$mI \leq \textstyle\sum_{t=t_q}^{t_q+N-1} \phi_i(t)\phi_i^\top(t) \leq MI,$$

where $0 < m \leq M < \infty$.

Assumption 5 (Step Size of Gradient Descent)**.** *The step size in* (18) *is fixed and satisfies: $0 < \alpha < 1/M$.*

Assumptions 2 and 3 are imposed to facilitate the information diffusion across the network. The excitation condition in Assumption 4 is to guarantee the convergence of policy evaluation. Along the same line as in [27, 8], this condition can be met by adding sinusoidal noise of various frequencies to $u_i(t)$. Moreover, we further assume that the input noise $\eta_i(t)$ is decaying for a more accurate system state estimation as the algorithm gradually converges.

For convenience, we restate in the following lemma the convergence result on distributed Q-learning with full global state observations [8, 9].

Lemma 1 (Convergence of Distributed Q-learning with Full Observation)**.** *Suppose that Assumptions 1, 4, 5 are satisfied, and K_{i1} is a stabilizing controller. There exists $N < \infty$ such that the sequence of stabilizing controllers $\{K_{iq}\}_{q=1}^{\infty}$ generated by the Q-learning Policy Iteration mechanism with global state information converges, i.e., $\forall i \in [L]$:*

$$\lim_{q\to\infty} \|K_{iq} - K_i^*\| = 0,$$

where K_i^* is the optimal feedback controller.

Proof Sketch. First, we show that by replacing recursive least squares (RLS) with stochastic gradient descent (SGD) in the adaptive policy iteration algorithm proposed in [8], the policy iteration algorithm also generates a sequence of stabilizing controls converging to the optimal in the single agent case. Under the setting when agents have full observation of the global state, [9] considers a policy iteration algorithm with the RLS estimation method and provides the convergence proof by utilizing the result in [8]. Following the same line in [9] and the result obtained in the preceding step, this lemma can be proved. The full proof is in Appendix B. □

To establish the convergence of the proposed ST-Q learning approach, we first evaluate the estimation error of $\hat{\theta}_{iq}$ in Algorithm (1a) with respect to θ_{iq} by characterizing the convergence performance of the global state estimation obtained by state tracking.

Lemma 2 (Convergence of Parameter Estimation)**.** *Under Assumptions 1-5, there exists $N < \infty$, such that*

(a) the global state estimation error is bounded above by some arbitrarily small $\delta > 0$, i.e., $\forall i \in [L]$, $\forall q$:

$$\|Z_i(t_q + N) - X(t_q + N)\| \leq \delta,$$

(b) the estimation error of θ_i in (15) *is bounded above by some arbitrarily small $\xi > 0$ when q is large enough, i.e., $\forall i \in [L]$:*

$$\|\theta_{iq} - \hat{\theta}_{iq}\| \leq \xi,$$

where $\hat{\theta}_{iq}$ is an estimate obtained by the ST-based approach and θ_{iq} is obtained with full observations. Note that $\hat{\theta}_{iq} = \hat{\theta}_{iq}(N)$, $\theta_{iq} = \theta_{iq}(N)$.

Proof Sketch. (a) First define $\epsilon_{ik}(t) = \sum_{j\in\mathcal{N}_k} w_{ij}(x_k(t) - x_k(t-1))$ and $\bar{x}_{av,k}(t) = \frac{1}{L}\sum_{j=1}^{L} \bar{x}_{jk}(t)$. The global state estimation error can be shown as

$$\begin{aligned}
\|Z_i(t) - X(t)\| =& \sqrt{\sum_{k=1}^{L} \|x_k(t) - \bar{x}_{ik}(t)\|_2^2} \\
\leq& \sum_{k=1}^{L} \|x_k(t) - \bar{x}_{ik}(t)\| \\
\leq& \sum_{k=1}^{L} \|x_k(t) - \bar{x}_{av,k}(t)\| + \sum_{k=1}^{L} \|\bar{x}_{av,k}(t) - \bar{x}_{ik}(t)\|,
\end{aligned}$$

where the first term can be analyzed by bringing in the definition of $\bar{x}_{av,k}$ and using the stability property of the controller. The second term is analyzed by formulating a perturbed consensus problem following the same line as in [25, Lemma 3]:

$$\begin{aligned}
\overline{x}_{ik}(t) &= \textstyle\sum_j w_{ij}\bar{x}_{jk}(t-1) + \epsilon_{ik}(t), \\
\epsilon_{ik}(t) &= \textstyle\sum_{j\in\mathcal{N}_k} w_{ij}(x_k(t) - x_k(t-1)).
\end{aligned}$$

(b) Recall the p-th gradient descent step:

$$\begin{aligned}
\theta_{iq}(p+1) &= \theta_{iq}(p) - \alpha\phi_i(t)\cdot\big(\theta_{iq}(p)^\top\phi_i(t) - g_i(t)\big), \\
\hat{\theta}_{iq}(p+1) &= \hat{\theta}_{iq}(p) - \alpha\hat{\phi}_i(t)\cdot\big(\hat{\theta}_{iq}(p)^\top\hat{\phi}_i(t) - \hat{g}_i(t)\big).
\end{aligned}$$

For convenience, define

$$\begin{aligned}
\Phi_i(t_q + N - \tau) &\triangleq I - \alpha\phi_i(t_q + N - \tau)\phi_i^\top(t_q + N - \tau), \\
\Pi_i(N) &\triangleq \textstyle\prod_{\tau=1}^{N} \Phi_i(t_q + N - \tau), \\
G_i(t_q + N - \tau) &\triangleq \alpha\phi_i(t_q + N - \tau)g_i(t_q + N - \tau).
\end{aligned}$$

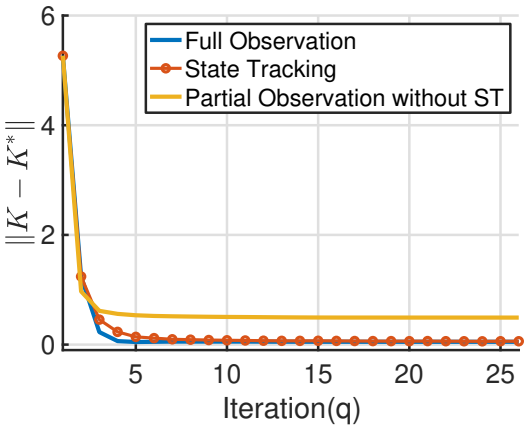
By using the deduction of $\theta_{iq}(p)$, we obtain the relationship between $\theta_{iq} = \theta_{iq}(N)$ and $\theta_{i(q-1)} = \theta_{iq}(0)$, as follows:

$$\begin{aligned}
\theta_{iq} &= \Pi_i(N)\theta_{i(q-1)} + \sum_{\tau=2}^{N} \Pi_i(\tau-1)G_i(t_q + N - \tau) + G_i(t_q + N - 1), \\
\hat{\theta}_{iq} &= \hat{\Pi}_i(N)\hat{\theta}_{i(q-1)} + \sum_{\tau=2}^{N} \hat{\Pi}_i(\tau-1)\hat{G}_i(t_q + N - \tau) + \hat{G}_i(t_q + N - 1).
\end{aligned}$$

It follows that the estimation error of $\|\hat{\theta}_{iq} - \theta_{iq}\|$ can be obtained by analyzing $\|\Phi_i - \hat{\Phi}_i\|$, $\|\Pi_i - \hat{\Pi}_i\|$ and $\|G_i - \hat{G}_i\|$ using the result from Lemma 2 (a).

The full proof of statements (a) and (b) is relegated to Appendices C and D, respectively. □

(a) Global controllers.



(b) Local controllers.

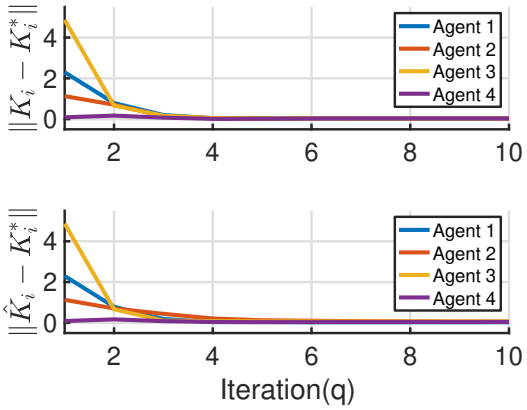


Figure 3: Convergence comparisons among three cases: ST based Q-learning, partial observation with no ST and full observation Q-learning. The controller obtained in three cases are denoted as $\hat{K}$, K_{partial} and K, respectively. The optimal controller is denoted as K^*. The upper figure in Figure (3b) is the local controllers convergence behavior with full observation while the lower one is the result with ST strategy.

Based on Lemma 2, we are now able to characterize the convergence performance of the ST based learning.

Theorem 1 (Convergence of the ST-Q learning)**.** *Suppose Assumption 1-5 are satisfied and K_{i1} is a stabilizing controller. Then, for any $\varepsilon_K > 0$, there exist $N < \infty$ and $q < \infty$, such that the ST-based Q-learning mechanism described in Algorithms (1a) and (1b) generates a sequence of stabilizing controllers $\{\hat{K}_{iq}\}_{q=1}^{\infty}$ that converge to the optimal controller, i.e., $\forall i \in [L]$:*

$$\|\hat{K}_{iq} - K_i^*\| \leq \varepsilon_K.$$

Proof. By Lemma 2, there exist $N < \infty$ and $q < \infty$ such that,

$$\|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\| \leq \xi.$$

Following [8], there exists a constant $k_0 > 0$, such that,

$$\|\hat{K}_{i(q)} - K_{i(q)}\| \leq k_0 \|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\|.$$

Hence, we obtain that

$$\|\hat{K}_{iq} - K_{iq}\| \leq k_0 \xi.$$

Besides, from Lemma 1, there exists $q < \infty$, such that

$$\|K_{iq} - K_i^*\| \leq \xi_k.$$

By using the triangle inequality, we obtain that

$$\|\hat{K}_{iq} - K_i^*\| = \|\hat{K}_{iq} - K_{iq} + K_{iq} - K_i^*\| \leq k_0 \xi + \xi_k \triangleq \varepsilon_K.$$

□

6 Experiments

In this section, we first introduce the experimental setup, and then evaluate the performance of the proposed ST-Q learning method. In particular, we compare our approach with two baselines: (i) distributed Q-learning with global state (DQG), and (ii) distributed Q-learning with partial observation of the global state (DQP) where the absent state information is set as 0, i.e., $x_j(t) = 0, \forall j \notin \mathcal{N}_i^c$. We further examine the impact of the hyper-parameters, including the step size α, interval N and the excitation noise η, on the performance of our approach to verify the assumptions made in Section 5.

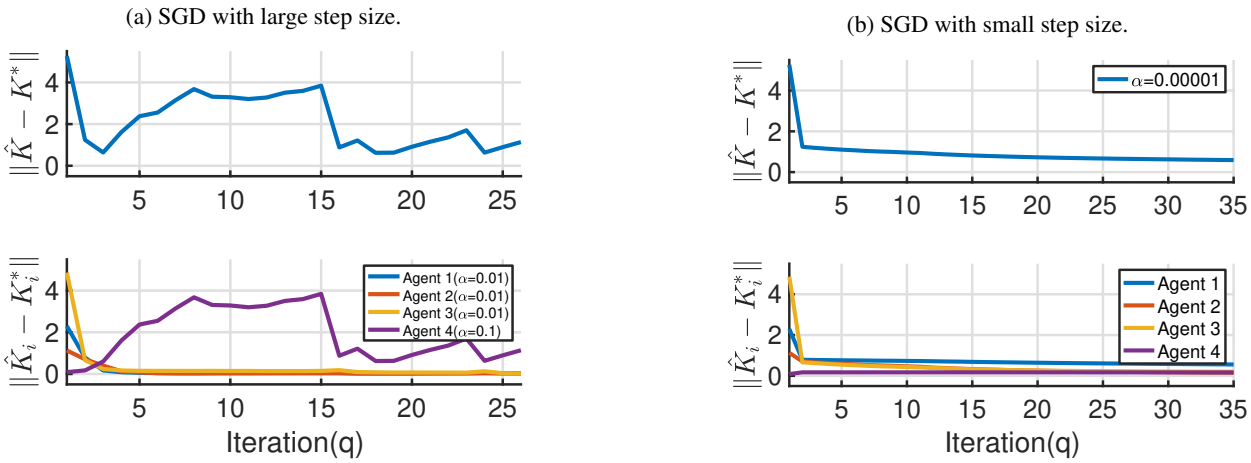


Figure 4: Comparison of different SGD step size in the ST based Q-learning. The upper figures in Figures (4a) and (4b) show the convergence behavior of the global controller while the lower figures are of the local controllers.

6.1 Experimental Setup

We consider the following global system with four agents, and the communication topology and interconnection topology as demonstrated in Fig. 1:

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ x_4(t+1) \end{bmatrix} = A \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} + B \begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \\ u_4(t) \end{bmatrix},$$

where the initial state for each agent is given as follows

$$x_i(0) = 0.01,\ i = 1, 2, 3, 4.$$

Note that the parameters can be chosen arbitrarily as long as they meet the assumptions in Section 5. The system parameters A and B are stabilizable and are set as

$$A = \begin{bmatrix} 0.2 & 0.4 & 0.1 & 0.01 \\ 0.4 & 0.2 & 0.3 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.4 \\ 0.2 & 0.1 & 0.5 & 0.3 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The weighting matrices for the LQR problem are selected as $P_i = R_i = 1$, for $i = 1, 2, 3, 4$. By solving the discrete Riccati equation, the optimal controller is obtained as,

$$K^* = \begin{bmatrix} 0.1223 & 0.2279 & 0.0779 & 0.0251 \\ 0.2267 & 0.1279 & 0.1823 & 0.0714 \\ 0.0796 & 0.1869 & 0.1944 & 0.2341 \\ 0.1212 & 0.0742 & 0.2838 & 0.1756 \end{bmatrix}.$$

Initial stable controller for Algorithms (1a) and (1b) is chosen to be:

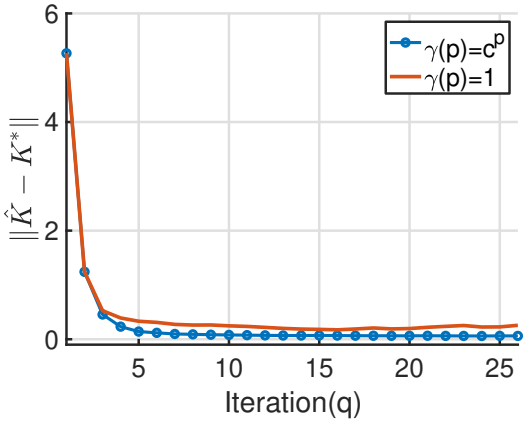
$$K_1 = \begin{bmatrix} 1 & 1 & 0.0004 & 2 \\ 1 & 0.2 & 1 & 0.1 \\ 4 & 0.1 & 1 & 3 \\ 0.2 & 0.1 & 0.3 & 0.2 \end{bmatrix}.$$

The weight matrix for the communication graph $\mathcal{N}_i^c$ is:

$$W = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.3 & 0.2 & 0 \\ 0 & 0.2 & 0.2 & 0.6 \\ 0 & 0 & 0.6 & 0.4 \end{bmatrix}.$$

The experiments are carried out with $N = 1000$ and step size $\alpha = 0.01$. Follow the approach in [27], the excitation noise $\eta_i(t)$ for Agent i is designed as $\big(b_i \cdot \text{rand}(-1, 1) + a_i \cdot \sum_{\omega=1}^{15} \sin(\omega t)^3 \cos(\omega t)\big) \cdot \upsilon(p)$, where the decaying factor $\upsilon(p) = 0.9999^p$ and $a_i, b_i \geq 0$ are constants.

(a) Decaying excitation noise.



(b) Convergence of $\hat{K} - K$.

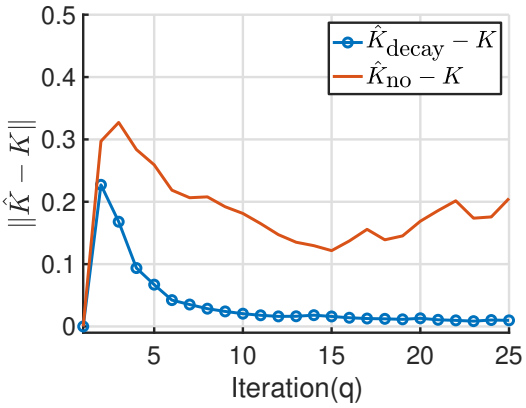
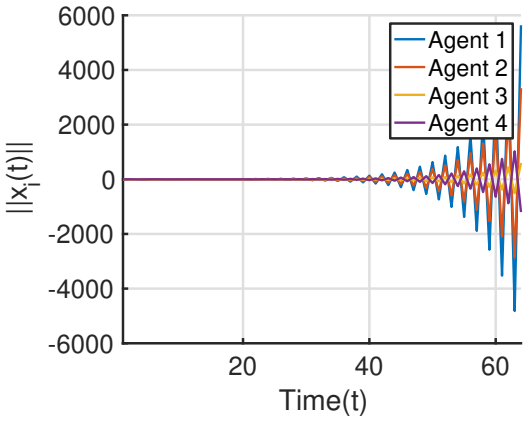


Figure 5: Comparison between two types of excitation noise in the ST based Q-learning. $\hat{K}_{\text{decay}}$ is the controller obtained with decaying factor and $\hat{K}_{\text{no}}$ is with no decaying factor.

(a) ST-Q with $N = 50$.



(b) ST-Q with different N.

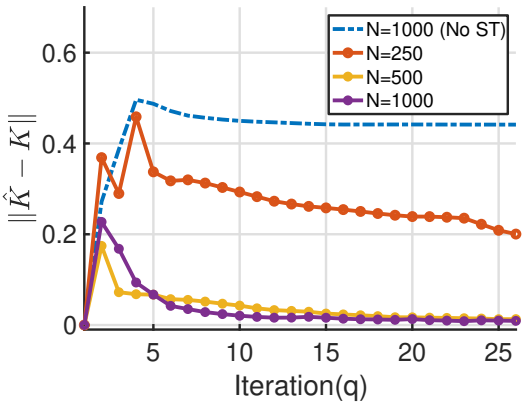


Figure 6: Comparisons of different N in the ST-Q learning algorithms.

6.2 Convergence of the ST-Q learning

We first characterize the convergence performance of the proposed ST-Q learning approach. As shown in Fig. 3a, the controller obtained by the ST-Q learning approach eventually converges to the optimal controller obtained by DQG, which clearly outperforms DQP. Note that all three approaches converge quickly. Moreover, Fig. 3b further demonstrates the convergence performance of the local controller $\hat{K}_i$ at each agent i compared with DQG, i.e., each agent in the ST-Q learning almost has the same convergence behaviour as in DQG. We also evaluate the gap between the controller $\hat{K}$ obtained by the ST-Q learning and the controller K obtained by DQG in the policy iteration, compared with that for the controller K_{partial} obtained by DQP. It can be seen from Fig. 6b that the gap $\|\hat{K} - K\|$ quickly converges to 0, while there exists a significant gap between K_{partial} and K (dashed line). These results together indicate that the proposed ST-Q learning approach can achieve comparable performance with DQG, corroborating the benefits by using state tracking to facilitate an accurate global state estimation in distributed Q-learning.

6.3 Impact of Hyper-Parameters

We first evaluate the impact of the step size α on the convergence of the controller $\hat{K}_i$ obtained by the ST-Q learning. In contrast to the step size 0.01 used in Fig 3, the divergence of the local controller obtained by the ST-Q learning may occur with a larger step size (Fig. 4a), while a smaller step size may result in slower convergence rate (Fig. 4b).

To examine the impact of the excitation noise, we compare the controller convergence performance under two different cases: (i) the noise is decaying as in Assumption 4, and (ii) the noise is not decaying. As demonstrated in Fig. 5a and Fig. 5b, the controller $\hat{K}$ obtained by the ST-Q learning may not converge when the excitation noise is not decaying, verifying the necessity of Assumption 4 to guarantee the convergence of the proposed ST-Q learning approach.

Clearly, the performance of the ST-Q learning depends on the estimation accuracy of $\hat{\theta}_{iq}$ in the policy evaluation step, which is directly affected by the value of N. Intuitively, as N increases, the estimation accuracy of $\hat{\theta}_{iq}$ improves accordingly, leading to a better performance of the ST-Q learning. Fig. 6 illustrates the impact of N on the performance of the ST-Q learning. As expected, when N is not large enough, the estimated controller may destabilize the system as shown in Fig. 6a due to the lack of adequate samples needed for achieving a better $\hat{\theta}_{iq}$. And Fig. 6b indicates that the larger N is, the better the performance of the ST-Q learning is. When N is large enough, the statement in Lemma 2 where the difference of the estimate $\hat{\theta}_{iq}$ obtained by the ST-based method and the estimate θ_{iq} obtained by the full observation method is decreasing along with the policy update, is verified in Fig. 6b.

7 Conclusions and Future Work

This work investigates a distributed multi-agent LQR control setup in a networked environment, in which the system dynamics, including the dynamics coupling graph is unknown. Each agent makes individual decisions based on its local observation and messages passed by its neighbors over the communication graph. Within this setting, we propose a multi-agent State Tracking based Q-learning method. Further, the asymptotic analysis on the convergence of the proposed algorithm is provided under mild assumptions. Empirically, in evaluation on an interconnected system, we demonstrate that the proposed ST-Q learning method outperforms the classic Q-learning with only the partial observation and yields the same optimal controller as the full observation setting. In future work, we shall consider more complicated communication settings, e.g., (i) communication delay in the network; (ii) time-varying graph. Moreover, it is also of interest to quantify the sampling complexity and the convergence rate of the proposed algorithm.

References

[1] Maria Carmela De Gennaro and Ali Jadbabaie. Decentralized control of connectivity for multi-agent systems. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 3628–3633. IEEE, 2006.

[2] Derui Ding, Qing-Long Han, Zidong Wang, and Xiaohua Ge. A survey on model-based distributed control and filtering for industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 15(5):2483–2499, 2019.

[3] Paolo Massioni and Michel Verhaegen. Distributed control for identical dynamically coupled systems: A decomposition approach. *IEEE Transactions on Automatic Control*, 54(1):124–135, 2009.

[4] Gianluca Antonelli. Interconnected dynamic systems: An overview on distributed control. *IEEE Control Systems Magazine*, 33(1):76–88, 2013.

[5] Christian Conte, Colin N Jones, Manfred Morari, and Melanie N Zeilinger. Distributed synthesis and stability of cooperative distributed model predictive control for linear systems. *Automatica*, 69:117–125, 2016.

[6] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[7] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[8] Steven J Bradtke, B Erik Ydstie, and Andrew G Barto. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pages 3475–3479. IEEE, 1994.

[9] Vignesh Narayanan and Sarangapani Jagannathan. Distributed adaptive optimal regulation of uncertain large-scale interconnected systems using hybrid q-learning approach. *IET Control Theory & Applications*, 10(12):1448–1457, 2016.

[10] Amirhassan Fallah Dizche, Aranya Chakrabortty, and Alexandra Duel-Hallen. Sparse wide-area control of power systems using data-driven reinforcement learning. In *2019 American Control Conference (ACC)*, pages 2867–2872. IEEE, 2019.

[11] Ali Bidram, Frank L Lewis, and Ali Davoudi. Distributed control systems for small-scale power networks: Using multiagent cooperative control theory. *IEEE Control Systems Magazine*, 34(6):56–77, 2014.

[12] Francesco Borrelli and Tamás Keviczky. Distributed lqr design for identical dynamically decoupled systems. *IEEE Transactions on Automatic Control*, 53(8):1901–1912, 2008.

[13] Eleftherios E Vlahakis, Leonidas D Dritsas, and George D Halikias. Distributed lqr design for identical dynamically coupled systems: Application to load frequency control of multi-area power grid. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 4471–4476. IEEE, 2019.

[14] Wenjie Dong. Distributed optimal control of multiple systems. *International Journal of Control*, 83(10):2067–2079, 2010.

[15] Daniel Görges. Distributed adaptive linear quadratic control using distributed reinforcement learning. *IFAC-PapersOnLine*, 52(11):218–223, 2019.

[16] Yi Cheng and V Ugrinovskii. Gain-scheduled leader-follower tracking control for interconnected parameter varying systems. *International Journal of Robust and Nonlinear Control*, 26(3):461–488, 2016.

[17] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

[18] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018.

[19] Yan Zhang and Michael M Zavlanos. Distributed off-policy actor-critic reinforcement learning with policy consensus. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 4674–4679. IEEE, 2019.

[20] Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *arXiv preprint arXiv:1912.09135*, 2019.

[21] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[22] Siavash Alemzadeh and Mehran Mesbahi. Distributed q-learning for dynamically decoupled systems. In *2019 American Control Conference (ACC)*, pages 772–777. IEEE, 2019.

[23] Frank L Lewis and Draguna Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE circuits and systems magazine*, 9(3):32–50, 2009.

[24] Jan Willems. Least squares stationary optimal control and the algebraic riccati equation. *IEEE Transactions on Automatic Control*, 16(6):621–634, 1971.

[25] Angelia Nedić and Ji Liu. Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:77–103, 2018.

[26] Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation for multi-agent reinforcement learning. *arXiv preprint arXiv:1902.07393*, 2019.

[27] Graham C Goodwin and Kwai Sang Sin. *Adaptive filtering prediction and control*. Courier Corporation, 2014.

Appendix

Assumption 1 (Stabilizability)**.** *The system parameters* (A, B) *in* (2) *are stabilizable.*

Assumption 2 (Communication Connectivity)**.** *The communication graph* $\mathcal{G}^c$ *is connected and static.*

Assumption 3 (Weight Matrix)**.** *There exists a positive constant* η *such that the weight matrix* $W = [w_{ij}] \in \mathbb{R}^{L\times L}$ *is doubly stochastic and* $w_{ij} \geq \eta$ *if* $j \in \mathcal{N}_i^c$*. In particular,* $w_{ii} \geq \eta$*,* $\forall i \in [L]$*.*

Assumption 4 (Decaying Excitation Noise)**.** *The input noise* $\eta(t)$ *is with the decaying factor* $\upsilon(p)$*,*

$$\begin{aligned} \upsilon(p) &= c^p, \quad 0 < c < 1, \\ \eta(t) &= \upsilon(p)\beta(t), \quad t_q \leq t < t_{q+1}. \end{aligned}$$

The input noise for the global system is $E(t) = \Upsilon(p)\boldsymbol{\beta}(t)$*. The system is further persistently excited with the input noise, i.e.,* $\forall i \in [L]$*,* $\forall q$*:*

$$mI \leq \textstyle\sum_{t=t_q}^{t_q+N-1} \phi_i(t)\phi_i^\top(t) \leq MI,$$

where $0 < m \leq M < \infty$*.*

Assumption 5 (Step Size of Gradient Descent)**.** *The step size in* (18) *is fixed and satisfies:* $0 < \alpha < 1/M$*.*

A Quadratic Structure of the Q-function

Substituting the system dynamics (1) into the definition of the Q-factor (8), we can obtain the quadratic structure of the Q-factor. Given a full observation of the global state, we have that

$$\begin{aligned} Q_i(x_i(t), u_i(t)) &= g_i(x_i(t), u_i(t)) + Q_i(x_i(t), u_i(t+1)) \\ &= x_i(t)^\top P_i x_i(t) + u_i(t)^\top R_i u_i(t) + x_i(t+1)^\top S_i x_i(t+1) \\ &= \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix}^\top \begin{pmatrix} \mathcal{A}_i{}^\top S_i \mathcal{A}_i + \tilde{P}_i & \mathcal{A}_i{}^\top S_i B_i \\ B_i{}^\top S_i \mathcal{A}_i & B_i{}^\top S_i B_i + R_i \end{pmatrix} \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix} \\ &= \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix}^\top \begin{pmatrix} H_{11,i} & H_{12,i} \\ H_{21,i} & H_{22,i} \end{pmatrix} \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix} \\ &= \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix}^\top H_i \begin{pmatrix} X(t) \\ u_i(t) \end{pmatrix} \\ &\triangleq y_i(t)^\top \theta_i, \\ \tilde{P}_i &= \begin{bmatrix} 0_{n\times n} & \dots & 0 \\ \vdots & P_i & \vdots \\ 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{Ln\times Ln}, \end{aligned}$$

where $\tilde{P}_i$ is a diagonal matrix with the (i,i)-th block set to be P_i. $y_i(t) = [x_1^2(t), x_1(t)x_2(t), \cdots, x_L(t)u_i(t), u_i^2(t)]$ is a vector consisting of all the quadratic basis over the elements in $[X(t); u_i(t)]$. Since H_i is symmetric, it suffices to use $\theta_i \in \mathbb{R}^{(Ln+m)(Ln+m+1)/2}$ to represent the unknown parameters, i.e., the elements of θ_i are the upper right triangle of H in the correct order. Moreover, $\mathcal{A}_i$ is a row vector which stacks subvectors $\{A_{ij}\}_{j\in[L]}$.

B Proof of Lemma 1

Lemma 1 (Convergence of Distributed Q-learning with Full Observation)**.** *Suppose that Assumptions 1, 4, 5 are satisfied, and* K_{i1} *is a stabilizing controller. There exists* $N < \infty$ *such that the sequence of stabilizing controllers* $\{K_{iq}\}_{q=1}^\infty$ *generated by the Q-learning Policy Iteration mechanism with global state information converges, i.e.,* $\forall i \in [L]$*:*

$$\lim_{q\to\infty} \|K_{iq} - K_i^*\| = 0,$$

where K_i^* *is the optimal feedback controller.*

Proof. The proof of this lemma includes two steps. First, we demonstrate that by replacing recursive least squares (RLS) with stochastic gradient descent (SGD) in the adaptive policy iteration algorithm proposed in [8], the policy iteration algorithm also generates a sequence of stabilizing controls converging to the optimal in the single agent case.

Observe that in [8], only the intermediate result Lemma 2 requires the property of the RLS estimation. Thus we need to prove that the SGD estimation also has the same property. Recall Lemma 2 in [8],

Lemma (Lemma 2 [8]). *If $\phi_i(t)$ is persistently excited and $N > N_0$, then we have*

$$\|\hat{\theta}_{iq} - \theta_{iq}\| \leq \varepsilon_N\big(\|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\| + \|\theta_{i(q-1)} - \theta_{iq})\|\big),$$

where $\lim_{N\to\infty} \varepsilon_N = 0$.

This lemma still holds when replacing RLS with SGD. Consider the q-th policy iteration where θ_{iq} is the true parameter vector for the Q-factor with control policy K_i. $\hat{\theta}_{iq} = \hat{\theta}_{iq}(N)$ is the estimate of θ_{iq} at the end of the q-th policy iteration. The initial estimate is the final value from the previous policy iteration, i.e., $\hat{\theta}_{iq}(0) = \hat{\theta}_{i(q-1)}(N)$. Recall the SGD algorithm,

$$\begin{aligned}
\hat{\theta}_{iq}(N) &= \hat{\theta}_{iq}(N-1) - \alpha\phi_i(t_q+N)\cdot\Big(\hat{\theta}_{iq}(N-1)^\top\phi_i(t_q+N) - g_i(t_q+N)\Big),\\
\hat{\theta}_{iq}(N) - \theta_{iq} &= \big(I - \alpha\phi_i(t_q+N)\phi_i^\top(t_q+N)\big)(\hat{\theta}_{iq}(N-1) - \theta_{iq}) - \underbrace{\alpha\phi_i(t_q+N)\cdot\Big(\phi_i^\top(t_q+N)\theta_{iq} - g_i(t_q+N)\Big)}_{=0}\\
&= \big(I - \alpha\phi_i(t_q+N)\phi_i^\top(t_q+N)\big)(\hat{\theta}_{iq}(N-1) - \theta_{iq})\\
&= \cdots\\
&= \prod_{\tau=t_q+1}^{t_q+N}\big(I - \alpha\phi_i(\tau)\phi_i^\top(\tau)\big)(\hat{\theta}_{i(q-1)} - \theta_{iq})\\
&= \prod_{\tau=t_q+1}^{t_q+N}\big(I - \alpha\phi_i(\tau)\phi_i^\top(\tau)\big)(\hat{\theta}_{i(q-1)} - \theta_{i(q-1)} + \theta_{i(q-1)} - \theta_{iq}),\\
\|\hat{\theta}_{iq}(N) - \theta_{iq}\| &\leq \|\prod_{\tau=t_q+1}^{t_q+N}\big(I - \alpha\phi_i(\tau)\phi_i^\top(\tau)\big)\|\cdot\big(\|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\| + \|\theta_{i(q-1)} - \theta_{iq})\|\big)\\
&\triangleq \varepsilon_N\big(\|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\| + \|\theta_{i(q-1)} - \theta_{iq})\|\big),
\end{aligned}$$

where step size α satisfies Assumption 5 and $\phi_i(t)$ satisfies Assumption 4. Hence we obtain that

$$\lim_{N\to\infty}\varepsilon_N = 0.$$

Second, we demonstrate that Lemma 1 holds when agents have full observation of the global state. Under this setting, [9] considers a policy iteration algorithm with the RLS estimation method and further provides the convergence proof by utilizing the single agent result in [8]. By following the same approach in [9] and the results obtained in the preceding step, we are able to obtain the desired result in Lemma 1. See [9, Theorem 1] for detailed proof.

□

C Proof of Lemma 2 (a)

Lemma 2 (Convergence of Parameter Estimation). *Under Assumptions 1-5, there exists $N < \infty$, such that*

(a) the global state estimation error is bounded above by some arbitrarily small $\delta > 0$, i.e., $\forall i \in [L]$, $\forall q$:

$$\|Z_i(t_q+N) - X(t_q+N)\| \leq \delta,$$

(b) the estimation error of θ_i in (15) *is bounded above by some arbitrarily small $\xi > 0$ when q is large enough, i.e., $\forall i \in [L]$:*

$$\|\theta_{iq} - \hat{\theta}_{iq}\| \leq \xi,$$

where $\hat{\theta}_{iq}$ is an estimate obtained by the ST-based approach and θ_{iq} is obtained with full observations. Note that $\hat{\theta}_{iq} = \hat{\theta}_{iq}(N)$, $\theta_{iq} = \theta_{iq}(N)$.

Proof. Before proceeding to the proof of Lemma 2, we first characterize the change of the agents' state (i.e., $x_i(t+1) - x_i(t)$) during the policy iteration. Consider Agent i's state in the q-th policy iteration. Notice that $t = t_q + p$, where t_q is the time index at the start of the q-th policy iteration and p counts the number of time steps from the start of the q-th policy iteration, such that $p \in [1, N]$ and $t \in [t_q, t_q + N]$. Assume that $X_q = X(t_q)$ is the initial global state for the q-th policy iteration. Following the system dynamics defined in (2), we obtain the global state after p steps of policy evaluation as,

$$\begin{aligned} X(t_q) &= X_q, \\ X(t_q+1) &= AX_q + B(K_qX_q + E(1)) = (A+BK_q)X_q + BE(1), \\ X(t_q+p) &= (A+BK_q)^pX_q + \sum_{\tau=1}^{p}(A+BK_q)^{p-\tau}BE(\tau). \end{aligned}$$

Suppose Assumption 4 holds and correspondingly we have

$$\begin{aligned} X_q &= X(t_{q-1}+N) \\ &= (A+BK_{q-1})^N X_{q-1} + \sum_{\tau=1}^{N}(A+BK_{q-1})^{N-\tau}BE(\tau) \\ &= (A+BK_{q-1})^N X_{q-1} + \sum_{\tau=1}^{N}(A+BK_{q-1})^{N-\tau}B\Upsilon(\tau)\boldsymbol{\beta}(\tau) \\ &= (A+BK_{q-1})^N X_{q-1} + \sum_{\tau=1}^{N}(A+BK_{q-1})^{N-\tau}c^{\tau}B\boldsymbol{\beta}(\tau), \end{aligned}$$

which indicates that $\|X_q\| \to 0$ as $N \to \infty$.

Hence, we can obtain that

$$\begin{aligned} X(t_q+p+1) - X(t_q+p) =&((A+BK_q)^{p+1} - (A+BK_q)^p)X_q \\ &+ (A+BK_q)^pBE(1) + \sum_{\tau=1}^{p}(A+BK_q)^{p-\tau}B(E(\tau+1) - E(\tau)). \end{aligned}$$

We further define

$$\begin{aligned} r &\triangleq rank(X(t_q+p+1) - X(t_q+p)), \\ \|(A+BK_q)^k\|_2 &= \sigma_{max}^k(A+BK_q) < 1. \end{aligned}$$

Let $\|C\|_F$ denote the Frobenius norm of a matrix $C \in \mathbb{R}^{m\times n}$, i.e., $\|C\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2}$. Now, we have,

$$\begin{aligned} \|X(t_q+p+1) - X(t_q+p)\|_F \leq& \sqrt{r}\|X(t_q+p+1) - X(t_q+p)\|_2 \\ \leq& \sqrt{r}\|(A+BK_q)^{p+1} - (A+BK_q)^p\|_2\|X_q\|_2 \\ &+ \sqrt{r}\|(A+BK_q)^p\|_2\|E(1)\|_2 + \sqrt{r}\|\sum_{\tau=1}^{p}(A+BK_q)^{p-\tau}B(E(\tau+1) - E(\tau))\|_2 \\ \leq& \sqrt{r}\|(A+BK_q)^{p+1} - (A+BK_q)^p\|_2\|X_q\|_2 + \sqrt{r}\|(A+BK_q)^p\|_2\|E(1)\|_2 \\ &+ \sqrt{r}\sum_{\tau=1}^{p}\left(\|(A+BK_q)^{p-\tau}\|_2\|B(E(\tau+1) - E(\tau))\|_2\right) \\ \leq& \sqrt{r}\|(A+BK_q)^{p+1} - (A+BK_q)^p\|_2\|X_q\|_2 + \sqrt{r}\sigma_{max}^t\Upsilon(1)\|\boldsymbol{\beta}(1)\|_2 \\ &+ \sqrt{r}\sum_{\tau=1}^{p}(\sigma_{max}^{p-\tau}\|B(c^{\tau+1}\boldsymbol{\beta}(\tau+1) - c^{\tau}\boldsymbol{\beta}(\tau))\|_2) \\ \leq& \sqrt{r}\|(A+BK_q)^{p+1} - (A+BK_q)^p\|_2\|X_q\|_2 + \sqrt{r}\sigma_{max}^p\Upsilon(1)\|\boldsymbol{\beta}(1)\|_2 \\ &+ c_\beta\sqrt{r}p\cdot\max\{c,\sigma_{max}\}^p \\ \triangleq& \delta_x, \end{aligned}$$

where $c_\beta = \max_\tau \|B(c\boldsymbol{\beta}(\tau+1) - \boldsymbol{\beta}(\tau))\|$. c_β is finite since the injected noise satisfies the PE condition in Assumption 4. Notice that with a fixed q, when $p \to \infty$, $\delta_x \to 0$. Note that K_q is a stable controller, such that $\sigma_{max} < 1$.

Therefore, there exists $N < \infty$ and we can choose $N_0 \in (0, N)$ such that when $t \in [t_q + N_0, t_{q+1}]$ (p is large enough), the difference between two adjacent states of Agent i is bounded from above by some arbitrary small $\delta_x > 0$, i.e., $\forall i \in [L]$:

$$\|x_i(t) - x_i(t-1)\| \leq \delta_x. \tag{19}$$

Now we are ready to prove Lemma 2 (a).

Define $\epsilon_{ik}(t) = \sum_{j \in \mathcal{N}_k} w_{ij}(x_k(t) - x_k(t-1))$ and $\bar{x}_{av,k}(t) = \frac{1}{L}\sum_{j=1}^{L} \bar{x}_{jk}(t)$ (the average estimation towards Agent k at time instance t). Further, let $d_k = |\mathcal{N}_k|$ denote the amount of communication neighbors of Agent k in the communication network. The norm of the difference between the ST-based global state estimates $Z_i(t)$ and the true global state $X(t)$ can be shown as

$$\begin{aligned}
\|Z_i(t) - X(t)\|_F &= \sqrt{\sum_{k=1}^{L} \|x_k(t) - \bar{x}_{ik}(t)\|_2^2} \\
&\leq \sum_{k=1}^{L} \sqrt{\|x_k(t) - \bar{x}_{ik}(t)\|^2} \\
&= \sum_{k=1}^{L} \|x_k(t) - \bar{x}_{av,k}(t) + \bar{x}_{av,k}(t) - \bar{x}_{ik}(t)\| \\
&\leq \Big(\underbrace{\sum_{k=1}^{L} \|x_k(t) - \bar{x}_{av,k}(t)\|}_{①} + \underbrace{\sum_{k=1}^{L} \|\bar{x}_{av,k}(t) - \bar{x}_{ik}(t)\|}_{②} \Big).
\end{aligned} \tag{20}$$

First we consider term ① in (20),

$$\begin{aligned}
\sum_{k=1}^{L} \|x_k(t) - \bar{x}_{av,k}(t)\| &= \sum_{k=1}^{L} \|x_k(t) - \frac{1}{L}\sum_{j=1}^{L} \bar{x}_{jk}(t)\| \\
&= \sum_{k=1}^{L} \|x_k(t) - \Big(\bar{x}_{av,k}(0) + \frac{d_k}{L}x_k(t) - \frac{d_k}{L}x_k(0)\Big)\| \\
&\leq L \max_k \{\|x_k(t)\|\} \triangleq w(t),
\end{aligned} \tag{21}$$

where we use the deduction of $\bar{x}_{av,k}(t)$:

$$\begin{aligned}
\bar{x}_{av,k}(t) &= \frac{1}{L}\sum_{j=1}^{L} \bar{x}_{jk}(t) \\
&= \frac{1}{L}\sum_{j=1}^{L}\sum_{u=1}^{L} w_{ju}\hat{x}_{uk}(t) \\
&= \frac{1}{L}\sum_{u=1}^{L} \hat{x}_{uk}(t) \\
&= \frac{1}{L}\Big(\sum_{u \in \mathcal{N}_k} x_k(t) + \sum_{u \notin \mathcal{N}_k} \bar{x}_{uk}(t-1)\Big) \\
&= \bar{x}_{av,k}(t-1) + \frac{d_k}{L}x_k(t) - \frac{d_k}{L}x_k(t-1) \\
&= \bar{x}_{av,k}(0) + \frac{d_k}{L}(x_k(t) - x_k(0)).
\end{aligned} \tag{22}$$

Now, consider term (2): Let Assumptions 2, 3 hold and rewrite $\bar{x}_{ik}(t)$ into the following format,

$$\begin{aligned}\bar{x}_{ik}(t) &= \sum_{j=1}^{L} w_{ij}\hat{x}_{jk}(t) \\ &= \sum_{j\notin\mathcal{N}_k} w_{ij}\hat{x}_{jk}(t) + \sum_{j\in\mathcal{N}_k} w_{ij}\hat{x}_{jk}(t) \\ &= \underbrace{\sum_{j\notin\mathcal{N}_k} w_{ij}\bar{x}_{jk}(t-1)}_{\text{Weighted Estimation}} + \underbrace{\sum_{j\in\mathcal{N}_k} w_{ij}x_k(t)}_{\text{Weighted True State}} \\ &= \sum_{j} w_{ij}\bar{x}_{jk}(t-1) + \underbrace{\sum_{j\in\mathcal{N}_k} w_{ij}(x_k(t)-x_k(t-1))}_{\text{Perturbation } \triangleq \epsilon_{ik}(t)}.\end{aligned} \tag{23}$$

Following the same line as in [25], we reformulate (23) as a perturbed consensus problem:

$$\begin{aligned}\overline{x}_{ik}(t) &= \sum_{j} w_{ij}\bar{x}_{jk}(t-1) + \epsilon_{ik}(t), \\ \epsilon_{ik}(t) &= \sum_{j\in\mathcal{N}_k} w_{ij}(x_k(t)-x_k(t-1)).\end{aligned} \tag{24}$$

For ease of exposition, we rewrite the evolution of the iterates $\overline{x}_{ik}(t)$ in a matrix form. For any coordinate index $l \in [n]$ (n is the dimension of the state vector), we can have the following for the $l-$th coordinate(denoted by a superscripts):

$$\overline{x}^l_{ik}(t) = \sum_{j} w_{ij}\bar{x}^l_{jk}(t-1) + \epsilon^l_{ik}(t), \quad \forall l \in [n].$$

Define $\bar{x}_k(t-1) = \begin{bmatrix}\bar{x}_{1k}(t-1)\\ \vdots \\ \bar{x}_{Lk}(t-1)\end{bmatrix}$, $\epsilon_p(t) = \begin{bmatrix}\epsilon_{1k}(t)\\ \vdots \\ \epsilon_{Lk}(t)\end{bmatrix}$.

Next, we stack all of the $l-$th coordinates in a column vector, denoted by $\overline{x}^l_k(t)$, i.e.,

$$\overline{x}^l_k(t) = \begin{bmatrix}\overline{x}^l_{1k}(t)\\ \overline{x}^l_{2k}(t)\\ \vdots \\ \overline{x}^l_{Lk}(t)\end{bmatrix} = W\bar{x}^l_k(t-1) + \epsilon^l_k(t).$$

Moreover, by stacking the column vectors $\overline{x}^l_k(t),\ l \in [n]$ into a matrix $\overline{\mathbf{x}}_k(t)$, we further build up the perturbation matrix $\mathbf{e}_k(t)$ from $\epsilon^l_k(t),\ l \in [n]$

$$\overline{\mathbf{x}}_k(t) = \begin{bmatrix}\overline{x}^1_k(t) & \overline{x}^2_k(t) & \cdots & \overline{x}^n_k(t)\end{bmatrix} = W\overline{\mathbf{x}}_k(t-1) + \mathbf{e}_k(t) \quad \forall t \geq 0. \tag{25}$$

Using the recursion, from Eqn. (25) we see that, for all $t_q \leq t \leq t_{q+1}$,

$$\begin{aligned}\overline{\mathbf{x}}_k(t) &= W\overline{\mathbf{x}}_k(t-1) + \mathbf{e}_k(t) \\ &= W(W\overline{\mathbf{x}}_k(t-2) + \mathbf{e}_k(t-1)) + \mathbf{e}_k(t) \\ &= (W)^2\overline{\mathbf{x}}_k(t-2) + (W)^1\mathbf{e}_k(t-1) + \mathbf{e}_k(t) \\ &= \cdots \\ &= (W)^p\overline{\mathbf{x}}_k(t_q) + \sum_{\tau=1}^{p-1}(W)^\tau\mathbf{e}_k(t-\tau) + \mathbf{e}_k(t).\end{aligned} \tag{26}$$

By multiplying both sides of (26) with matrix $\frac{1}{L}\mathbf{1}\mathbf{1}^\top$, we have

$$\begin{aligned}\frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t) &= \frac{1}{L}\mathbf{1}\mathbf{1}^\top(W)^p\overline{\mathbf{x}}_k(t_q) + \Big(\sum_{\tau=1}^{p-1}\frac{1}{L}\mathbf{1}\mathbf{1}^\top(W)^\tau\mathbf{e}_k(t-\tau))\Big) + \frac{1}{L}\mathbf{1}\mathbf{1}^\top\mathbf{e}_k(t)\\ &= \frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t_q) + \sum_{\tau=1}^{p-1}\frac{1}{L}\mathbf{1}\mathbf{1}^\top\mathbf{e}_k(t-\tau) + \frac{1}{L}\mathbf{1}\mathbf{1}^\top\mathbf{e}_k(t).\end{aligned}$$

Now, consider $\overline{\mathbf{x}}_k(t) - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t)$

$$\begin{aligned}\overline{\mathbf{x}}_k(t) - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t) =& \Big((W)^p - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\Big)\overline{\mathbf{x}}_k(t_q) + \sum_{\tau=1}^{p-1}\Big((W)^\tau - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\Big)\mathbf{e}_k(t-\tau)\\ &+ \Big(I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\Big)\mathbf{e}_k(t).\end{aligned} \tag{27}$$

By taking the F-norm of the both sides of (27), we obtain,

$$\begin{aligned}\|\overline{\mathbf{x}}_k(t) - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t)\|_F \leq& \|\big((W)^p - \mathbf{1}\mathbf{1}^\top\big)\overline{\mathbf{x}}_k(t_q)\|_F + \sum_{\tau=1}^{p-1}\|\big((W)^\tau - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\big)\mathbf{e}_k(t-\tau)\|_F\\ &+ \|\big(I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\big)\mathbf{e}_k(t)\|_F\\ \leq& \|\big((W)^p - \mathbf{1}\mathbf{1}^\top\big)\|_F\|\overline{\mathbf{x}}_k(t_q)\|_F + \sum_{\tau=1}^{p-1}\Big(\|(W)^\tau - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\|_F\|\mathbf{e}_k(t-\tau)\|_F\Big)\\ &+ \|\big(I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\big)\|_F\|\mathbf{e}_k(t)\|_F.\end{aligned} \tag{28}$$

The following lemma (Lemma 5 [25]) is required here.

Lemma 3. *Let the graph G^c satisfy Assumption 2 and let the weight matrix W satisfy Assumption 3. Then, for all $s \geq 0$,*

$$\Big([W^s]_{ij} - \frac{1}{L}\Big)^2 \leq \Big(1 - \frac{\eta}{2L^2}\Big)^{s-1}, \quad \forall i,j \in [L].$$

Based Lemma 3, we have

$$\|(W)^\tau - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\|_F = \sqrt{\sum_{i=1}^{L}\sum_{j=1}^{L}\Big([W^\tau]_{ij} - \frac{1}{L}\Big)^2} \leq L\sqrt{\Big(1 - \frac{\eta}{2L^2}\Big)^{\tau-1}}.$$

Following the fact that $\sqrt{1-\mu} \leq 1 - \frac{\mu}{2}, \forall \mu \in (0,1)$, we further have,

$$\|(W)^\tau - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\|_F \leq Lc_w^{\tau-1},\ c_w = 1 - \frac{\eta}{4L^2}.$$

For the norm $\|I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\|_F$, we have

$$\|I - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\|_F = \sqrt{L\Big(1 - \frac{1}{L}\Big)^2 + (L-1)L\frac{1}{L^2}} = \sqrt{L-1}. \tag{29}$$

Now, we obtain that

$$\|\overline{\mathbf{x}}_k(t) - \frac{1}{L}\mathbf{1}\mathbf{1}^\top\overline{\mathbf{x}}_k(t)\|_F \leq Lc_w^{p-1}\|\overline{\mathbf{x}}_k(t_q)\|_F + L\Big(\sum_{\tau=1}^{p-1}c_w^{\tau-1}\|\overline{\mathbf{e}}_k(t-\tau)\|_F\Big) + \sqrt{L-1}\|\overline{\mathbf{e}}_k(t)\|_F. \tag{30}$$

This equation is equivalent to

$$\sqrt{\sum_{i=1}^{L}\|\bar{x}_{ik}(t)-\bar{x}_{av,k}(t)\|^2}\leq Lc_w^{p-1}\sqrt{\sum_{i=1}^{L}\|\bar{x}_{ik}(t_q)\|^2}+L\Big(\sum_{\tau=t-p+1}^{t-1}c_w^{t-\tau-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(\tau)\|^2}\Big)+\sqrt{L-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(t)\|^2}. \tag{31}$$

Recall that $\epsilon_{ik}(t)=\sum_{j\in\mathcal{N}_k}w_{ij}(x_k(t)-x_k(t-1))$, $\bar{x}_{av,k}(t)=\frac{1}{L}\sum_{j=1}^{L}\bar{x}_{jk}(t)$ (the average estimation towards Agent k at time instance t). Now, we are ready to obtain the upper bound on $\sum_{k=1}^{L}\|\bar{x}_{av,k}(t)-\bar{x}_{ik}(t)\|$:

$$\begin{aligned}\sum_{k=1}^{L}\|\bar{x}_{av,k}(t)-\bar{x}_{ik}(t)\|&=\sum_{k\notin\mathcal{N}_i}\|\bar{x}_{av,k}(t)-\bar{x}_{ik}(t)\|+\sum_{k\in\mathcal{N}_i}\|\bar{x}_{av,k}(t)-x_k(t)\|\\&\leq L^{\frac{1}{2}}\sqrt{\sum_{i=1}^{L}\|\bar{x}_{ik}(t)-\bar{x}_{av,k}(t)\|^2}+w(t)\\&\leq L^{\frac{3}{2}}c_w^{p-1}\sqrt{\sum_{i=1}^{L}\|\bar{x}_{ik}(t_q)\|^2}+L^{\frac{3}{2}}\Big(\sum_{\tau=t-p+1}^{t-1}c_w^{t-\tau-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(\tau)\|^2}\Big)\\&\quad+L^{\frac{1}{2}}\sqrt{L-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(t)\|^2}+(L-d_k)\|x_k(t)\|\\&\triangleq v(t),\end{aligned} \tag{32}$$

where the first inequality is based on the Hölder's inequality. Note that t denotes the time instance from the start of the algorithm, i.e., $t=t_q+p$. Thus, we can obtain the following result from (19)

$$\lim_{p\to\infty}\|\epsilon_{ik}(t)\|=\lim_{p\to\infty}\|\epsilon(p)\|=0.$$

Besides, the second term of (31) satisfies

$$\begin{aligned}&\sum_{\tau=t-p+1}^{t-1}c_w^{t-\tau-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(\tau)\|^2}\\&\leq\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}\|\epsilon_{ik}(\tau)\|\\&=\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}\frac{1}{\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}}\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}\|\epsilon_{ik}(\tau)\|\\&=\frac{L(1-c_w^{p-1})}{1-c_w}\Big(\frac{1}{\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}}\sum_{\tau=t-p+1}^{t-1}Lc_w^{t-\tau-1}\|\epsilon_{ik}(\tau)\|\Big).\end{aligned}$$

Let $p\to\infty$. We have

$$\lim_{p\to\infty}\sum_{\tau=t-p+1}^{t-1}c_w^{t-\tau-1}\sqrt{\sum_{i=1}^{L}\|\epsilon_{ik}(\tau)\|^2}\leq\frac{L}{1-c_w}\lim_{p\to\infty}\|\epsilon(p)\|=0, \tag{33}$$

where the right side follows Mazur's Lemma that any convex combination of a convergent sequence $\{\epsilon(p)\}$ converges to the same limit as the sequence itself.

Following the preceding results, finally we obtain that

$$\|Z_i(t)-X(t)\|\leq\Big(w(t)+v(t)\Big)\triangleq\delta(t),\quad t_q\leq t\leq t_{q+1}. \tag{34}$$

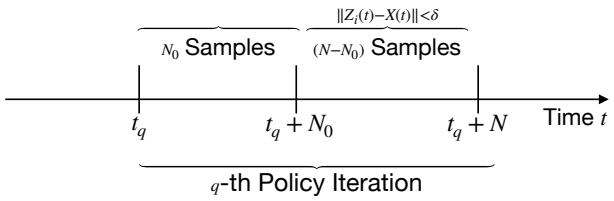


Figure 7: Illustration of the result in (35).

Note that when $t \to \infty$ ($p \to \infty$), $w(t) \to 0$ (see (21)), $v(t) \to 0$ (see (32)). Thus, for all $\delta > 0$, there exist N and $N_0 \in (0, N)$ such that, $\forall t \in [t_q + N_0, t_{q+1}]$,

$$\|Z_i(t) - X(t)\| \leq \delta, \ \forall i \in [L], \tag{35}$$

which indicates that with large enough N, we are able to collect $(N - N_0)$ samples with relative error δ (See Fig. 7).

□

D Proof of Lemma 2 (b)

Lemma 2 (Convergence of Parameter Estimation)**.** *Under Assumptions 1-5, there exists $N < \infty$, such that*

(a) the global state estimation error is bounded above by some arbitrarily small $\delta > 0$, i.e., $\forall i \in [L]$, $\forall q$:

$$\|Z_i(t_q + N) - X(t_q + N)\| \leq \delta,$$

(b) the estimation error of θ_i in (15) *is bounded above by some arbitrarily small $\xi > 0$ when q is large enough, i.e., $\forall i \in [L]$:*

$$\|\theta_{iq} - \hat{\theta}_{iq}\| \leq \xi,$$

where $\hat{\theta}_{iq}$ is an estimate obtained by the ST-based approach and θ_{iq} is obtained with full observations. Note that $\hat{\theta}_{iq} = \hat{\theta}_{iq}(N)$, $\theta_{iq} = \theta_{iq}(N)$.

Proof. Recall the SGD update rule in (18),

$$\begin{aligned}
\theta_{iq}(p+1) &= \theta_{iq}(p) - \alpha\phi_i(t) \cdot \Big(\theta_{iq}(p)^\top \phi_i(t) - g_i(t)\Big), \\
\hat{\theta}_{iq}(p+1) &= \hat{\theta}_{iq}(p) - \alpha\hat{\phi}_i(t) \cdot \Big(\hat{\theta}_{iq}(p)^\top \hat{\phi}_i(t) - \hat{g}_i(t)\Big).
\end{aligned} \tag{36}$$

We first define

$$\begin{aligned}
\Phi_i(t_q + N - \tau) &\triangleq I - \alpha\phi_i(t_q + N - \tau)\phi_i^\top(t_q + N - \tau), \\
\Pi_i(M) &\triangleq \prod_{m=1}^{M} \Phi_i(t_q + N - m), \\
G_i(t_q + N - \tau) &\triangleq \alpha\phi_i(t_q + N - \tau)g_i(t_q + N - \tau).
\end{aligned} \tag{37}$$

Next, we use recursion to obtain the relationship between $\theta_{iq} = \theta_{iq}(N)$ and $\theta_{i(q-1)} = \theta_{iq}(1)$,

$$\begin{aligned}
\theta_{iq} &= \Pi_i(N)\theta_{i(q-1)} + \sum_{\tau=2}^{N} \Pi_i(\tau - 1)G_i(t_q + N - \tau) + G_i(t_q + N - 1), \\
\hat{\theta}_{iq} &= \hat{\Pi}_i(N)\hat{\theta}_{i(q-1)} + \sum_{\tau=2}^{N} \hat{\Pi}_i(\tau - 1)\hat{G}_i(t_q + N - \tau) + \hat{G}_i(t_q + N - 1).
\end{aligned} \tag{38}$$

Therefore, we have

$$\begin{aligned}\|\theta_{iq}-\hat{\theta}_{iq}\| =&\|\Pi_i(N)\theta_{i(q-1)}-\hat{\Pi}_i(N)\hat{\theta}_{i(q-1)}\| \to \textcircled{1}\\ &+\|G_i(t_q+N-1)-\hat{G}_i(t_q+N-1)\| \to \textcircled{2}\\ &+\|\sum_{\tau=2}^{N}\Big(\Pi_i(\tau-1)G_i(t_q+N-\tau)-\hat{\Pi}_i(\tau-1)\hat{G}_i(t_q+N-\tau)\Big)\| \to \textcircled{3}.\end{aligned} \tag{39}$$

In order to utilize the result in Appendix C, we first explore the relationship between $\phi_i(t)$ and $\|Z_i(t)-X(t)+\hat{u}_i(t)-u_i(t)\|$. It can be shown that the norm of $y_i(t)$ is equivalent to the $F-$norm of the product of two matrices, i.e. $E_i(t)M_i(t)$,

$$\begin{aligned}y_i(t) =& \begin{bmatrix}x_1^2(t) & x_1(t)x_2(t) & x_1(t)x_3(t)\cdots u_i^2(t)\end{bmatrix}\\ =&\underbrace{\begin{bmatrix}x_1(t) & & & & \\ & x_2(t) & & & \\ & & \ddots & & \\ & & & x_L(t) & \\ & & & & u_i(t)\end{bmatrix}}_{\triangleq E_i(t)}\underbrace{\begin{bmatrix}x_1(t) & x_2(t) & \cdots & x_L(t) & u_i(t)\\ x_2(t) & x_3(t) & \cdots & u_i(t) & \\ \vdots & \vdots & & & \\ u_i(t) & & & & \end{bmatrix}}_{\triangleq M_i(t)}.\end{aligned} \tag{40}$$

Then, we can apply the result in Appendix C ($\delta(t)$ is defined in (34), $\hat{K}_{i(q-1)}$ denotes the controller obtained by using estimated global state, $(q-1)$ denotes the controller is updated in the last policy update). Note that $t\in[t_q,t_{q+1}]$ denotes the $q-$th policy evaluation time instant.

$$\begin{aligned}T_{q-1} &\triangleq \|\theta_{i(q-1)}-\hat{\theta}_{i(q-1)}\|,\\ \|u_i(t)-\hat{u}_i(t)\| &= \|\hat{K}_{i(q)}Z_i(t)-K_{i(q)}X(t)\|,\\ &\le \|\hat{K}_{i(q)}-K_{i(q)}\|\|X(t)\|+\|X(t)-Z_i(t)\|\|\hat{K}_{i(q)}\|,\\ &\le \kappa\|X(t)\|T_{q-1}+b_k\delta(t),\end{aligned} \tag{41}$$

where $b_k<\infty$ is the upper bound of $\|\hat{K}_{i(q-1)}\|$. The last inequality follows that

$$\begin{aligned}\hat{K}_{i(q)} &= -\hat{H}^{-1}_{i(q-1),22}\hat{H}_{i(q-1),21},\\ K_{i(q)} &= -H^{-1}_{i(q-1),22}H_{i(q-1),21}.\end{aligned}$$

Then, we have

$$\begin{aligned}\hat{K}_{i(q)}-K_{i(q)} &= -\hat{H}^{-1}_{i(q-1),22}\hat{H}_{i(q-1),21}+H^{-1}_{i(q-1),22}H_{i(q-1),21}\\ &= H^{-1}_{i(q-1),22}((H_{i(q-1),21}-\hat{H}_{i(q-1),21})+(\hat{H}_{i(q-1),22}-H_{i(q-1),22})\hat{H}^{-1}_{i(q-1),22}\hat{H}_{i(q-1),21}),\\ \|\hat{H}_{i(q-1),22}-H_{i(q-1),22}\| &\le \|\hat{\theta}_{i(q-1)}-\theta_{i(q-1)}\|,\\ \|H_{i(q-1),22}\| &\le \|\theta_{i(q-1)}\|.\end{aligned}$$

Since the estimated parameters are bounded and $\kappa>0$ is a finite constant, it follows that

$$\|\hat{K}_{i(q)}-K_{i(q)}\|\le\kappa\|\hat{\theta}_{i(q-1)}-\theta_{i(q-1)}\|.$$

Therefore,

$$\begin{aligned}\|g_i(t)-\hat{g}_i(t)\| =& \|\langle x_i(t),x_i(t)\rangle_P+\langle u_i(t),u_i(t)\rangle_R-\langle\hat{x}_i(t),\hat{x}_i(t)\rangle_P+\langle\hat{u}_i(t),\hat{u}_i(t)\rangle_R\|\\ =&\|\langle x_i(t)-\hat{x}_i(t),x_i(t)\rangle_P+\langle\hat{x}_i(t),x_i(t)-\hat{x}_i(t)\rangle_P\\ &+\langle u_i(t)-\hat{u}_i(t),u_i(t)\rangle_R-\langle\hat{u}_i(t),u_i(t)-\hat{u}_i(t)\rangle_R\|\\ \le&\lambda_{max}(P)(2b_x\delta(t))+2\lambda_{max}(R)(\kappa\|X(t)\|T_{q-1}+b_k\delta(t))b_u\\ =&c_1\delta(t)+c_2\|X(t)\|T_{q-1},\\ c_1 \triangleq&2b_x\lambda_{max}(P)+b_kb_u\lambda_{max}(R)<\infty,\\ c_2 \triangleq&\lambda_{max}(R)\kappa<\infty,\\ \langle x_i(t),x_i(t)\rangle_P =&x_i(t)^TPx_i(t),\end{aligned} \tag{42}$$

where $b_x < \infty$ is the upper bound of the system state $\|x_i(t)\|$ and $b_u < \infty$ is the upper bound of the system input $\|u_i(t)\|$. $\lambda_{max}(P)$ is the largest eigenvalue of matrix P and $\lambda_{max}(R)$ is the largest eigenvalue of matrix R. Hence, we obtain the following inequality with regard to E_i and M_i

$$\begin{aligned}\|E_i(t) - \hat{E}_i(t)\| &= \|Z_i(t) - X(t)\| + \|u_i(t) - \hat{u}_i(t)\| \\ &\leq 2b_k\delta(t) + \kappa\|X(t)\|T_{q-1} \\ &\triangleq c_3\delta(t) + c_4\|X(t)\|T_{q-1}, \\ \|M_i(t) - \hat{M}_i(t)\| &\leq L\|Z_i(t) - X(t)\| + (L+1)\|u_i(t) - \hat{u}_i(t)\| \\ &\leq (2L+1)b_k\delta(t) + (L+1)\kappa\|X(t)\|T_{q-1} \\ &\triangleq c_5\delta(t) + c_6\|X(t)\|T_{q-1}.\end{aligned} \tag{43}$$

Furthermore, we can obtain that

$$\begin{aligned}\|\phi_i(t) - \hat{\phi}_i(t)\| &= \|E_i(t)M_i(t) - E_i(t+1)M_i(t+1) - \hat{E}_i(t)\hat{M}_i(t) + \hat{E}_i(t+1)\hat{M}_i(t+1)\| \\ &\leq \|E_i(t)M_i(t) - \hat{E}_i(t)\hat{M}_i(t)\| + \|E_i(t+1)M_i(t+1) - \hat{E}_i(t+1)\hat{M}_i(t+1)\| \\ &\leq (c_3b_m + c_5b_x)(\delta(t) + \delta(t+1)) + (c_4b_m + c_6b_x)(\|X(t)\| + \|X(t+1)\|))T_{q-1} \\ &\triangleq c_7(\delta(t) + \delta(t+1)) + c_8(\|X(t)\| + \|X(t+1)\|)T_{q-1}.\end{aligned} \tag{44}$$

Thus,

$$\begin{aligned}\|(I - \alpha\hat{\phi}_i(t)\hat{\phi}_i^\top(t)) - (I - \alpha\phi_i(t)\phi_i^\top(t))\| &= \|\alpha\hat{\phi}_i(t)\hat{\phi}_i^\top(t) - \alpha\phi_i(t)\phi_i^\top(t)\| \\ &\leq c_9(\delta(t) + \delta(t+1)) + c_{10}(\|X(t)\| + \|X(t+1)\|)T_{q-1}.\end{aligned} \tag{45}$$

Now, we are ready to analyze term ①: $\|\Pi_i(N)\theta_{i(q-1)} - \hat{\Pi}_i(N)\hat{\theta}_{i(q-1)}\|$. Note that

$$\begin{aligned}\|\Pi(n)\| &\leq c_\pi^n \leq \bar{c}_\pi^n, \\ \|\hat{\Pi}(n)\| &\leq \hat{c}_\pi^n \leq \bar{c}_\pi^n, \\ \bar{c}_\pi &= \max\{c_\pi, \hat{c}_\pi\},\end{aligned} \tag{46}$$

where $c_\pi = \max_t\{\|\Phi_i(t)\|\} < 1$ and $\hat{c}_\pi = \max_t\{\|\hat{\Phi}_i(t)\|\} < 1$.

Hence, we have

$$\|\Pi_i(N) - \hat{\Pi}_i(N)\| = \|\prod_{m=1}^{N}\Phi_i(t_q + N - m) - \prod_{m=1}^{N}\hat{\Phi}_i(t_q + N - m)\| \leq \bar{c}_\pi^N. \tag{47}$$

Now, we consider

$$\begin{aligned}\|\Pi(N)\theta_{i(q-1)} - \hat{\Pi}(N)\hat{\theta}_{i(q-1)}\| &\leq \|\Pi(N) - \hat{\Pi}(N)\|\|\hat{\theta}_{i(q-1)}\| + \|\hat{\theta}_{i(q-1)} - \theta_{i(q-1)}\|\|\Pi(N)\| \\ &\leq c_{11}\bar{c}_\pi^N + \|\Pi(N)\|T_{q-1} \\ &\triangleq \xi_1.\end{aligned} \tag{48}$$

Notice that as $N \to \infty$, $\xi_1 \to 0$.

Similarly, we analyze term ②: $\|G(t_q + N - 1) - \hat{G}(t_q + N - 1)\|$,

$$\begin{aligned}\|G(t_q + N - 1) - \hat{G}(t_q + N - 1)\| =& \alpha\|\phi^\top(t_q + N - 1)g(t_q + N - 1) - \hat{\phi}^\top(t_q + N - 1)\hat{g}(t_q + N - 1)\| \\ \leq& \|\phi^\top(t_q + N - 1) - \hat{\phi}^\top(t_q + N - 1)\|\|g(t_q + N - 1)\| \\ &+ \|g(t_q + N - 1) - \hat{g}(t_q + N - 1)\|\|\hat{\phi}^\top(t_q + N - 1)\| \\ \leq& c_{12}\delta(t_q + N) + c_{13}\delta(t_q + N - 1) \\ &+ \big(c_{14}\|X(t_q + N)\| + c_{15}\|X(t_q + N - 1)\|\big)T_{q-1} \\ \triangleq& \xi_2,\end{aligned} \tag{49}$$

where the second inequality follows from (42) and (44). Notice that as $N \to \infty$, $\xi_2 \to 0$. (49) also indicates that for all $\epsilon_2 > 0$, there exists a $0 < N_2 < \infty$, when $N > N_2$, such that,

$$\begin{aligned}\|G(t_q + N) - \hat{G}(t_q + N)\| &< \epsilon_2, \\ \|G(t_q + N)\| &< \epsilon_2.\end{aligned} \tag{50}$$

Consider term ③: $\| \sum_{\tau=2}^{N} \Big(\Pi(\tau-1)G(t_q+N-\tau) - \hat{\Pi}(\tau-1)\hat{G}(t_q+N-\tau) \Big) \|$,

$$
\begin{aligned}
&\| \sum_{\tau=2}^{N} \Big(\Pi(\tau-1)G(t_q+N-\tau) - \hat{\Pi}(\tau-1)\hat{G}(t_q+N-\tau) \Big) \| \\
\leq &\sum_{\tau=2}^{N} \|\Pi(\tau-1) - \hat{\Pi}(\tau-1)\| \|G(t_q+N-\tau)\| + \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\| \|\hat{\Pi}(\tau-1)\| \\
= &\sum_{\tau=2}^{N'} \Big(\|\Pi(\tau-1) - \hat{\Pi}(\tau-1)\| \|G(t_q+N-\tau)\| + \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\| \|\hat{\Pi}(\tau-1)\| \Big) \\
&+ \sum_{\tau=N'}^{N} \Big(\|\Pi(\tau-1) - \hat{\Pi}(\tau-1)\| \|G(t_q+N-\tau)\| + \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\| \|\hat{\Pi}(\tau-1)\| \Big),
\end{aligned}
\tag{51}
$$

where $N' = N - N_2 - 1 > N_2$ (assuming N is large enough). Using the result from (46) and (50), we further obtain,

$$
\begin{aligned}
&\| \sum_{\tau=2}^{N} \Big(\Pi(\tau-1)G(t_q+N-\tau) - \hat{\Pi}(\tau-1)\hat{G}(t_q+N-\tau) \Big) \| \\
\leq &\sum_{\tau=2}^{N'} \Big(\|\Pi(\tau-1) - \hat{\Pi}(\tau-1)\| \|G(t_q+N-\tau)\| + \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\| \|\hat{\Pi}(\tau-1)\| \Big) \\
&+ \sum_{\tau=N'}^{N} \Big(\|\Pi(\tau-1) - \hat{\Pi}(\tau-1)\| \|G(t_q+N-\tau)\| + \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\| \|\hat{\Pi}(\tau-1)\| \Big) \\
< &\sum_{\tau=2}^{N'} \Big(\bar{c}_\pi^{\tau-1} \epsilon_2 \Big) + \sum_{\tau=N'}^{N} \Big(b_g \bar{c}_\pi^{\tau-1} \Big) \\
= &\frac{\bar{c}_\pi (1 - \bar{c}_\pi^{N'-1})}{1-\bar{c}_\pi} \epsilon_2 + \frac{\bar{c}_\pi^{N'-1}(1-\bar{c}_\pi^{N-N'+1})}{1-\bar{c}_\pi} b_g \\
= &\frac{\bar{c}_\pi (1 - \bar{c}_\pi^{N-N_2-2})}{1-\bar{c}_\pi} \epsilon_2 + \frac{\bar{c}_\pi^{N}(\bar{c}_\pi^{-2-N_2}-1)}{1-\bar{c}_\pi} b_g \\
\triangleq &\xi_3,
\end{aligned}
\tag{52}
$$

where $b_g = \max_\tau \{ \|G(t_q+N-\tau) - \hat{G}(t_q+N-\tau)\|, \|G(t_q+N-\tau)\| \} < \infty$. Notice that when $N \to \infty$, $\xi_3 \to 0$.

When $N > 2N_2 + 1$, by combing the results from (48), (49) and (52), we obtain the upper bound on $\|\theta_{iq} - \hat{\theta}_{iq}\|$:

$$
\begin{aligned}
\|\theta_{iq} - \hat{\theta}_{iq}\| = &① + ② + ③ \\
< &\xi_1 + \xi_2 + \xi_3 \\
= &c_{11}\bar{c}_\pi^N + \frac{\bar{c}_\pi (1 - \bar{c}_\pi^{N-N_2-2})}{1-\bar{c}_\pi} \epsilon_2 + \frac{\bar{c}_\pi^{N}(\bar{c}_\pi^{-2-N_2}-1)}{1-\bar{c}_\pi} b_g + c_{12}\delta(t_q+N) + c_{13}\delta(t_q+N-1) \\
&+ \big(c_{14}\|X(t_q+N)\| + c_{15}\|X(t_q+N-1)\| + \|\Pi(N)\| \big) T_{q-1} \\
= &\zeta(N) + \psi(N) T_{q-1},
\end{aligned}
\tag{53}
$$

$$
\zeta(N) \triangleq c_{11}\bar{c}_\pi^N + \frac{\bar{c}_\pi (1 - \bar{c}_\pi^{N-N_2-2})}{1-\bar{c}_\pi} \epsilon_2 + \frac{\bar{c}_\pi^{N}(\bar{c}_\pi^{-2-N_2}-1)}{1-\bar{c}_\pi} b_g + c_{12}\delta(t_q+N) + c_{13}\delta(t_q+N-1),
$$

$$
\psi(N) \triangleq c_{14}\|X(t_q+N)\| + c_{15}\|X(t_q+N-1)\| + \|\Pi(N)\|.
$$

Notice that when $N \to \infty$, $\psi(N) \to 0$ and $\zeta(N) \to 0$.

Further we consider,

$$
\begin{aligned}
T_q - T_{q-1} &= \zeta(N) + (\psi(N) - 1) T_{q-1} \\
&= \zeta(N) + (\psi(N) - 1)(\zeta(N) + \psi(N) T_{q-2}) \\
&= \psi(N)(\zeta(N) + (\psi(N) - 1) T_{q-2}) \\
&= \psi(N)(T_{q-1} - T_{q-2}).
\end{aligned}
$$

We observe that when N is large enough, $\psi(N) < 1$, such that

$$|T_q - T_{q-1}| < |T_{q-1} - T_{q-2}| < \cdots < |T_1 - T_0| \,. \tag{54}$$

Notice that when q is large enough, θ_{iq} converges to optimal (Lemma 1), thus, we can now draw the conclusion: for any $\xi > 0$, there exist $N < \infty$ and policy improvement step $q < \infty$ such that,

$$T_q = \|\theta_{iq} - \hat{\theta}_{iq}\| \leq \xi. \tag{55}$$

□