Rate Analysis of Coupled Distributed Stochastic Approximation for Misspecified Optimization ☆

Yaqun Yang[a], Jinlong Lei[b,*]

[a]Department of Control Science and Engineering, Tongji University, Shanghai, 201804, China
[b]Department of Control Science and Engineering, Tongji University, Shanghai, 201804, China; Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai, 200092, China

Abstract

We consider an n agents distributed optimization problem with imperfect information characterized in a parametric sense, where the unknown parameter can be solved by a distinct distributed parameter learning problem. Though each agent only has access to its local parameter learning and computational problem, they mean to collaboratively minimize the average of their local cost functions. To address the special optimization problem, we propose a coupled distributed stochastic approximation algorithm, in which every agent updates the current beliefs of its unknown parameter and decision variable by stochastic approximation method; then averages the beliefs and decision variables of its neighbors over network in consensus protocol. Our interest lies in the convergence analysis of this algorithm. We quantitatively characterize the factors that affect the algorithm performance, and prove that the mean-squared error of the decision variable is bounded by $O(\frac{1}{nk}) + O\left(\frac{1}{\sqrt{n}(1-\rho_w)}\right)\frac{1}{k^{1.5}} + O(\frac{1}{(1-\rho_w)^2})\frac{1}{k^2}$, where k is the iteration count and $(1-\rho_w)$ is the spectral gap of the network weighted adjacency matrix. It reveals that the network connectivity characterized by $(1-\rho_w)$ only influences the high order of convergence rate, while the domain rate still acts the same as the centralized algorithm. In addition, we analyze that the transient iteration needed for reaching its dominant rate $O(\frac{1}{nk})$ is $O(\frac{n}{(1-\rho_w)^2})$. Numerical experiments are carried out to demonstrate the theoretical results by taking different CPUs as agents, which is more applicable to real-world distributed scenarios.

Keywords: Distributed Coupled Optimization, Stochastic Approximation, Misspecification, Convergence Rate Analysis

1. Introduction

In recent years, distributed optimization has drawn much research attention in various fields including economic dispatch[1, 2], smart grids [3, 4, 5], automatic controls[6, 7, 8] and machine learning [9, 10]. In distributed scenarios, each agent only preserves its local information, while they can exchange information with others over a connected network to cooperatively minimize the average of all agents' cost functions [11, 12]. There are several approaches for resolving distributed optimization problems such as (primary) consensus-based, duality-based, and constraint exchange methods, where the primal approaches characterized by gradient-based algorithms have attracted the most research attention due to their satisfactory performance and well-scalable nature[13]. The distributed dual approaches based on Lagrange method regularly use dual decomposition like the alternating direction method of multipliers (ADMM)[14]. Constraint exchange method is another prevalent scheme where the information exchanged by agents amounts to constraints[15].

☆The paper was sponsored by the National Key Research and Development Program of China under No 2022YFA1004701, the National Natural Science Foundation of China under No. 72271187 and No. 62373283, and partially by Shanghai Municipal Science and Technology Major Project No. 2021SHZDZX0100, and National Natural Science Foundation of China (Grant No. 62088101).

*Corresponding author
Email address: yangyaqun@tongji.edu.cn, leijinlong@tongji.edu.cn (Jinlong Lei)

However, among various formulations in distributed optimization, a crucial assumption is that we need precise objective functions (or problem information), i.e., all parameters in the model are precisely known. Yet in many economic and engineering problems, parameters of the functions are unknown but we may have access to observations that can aid in resolving this misspecification. For example, in the Markowitz profile problem, it is routinely assumed that the expectation or covariance matrices associated with a collection of stocks are accurately available, but in reality, it needs empirical estimates via past data[16].

This paper is devoted to proposing distributed algorithms for resolving optimization problems with parametric misspecification, and quantitatively characterizing the influence of network properties, the heterogeneity of agents, initial states, etc. on the algorithm performance. This work is primarily centered around conducting a comprehensive theoretical analysis of convergence. We begin by initiating the problem formulation.

1.1. Problem Formulation

In this article, we consider a static misspecified distributed optimization problem defined as follows:

$$C_x(\theta_*): \qquad \min_{x\in\mathbb{R}^p} f(x,\theta_*) = \frac{1}{n}\sum\nolimits_{i=1}^{n} f_i(x,\theta_*), \tag{1}$$

where $f_i(x,\theta_*) \triangleq \mathbb{E}[\tilde{f}_i(x,\theta,\xi_i)]$ is the local cost function only accessible for agent $i \in \mathcal{N} \triangleq \{1,2,...,n\}$. Suppose that for any $i \in \mathcal{N}$, $\xi_i : \Omega_x \to \mathbb{R}^d$ are mutually independent random variables defined on a probability space $(\Omega_x, \mathcal{F}_x, \mathbb{P}_x)$. Here, $\theta_* \in \mathbb{R}^q$ denotes the unknown parameter, which is a solution to a distinct convex problem.

$$\mathcal{L}_\theta: \qquad \min_{\theta\in\mathbb{R}^q} h(\theta) = \frac{1}{n}\sum\nolimits_{i=1}^{n} h_i(\theta), \tag{2}$$

where $h_i(\theta) \triangleq \mathbb{E}[\tilde{h}_i(\theta,\zeta_i)]$ is the local parameter learning function only accessible for agent $i \in \mathcal{N}$, and for any $i \in \mathcal{N}$, $\zeta_i : \Omega_\theta \to \mathbb{R}^m$ are mutually independent random variables defined on a probability space $(\Omega_\theta, \mathcal{F}_\theta, \mathbb{P}_\theta)$. Problems in the form eq. (1) and eq. (2) jointly formulate an unknown coupled distributed optimization scheme consisting of both *computational problem* and *learning problem*, where the learning problem is independent of the computational one. We have depicted the problem setting in fig. 1.

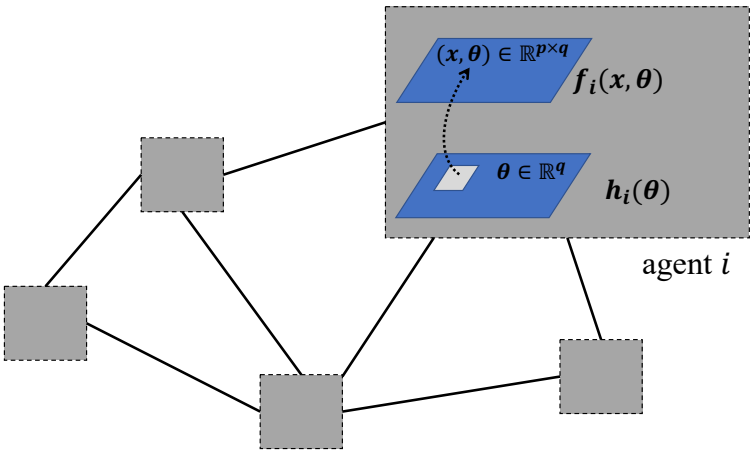


Figure 1: The problem setup: a connected network of communicating agents, where each agent preserving a local learning problem h_i and computational problem f_i correlated with h_i through the unknown parameter θ, while they cooperate to solve the distributed coupled optimization problem.

1.2. Prior Work

We now give a review of prior work for resolving optimization problems with unknown parameters.

Robust optimization approach. Robust optimization considers the optimization problem where the parameter θ is unavailable but one can have access to its uncertainty set, say $\mathcal{U}_\theta$[17]. The key idea is to optimize against the worst-case realization within this set, i.e.,

$$\min_{x\in\mathbb{R}^p} \max_{\theta\in\mathcal{U}_\theta} f(x,\theta).$$

Robust optimization is shown to be a useful technique in the resolution of problems arising from control, design, and optimization [18]. However, it usually produces conservative solutions and sometimes is intractable when poor set $\mathcal{U}_\theta$ is chosen (e.g. the set is given by unexplicit systems of non-convex inequalities)[19].

Stochastic optimization. Unlike robust optimization, in a stochastic optimization scenario one may obtain statistical or distributional information about the unknown parameter. For example, θ follows a probability distribution $\mathcal{D}$[20], the optimal solution is gained by minimizing the expectation of cost functions,

$$\min_{x\in\mathbb{R}^p} \mathbb{E}_{\theta\sim\mathcal{D}}[f(x,\theta)].$$

Stochastic optimization has been widely investigated in telecommunication, finance and machine learning [21]. In the scenario of a multi-agent network dealing with large datasets, stochastic optimization has become popular since it is challenging to calculate the exact gradient while the stochastic gradient is much easier to obtain. A key shortcoming in using stochastic optimization models for resolving optimization problems with unknown parameters lies in that it needs the distribution of θ, which might be a stringent requirement when the available data for estimating is limited or noisy. In such cases, the resulting distribution estimates may be unreliable or biased, leading to suboptimal solutions or even infeasible solutions[22]. Alternatively, suppose that θ_* can be learnt by a suitably defined estimation problem, then it brings about the following approach.

Data-driven learning approach. As data availability reaches hitherto unseen in recent years, we can use data-driven approaches to lessen or even eliminate the impact of model uncertainty. For example, the model parameter θ can be obtained by solving a suitably defined learning problem $l(\theta)$ (see e.g., [23]),

$$\min_{x\in\mathbb{R}^p} \left\{ f(x,\theta_*) : \theta_* \in \arg\min_{\theta\in\mathbb{R}^q} l(\theta) \right\}. \tag{3}$$

Computational evidence in portfolio management and queueing confirm that data-driven sets significantly outperform traditional robust optimization techniques[19].

A natural question is whether this problem could be solved in a sequential method, i.e., first accomplish estimating θ_* with high accuracy and then solve the core computational optimization problem with the achieved estimation $\hat{\theta}$. However, they have some disadvantages discussed in [23, 24, 25]: on the one hand, the large-scale parameter learning problem will lead to long time waiting for solving the original problem. On the other hand, this scheme provides an approximate solution $\hat{\theta}$, then the corrupt error might propagates into the computational problem. As such, sequential methods cannot provide asymptotically accurate convergence. Therefore, an alternative simultaneous approach is designed (see e.g., [23, 26]), which use observations to get an estimation θ_k of unknown parameters θ^* at each time instant k; then update the upper optimization problem by taking the estimated parameter θ_k as "true" parameter. This simultaneous approach can generate a sequence $\{(x_k,\theta_k)\}$ that converges to a minimizer of $f(x,\theta_*)$ and $l(\theta)$ respectively [23].

Such data-driven learning approaches for unknown parameter has gradually attracted research attention recently. For example, the authors of [23] presented a centralized coupled stochastic optimization scheme to solve problem (3) and showed the convergence properties in regimes when the function is either strongly convex or merely convex. Then [25] extended it smooth or nonsmooth functions f and presented an averaging-based subgradient approach, but it is still a centralized scheme. In addition, the authors of [24] divided the optimization problem with uncertainty into two paradigms: robust optimization and joint estimation optimization, and they exploited these two problem structures in online convex optimization and gave regret analysis under different conditions. The recent work [16] investigated the misspecified conic convex programs, and developed a centralized first-order inexact augmented Lagrangian scheme for computing the optimal solution while simultaneously learning the unknown parameters. The aforementioned work [16, 24, 23, 25] all investigated centralized methods, while there are some other work exploit distributed approaches. For example, [27] considered the distributed stochastic optimization with imperfect information, while it only showed

that the generated iterates converge almost surely to the optimal solution. Though the work [28] presented a distributed problem with a composite structure consisting of an exact engineering part and an unknown personalized part, it exhibits a bounded regret under certain conditions.

1.3. Gaps and Motivation

Recalling the problem setup in section 1.1, our research falls into distributed data-driven stochastic optimization scenario. Taking into account the research that is most pertinent to this paper, the majority of previous studies have primarily concentrated on centralized inquiries (see e.g. [16, 24, 23, 25]), while the distributed schemes [27, 28] mainly investigated the asymptotic convergence. It remains unknown how to design an efficient distributed algorithm, how does the network connectivity influence the algorithm performance, and whether the rate can reach the same order as the centralized scheme? To be specific, this paper is motivated by the following gaps: (i) previous work on unknown parameter learning problems focused on the centralized scheme, the distributed data-driven stochastic approximation method is less studied; (ii) the discussion of convergence analysis especially how factors such as the number of agents, the network connectivity, and the heterogeneity of agents influence the rate of algorithm is rarely studied in details; (iii) the gap between centralized and distributed algorithm under imperfect information need to be specified, or in other words can we find the transient time when the rate of distributed algorithms reach the same order as that of the centralized scheme?

1.4. Outline and Contributions

To address these gaps, we propose a data-driven coupled distributed stochastic approximation method to resolve this special optimization problem and give a precise convergence rate analysis of our algorithm. The main contributions are summarized as follows, and the comparison with previous works is shown in table 1.

Table 1: Work comparation with previous studies

Paper	Distributed	Imperctect Information	Stochastic	Rate	Factor Influence
[24, 23, 25]	✗	✓	✗	$O\left(\frac{1}{k}\right)$	✗
[29, 30]	✓	✗	✓	$O\left(\frac{1}{k}\right)$	✓
[28]	✓	✓	✗	\	✓
[27]	✓	✓	✓	\	✗
[31]	✓	✗	✗	$O\left(\frac{1}{\sqrt{k}}\right)$	✗
Our Work	✓	✓	✓	$O\left(\frac{1}{k}\right)$	✓

(1) We propose a coupled distributed stochastic approximation algorithm that generates iterates $\{(\mathbf{x}(k), \boldsymbol{\theta}(k))\}$ for the distributed stochastic optimization problem (1) with the unknown parameter learning prescribed by a separate distributed stochastic optimization problem (2). Our model framework builds upon previous research involving deterministic and stochastic gradient schemes. This is particularly relevant for certain studies where waiting for parameter learning to complete over an extended period is not feasible, or for real-world problems in which parameter learning and objective optimization are intertwined.

(2) We characterize the convergence rate of the presented algorithm that combined the distributed consensus protocol with stochastic gradient descent methods. On the one hand, we prove that the upper bound of expected consensus error for every agent decay at rate $O(\frac{1}{k^2})$; on the other hand, we also show that the upper bounded of expected optimization error is $O(\frac{1}{k})$. We then give the sublinear convergence rate and quantitatively characterize some factors affecting the convergence rate, such as the network size, spectral gap of the weighted adjacency matrix, heterogenous of individual function, and initial values. We emphasize that the mean-squared error of the decision variable is bounded by $O(\frac{1}{nk}) + O\left(\frac{1}{\sqrt{n}(1-\rho_w)}\right)\frac{1}{k^{1.5}} + O(\frac{1}{(1-\rho_w)^2})\frac{1}{k^2}$, which indicates that the network connectivity characterized by $(1-\rho_w)$ only influences the high order of convergence rate, while the domain rate $O(\frac{1}{k})$ still acts the same as the centralized algorithm.

(3) We analyze the transient time K_T for the proposed algorithm, namely, the number of iterations before the algorithm reaches its dominant rate. Specially, we show that when the iterate $k \geq K_T$, the dominant factor influencing

the convergence rate is related to stochastic gradient descent, while for small $k < K_T$, the main factor influencing the convergence rate originates from the distributed average consensus method. Finally, we show that the algorithm asymptotically achieves the same network-independent convergence rate as the centralized scheme.

The paper is organized as follows. We present the algorithm and the related assumptions in section 2. In section 3, the auxiliary results supporting the convergence rate analysis is proved. Our main results are in section 4. Experimental results are implemented in section 5, while the concluding remarks are given in section 6.

Notation. All vectors in this paper are column vectors. The structure of the communication network is modeled by an undirected weighted graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$ in which $\mathcal{N} = \{1, 2, ..., n\}$ represents the set of vertices. $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. $W = [w_{ij}]_{n\times n} \in \mathbb{R}^{n\times n}$ denotes the weighted adjacency matrix, $w_{ij} > 0$ if and only if agent i and agent j are connected, $w_{ij} = w_{ji} = 0$ otherwise. Each agent(vertice) has a set of neighbors $\mathcal{N}_i = \{j|(i, j) \in \mathcal{E}\}$. The graph is connected means for every pair of nodes (i, j) there exists a path of edges that goes from i to j. $\|\cdot\|$ denotes $\mathcal{L}_2$-norm for vectors and Euclidean norm for matrices. The optimal solution denote as (x_*, θ_*).

2. Algorithm and Assumptions

To solve this special optimization problem consisting of the *computational problem* eq. (1) and the *learning problem* eq. (2), we will propose a *Coupled Distributed Stochastic Approximation (CDSA) Algorithm* and impose some conditions for rate analysis in this section.

2.1. Algorithm Set Up

As mentioned previously, each agent i only knows its local core computational function $f_i(x, \theta)$ and parameter learning function $h_i(\theta)$, while they are connected by a network $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{W})$ in which agents may communicate and exchange information with their neighbors $\mathcal{N}_i = \{j|(i, j) \in \mathcal{E}\}$. At each step $k \geq 0$, every agent i holds an estimate of the decision variable and unknown parameter, denoted by $x_i(k)$ and $\theta_i(k)$, respectively. Suppose that every agent has access to a stochastic first-order oracle that can generate stochastic gradients $g_i(x_i(k), \theta_i(k), \xi_i(k)) \triangleq \nabla_x f_i(x_i(k), \theta_i(k), \xi_i(k))$ and $\phi_i(\theta_i(k), \zeta_i(k)) \triangleq \nabla_\theta h_i(\theta_i(k), \zeta_i(k))$ respectively (where $\xi_i, \zeta_i, i = 1, 2, ..., n$ are independent random variables). Then, every agent updates its parameters through stochastic gradient descent method to obtain temporary estimates $\widetilde{x}_i(k)$ and $\widetilde{\theta}_i(k)$. Next, each agent communicates with its local neighbors and gathers temporary parameters information over a static connected network to renew the iterates $x_i(k + 1)$ and $\theta_i(k + 1)$ based on the consensus protocol. We summarize the pseudo-code is in algorithm 1.

Algorithm 1 Coupled Distributed Stochastic Approximation (CDSA)

Initialization: $W = [w_{ij}]_{n\times n}$; $(x_i(0), \theta_i(0)), \forall i \in \mathcal{N}$

Evolution: for $k = 0, 1, 2, ...; \forall i \in \mathcal{N}$

Compute: stochastic gradient $\phi_i(\theta_i(k), \zeta_i(k))$ and $g_i(x_i(l), \theta_i(k), \xi_i(k))$

Choose: stepsize α_k and γ_k (To be introduced in section 3.3)

Update according to the following stochastic gradient descent method.

$$\widetilde{x}_i(k) = x_i(k) - \alpha_k g_i(x_i(k), \theta_i(k), \xi_i(k))$$
$$\widetilde{\theta}_i(k) = \theta_i(k) - \gamma_k \phi_i(\theta_i(k), \zeta_i(k))$$

Gather information $\widetilde{x}_j(k), \widetilde{\theta}_j(k)$ from its neighbors $j \in \mathcal{N}_i$ and renew the iterates by the consensus protocol below.

$$x_i(k + 1) = \sum\nolimits_{j\in\mathcal{N}_i} w_{ij}\widetilde{x}_j(k)$$
$$\theta_i(k + 1) = \sum\nolimits_{j\in\mathcal{N}_i} w_{ij}\widetilde{\theta}_j(k)$$

We can rewrite Algorithm 1 in a more compact form as follows.

$$x_i(k+1) = \sum\nolimits_{j\in\mathcal{N}_i} w_{ij}(x_j(k) - \alpha_k g_j(x_i(l), \theta_i(k), \xi_i(k))), \tag{4}$$
$$\theta_i(k+1) = \sum\nolimits_{j\in\mathcal{N}_i} w_{ij}\left(\theta_j(k) - \gamma_k \phi_j\left(\theta_i(k), \zeta_i(k)\right)\right). \tag{5}$$

Define

$$\mathbf{x} \triangleq [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^{n\times p}, \boldsymbol{\theta} \triangleq [\theta_1, \theta_2, \cdots, \theta_n]^T \in \mathbb{R}^{n\times q}, \tag{6}$$
$$\boldsymbol{\xi} \triangleq [\xi_1, \xi_2, \cdots, \xi_n]^T \in \mathbb{R}^n,\ \boldsymbol{\zeta} \triangleq [\zeta_1, \zeta_2, \cdots, \zeta_n]^T \in \mathbb{R}^n, \tag{7}$$
$$\mathbf{g}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\xi}) \triangleq [g_1(x_1, \theta_1, \xi_1), g_2(x_2, \theta_2, \xi_2), \cdots, g_n(x_n, \theta_n, \xi_n)]^T \in \mathbb{R}^{n\times p}, \tag{8}$$
$$\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{\zeta}) \triangleq [\phi_i(\theta_1, \zeta_1), \phi_i(\theta_2, \zeta_2), \cdots, \phi_i(\theta_n, \zeta_n)]^T \in \mathbb{R}^{n\times q}. \tag{9}$$

Then equation (4) and (5) can be reformulated in the following vector formula.

$$\mathbf{x}(k+1) = W(\mathbf{x}(k) - \alpha_k \mathbf{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k))), \tag{10}$$
$$\boldsymbol{\theta}(k+1) = W(\boldsymbol{\theta}(k) - \alpha_k \boldsymbol{\phi}(\boldsymbol{\theta}(k), \boldsymbol{\zeta}(k))). \tag{11}$$

2.2. Assumptions

In this subsection, we will specify the conditions for rate analysis of the CDSA algorithm. We need to make some assumptions about the properties of objective functions in both learning and computation metrics to get the global optimal solution. Besides, we impose some constraints on conditional first and second moments of "stochastic gradient". Last but not least, we inherit the typical assumptions about communication networks as that of distributed algorithms.

Assumption 2.2.1 (Function properties) *(i) For every $\theta \in \mathbb{R}^q$, $f_i(x, \theta), i = 1, \cdots, n$ is strongly convex and Lipschitz smooth in x with constants μ_x and L_x, i.e.*

$$(\nabla_x f_i(x', \theta) - \nabla_x f_i(x, \theta))^T (x' - x) \geq \mu_x \|x' - x\|^2, \forall x, x' \in \mathbb{R}^p,$$
$$\|\nabla_x f_i(x', \theta) - \nabla_x f_i(x, \theta)\| \leq L_x \|x' - x\|, \forall x, x' \in\in \mathbb{R}^p.$$

(ii) For every $x \in \mathbb{R}^p$, $f_i(x, \theta), i = 1, \cdots, n$ is strongly convex and Lipschitz smooth in θ with constants μ_θ and L_θ respectively, i.e.

$$(\nabla_x f_i(x, \theta') - \nabla_x f_i(x, \theta))^T (\theta' - \theta) \geq \mu_\theta \|\theta' - \theta\|^2, \forall \theta, \theta' \in \mathbb{R}^q,$$
$$\|\nabla_x f_i(x, \theta') - \nabla_\theta f(x, \theta)\| \leq L_\theta \|\theta' - \theta\|, \forall \theta, \theta' \in \mathbb{R}^q.$$

(iii)The learning metric $h_i(\theta)$ for every $i \in \{1, 2, ..., n\}$ is strongly convex and Lipschitz smooth with constants ν_θ and C_θ, i.e.

$$(h(\theta) - h(\theta'))^T (\theta - \theta') \geq \nu_\theta \|\theta - \theta'\|^2, \forall \theta, \theta' \in \mathbb{R}^q,$$
$$\|\nabla h(\theta) - \nabla h(\theta')\| \leq C_\theta \|\theta - \theta'\|, \forall \theta, \theta' \in \mathbb{R}^q.$$

Strong convexity assumptions indicate that both computational problem and learning problem have a unique optimal solution $x_* \in \mathbb{R}^p$ and $\theta_* \in \mathbb{R}^q$ [32]. The Lipschitz continuity of gradient functions ensure that the gradient doesn't change arbitrarily fast concerning the corresponding parameter vector. It is widely used in the convergence analyses of most gradient-based methods, without it, the gradient wouldn't provide a good indicator for how far to move to decrease the objective function[32]. These assumptions are satisfied for many machine learning problems, such as logistic regression, linear regression, and support vector machine (SVM).

Next, we define a new probability space $(\mathcal{Z}, \mathcal{F}, \mathbb{P})$, where $\mathcal{Z} \triangleq \Omega_x \times \Omega_\theta, \mathcal{F} \triangleq \mathcal{F}_x \times \mathcal{F}_\theta$ and $\mathbb{P} \triangleq \mathbb{P}_x \times \mathbb{P}_\theta$. We use $\mathcal{F}(k)$ to denote the σ-algebra generated by $\{(x_i(0), \theta_i(0)), (x_i(1), \theta_i(1)), ..., (x_i(k), \theta_i(k)) | i \in \mathcal{N}\}$. Then give the following assumptions related to the stochastic gradient estimator, which assume that the stochastic gradient is an unbiased estimator of the true gradient, and the variance of the stochastic gradient is restricted.

Assumption 2.2.2 (Conditional first and second moments) *For all $k \geq 0$ and $i \in \mathcal{N}$,there exist $\sigma_x > 0, \sigma_\theta > 0, M_x > 0, M_\theta > 0$, such that*

$$
\begin{aligned}
&(a) \quad \mathbb{E}_{\xi_i(k)}[g_i(x_i(k), \theta_i(k), \xi_i(k))|\mathcal{F}(k)] = \nabla_x f_i(x_i(k), \theta_i(k)), \quad a.s.,\\
&(b) \quad \mathbb{E}_{\zeta_i(k)}[\phi_i(\theta_i(k), \zeta_i(k))|\mathcal{F}(k)] = \nabla h_i(\theta_i(k)), \quad a.s.,\\
&(c) \quad \mathbb{E}_{\xi_i(k)}[\|g_i(x_i(k), \theta_i(k), \xi_i(k)) - \nabla_x f_i(x_i(k), \theta_i(k))\|^2|\mathcal{F}(k)],\\
&\qquad\qquad \leq \sigma_x^2 + M_x\|\nabla_x f_i(x_i(k), \theta_i(k))\|^2 \quad a.s.,\\
&(d) \quad \mathbb{E}_{\zeta_i(k)}[\|\phi_i(\theta_i(k), \zeta_i(k)) - \nabla h_i(\theta_i(k))\|^2|\mathcal{F}(k)] \leq \sigma_\theta^2 + M_\theta\|\nabla h_i(\theta_i(k))\|^2, \quad a.s.,
\end{aligned}
$$

Next, we impose the connectivity condition on the graph, which indicates that after multiple rounds of communication, information can be exchanged between any two agents. This inherits the typical assumptions on consensus protocols [33].

Assumption 2.2.3 (Graph and weighted matrix) *The graph $\mathcal{G}$ is static, undirected, and connected. The weighted adjacency matrix W is nonnegative and doubly stochastic, i.e.,*

$$W\mathbf{1} = \mathbf{1}, \mathbf{1}^T W = \mathbf{1}^T \tag{12}$$

where $\mathbf{1}$ is the vector of all ones.

Next, we state two lemmas that partially explain the practicability of algorithm 1 based on the aforementioned assumptions.

Lemma 2.2.1 *[34, Lemma 10] For any $x \in \mathbb{R}^p$, define $x^+ = x - \alpha\nabla f(x)$. Suppose that f is strongly convex with constant μ and its gradient function is Lipschitz continuous with constant L. If $\alpha \in (0, 2/L)$, we then have $\|x^+ - x_*\| \leq \lambda\|x - x_*\|$, where $\lambda \triangleq \max(|1 - \alpha\mu|, |1 - \alpha L|)$.*

It can be observed from the above lemma that as long as we choose a proper stepsize ($0 < \alpha < 2/L$), the distance to optimizer shrinks by a ratio $\lambda < 1$ at each step for strongly convex and smooth functions. While the following lemma reveals that under distributed algorithm with linear iteration, the gap between the current iteration and consensus optimal solution is decreased by a ratio $\rho_w < 1$ compared to the last iteration.

Lemma 2.2.2 *[33, Theorem 1] Let Assumption 2.2.3 hold, and ρ_w denote the spectural norm of matrix $W - \frac{\mathbf{1}\mathbf{1}^T}{n}$. Thus $\rho_w < 1$. Define $\boldsymbol{\omega}^+ = W\boldsymbol{\omega}$ for any $\boldsymbol{\omega} \in \mathbb{R}^{n\times p}$. We then have $\|\boldsymbol{\omega}^+ - \mathbf{1}\bar{\omega}\| \leq \rho_w\|\boldsymbol{\omega} - \mathbf{1}\bar{\omega}\|$, where $\bar{\omega} \triangleq \frac{1}{n}\mathbf{1}^T\boldsymbol{\omega}$.*

The aforementioned lemmas show that both the gradient descent method and distributed linear iteration can move the decision variable towards the optimal solution with linear decaying rates. Thus, our algorithm consisting of both approaches might find the optimal solution efficiently. We will rigorously prove the convergence rate of algorithm 1 in the following two sections.

3. Auxiliary Results

In this section, we will present some results to assist subsequent convergence rate analysis. We first give some preliminary bound which will be used for later proof, then present the supporting lemmas concerning recursions for expected optimization error and expected consensus error, and finally, we prove that under diminishing stepsize, the mean-squared distance between the current iterate and the optimal solution is uniformly bounded.

3.1. Preliminary Bound

For simplicity, we denote

$$\bar{x}(k) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i(k), \ \bar{\theta}(k) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} \theta_i(k), \tag{13}$$

$$\bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k)) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} g_i(x_i(k), \theta_i(k), \xi_i(k)), \tag{14}$$

$$\bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) \triangleq \frac{1}{n}\sum\nolimits_{i=1}^{n} \nabla_x f_i\left(x_i(k), \theta_i(k)\right). \tag{15}$$

We will show in the following lemma that with Assumptions 2.2.1 and 2.2.2, the conditional squared distance between the gradient $\bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))$ and its estimate can be bounded by linear combinations of squared errors $\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2$ and $\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2$. For completeness, its proof is given in appendix Appendix A.

Lemma 3.1.1 *Let Assumption 2.2.1 and 2.2.2 hold. Then for any $k \geq 0$,*

$$\begin{aligned}\mathbb{E}[\|\bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\|^2 | \mathcal{F}(k)] \\ \leq \frac{3M_x L_x^2}{n^2}\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2 + \frac{3M_x L_\theta^2}{n^2}\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2 + \frac{\bar{M}}{n}, \quad (16)\end{aligned}$$

$$\textit{where} \quad \bar{M} = \frac{3M_x \sum_{i=1}^n \|\nabla_x f_i(x_*, \theta_*)\|^2}{n} + \sigma_x^2. \quad (17)$$

The following lemma shows the gap between the gradient of objective function at the consensual points $(\bar{x}(k), \bar{\theta}(k))$, denoted by $\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(\bar{x}(k), \bar{\theta}(k))$, and at current iterates $\frac{1}{n}\sum_{i=1}^n \nabla_x f_i(x_i(k), \theta_i(k))$ can also be bounded by linear combinations of $\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2$ and $\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2$. The precise proof is in appendix Appendix B.

Lemma 3.1.2 *Let Assumption 2.2.1 hold. Then for any $k \geq 0$,*

$$\|\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\| \leq \frac{L_x}{\sqrt{n}}\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\| + \frac{L_\theta}{\sqrt{n}}\|\boldsymbol{\theta}(k) - \mathbf{1}\bar{\theta}(k)^T\|. \quad (18)$$

The above two lemmas, providing the related upper bounds of functions, are derived by virtue of the Lipschitz smooth assumption. They are essential for the subsequent convergence analysis.

3.2. Supporting Lemmas

In this subsection, we present some results concerning *expected optimization error* $\mathbb{E}[\|\bar{x}(k) - x_*\|^2]$ and *expected consensus error* $\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|^2]$ for core computational problem, while the discussion of parameter learning problem can be found in [30]. For ease of presentation, we denote

$$U_1(k) \triangleq \mathbb{E}[\|\bar{x}(k) - x_*\|^2], V_1(k) \triangleq \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|^2], \quad (19)$$

$$U_2(k) \triangleq \mathbb{E}[\|\bar{\theta}(k) - \theta_*\|^2], V_2(k) \triangleq \mathbb{E}[\|\boldsymbol{\theta}(k) - \mathbf{1}\bar{\theta}(k)^T\|^2]. \quad (20)$$

Next we will bound $U_1(k+1)$ and $V_1(k+1)$ by error terms at iteration k. The precise proof of Lemma 3.2.1 is in appendix Appendix C.

Lemma 3.2.1 *Let Assumption 2.2.1~2.2.3 hold, under algorithm 1,*
*(**A**) Supposing stepsize $\alpha_k \leq \frac{1}{L_x}$, we have*

$$\begin{aligned}U_1(k+1) \leq (1 - \alpha_k\mu_x)^2 U_1(k) + \frac{\alpha_k^2 L_x^2}{n} V_1(k) + \frac{\alpha_k^2 L_\theta^2}{n} V_2(k) \\ + \frac{2L_x L_\theta \alpha_k^2}{n}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|\|\boldsymbol{\theta}(k) - \mathbf{1}\bar{\theta}(k)^T\|] \\ + \frac{2\alpha_k L_x}{\sqrt{n}}(1 - \alpha_k\mu_x)\mathbb{E}[\|\bar{x}(k) - x_*\|\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|] \\ + \frac{2\alpha_k L_\theta}{\sqrt{n}}(1 - \alpha_k\mu_x)\mathbb{E}[\|\bar{x}(k) - x_*\|\|\boldsymbol{\theta}(k) - \mathbf{1}\bar{\theta}(k)^T\|] \\ + \alpha_k^2\left(\frac{3M_x L_x^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|] + \frac{3M_x L_\theta^2}{n^2}\mathbb{E}[\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|] + \frac{\bar{M}}{n}\right), \quad (21)\end{aligned}$$

*(**B**) Supposing stepsize $\alpha_k \leq \min\{\frac{1}{L_x}, \frac{1}{3\mu_x}\}$, we have*

$$U_1(k+1) \leq (1 - \frac{3}{2}\alpha_k\mu_x)U_1(k) + \frac{6\alpha_k L_x^2}{n\mu_x}V_1(k) + \frac{6\alpha_k L_\theta^2}{n\mu_x}V_2(k) + \alpha_k^2\left(\frac{3M_x L_x^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] + \frac{3M_x L_\theta^2}{n^2}\mathbb{E}[\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2] + \frac{\bar{M}}{n}\right). \tag{22}$$

Result **B** restricts the stepsize to smaller one than that of Result **A**. It thus simplifies Result **A** eq. (21) by removing the cross term to facilitate the later analysis. We can revisit Inequality eq. (22) and reformulate it as $U_1(k+1) \leq (1 - \frac{3}{2}\alpha_k\mu_x)U_1(k) + error(\alpha_k)$, where $error(\alpha)$ means an error function that is proportional to α. We should mention that, since $\alpha_k > 0$ and $\mu_x > 0$, expected optimization error $U_1(k)$ roughly shrinks by a ratio $(1 - \frac{3}{2}\alpha_k\mu_x) < 1$. Though there is an error term related to α_k, when we choose diminishing stepsize policy and the consensus errors $V_1(k), V_2(k)$ as well as $\mathbb{E}(\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2)$, $\mathbb{E}(\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2)$ are bounded, the error may decrease to 0, which indicates the convergence of $U_1(k)$.

We define

$$\nabla_x \boldsymbol{F}(\mathbf{x}, \boldsymbol{\theta}) \triangleq [\nabla_x f_1(x_1, \theta_1), \nabla_x f_2(x_2, \theta_2), \cdots, \nabla_x f_n(x_n, \theta_n)]^T \in \mathbb{R}^{n \times p}. \tag{23}$$

In the next lemma, we will show the recursive formulation of expected consensus error $V_1(k)$, which is critical for convergence analysis. For completeness, we give its proof in appendix Appendix D.

Lemma 3.2.2 *Let Assumption 2.2.1~2.2.3 hold, and consider algorithm 1. Then for any $k \geq 0$, we have*

$$V_1(k+1) \leq \frac{3+\rho_w^2}{4}V_1(k) + \alpha_k^2\rho_w^2 n\sigma_x^2 + 3\alpha_k^2\rho_w^2\left(\frac{3}{1-\rho_w^2} + M_x\right)\Big(L_x^2\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] + L_\theta^2\mathbb{E}[\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2] + \|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T, \mathbf{1}\theta_*^T)\|^2\Big). \tag{24}$$

The recursion of expected consensus error can be reformulate as $V_1(k+1) \leq \frac{3+\rho_w^2}{4}V_1(k) + error(\alpha_k^2\rho_w^2)$. It is worth mentioning that $V_1(k)$ can roughly shrink by $\frac{3+\rho_w^2}{4} < 1$ since $\rho_w < 1$. Note that the extra error term in the consensus error is proportional to α_k^2, compared to $U_1(k)$ with an error term proportional to α_k. We might obtain a qualitative conclusion that expected consensus error decrease faster than expected optimization error. We will present the precise proof in the next part that consensus error decrease to 0 at an order $O(\frac{1}{k^2})$ while optimization error at $O(\frac{1}{k})$.

Remark 1 *Recalling recursion of $U_1(k)$ in (22) and recursion of $V_1(k)$ in (24), we could notice that the expected consensus error is more related to the network connectivity ρ_w, which is natural because "consensus" is induced from the distributed algorithm, while "optimization" mainly comes from original optimization method such as stochastic gradient descent.*

3.3. Uniform Bound

From now on, we consider stepsize policy as follows

$$\alpha_k \triangleq \frac{\beta}{\mu_x(k+K)}, \gamma_k \triangleq \frac{\beta}{\mu_\theta(k+K)}, \quad \forall k, \tag{25}$$

where the β is a positive constant , and

$$K \triangleq \left\lceil \max\left\{\frac{3\beta(1+M_x)L_x^2}{\mu_x^2}, \frac{3\beta(1+M_\theta)L_\theta^2}{\mu_\theta^2}\right\}\right\rceil \tag{26}$$

with $\lceil\cdot\rceil$ denoting the ceiling function.

Next, We present a lemma that derives a uniform bound on the iterates $\{\boldsymbol{\theta}(k)\}, \{\boldsymbol{x}(k)\}$ generated by algorithm 1. Such a result is helpful for bounding the expected optimization error and expected consensus error.

Lemma 3.3.1 *Let Assumption 2.2.1~2.2.3 hold. Consider algorithm 1 with stepsize policy (25). We then obtain from [30, Lemma 8] that for all $k \geq 0$,*

$$\mathbb{E}[\|\theta_i(k) - \theta_*\|^2] \leq \hat{\Theta}_i \triangleq \max\left\{\|\theta_i(0) - \theta_*\|^2, \frac{9\|\nabla h_i(\theta_*)\|^2}{\mu_\theta} + \frac{\sigma_\theta^2}{(1+M_\theta)L_\theta^2}\right\}. \tag{27}$$

Based on (27)*, we can obtain the following result with $\hat{\Theta} \triangleq \sum_{i=1}^n \hat{\Theta}_i$,*

$$\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] \leq \hat{X}, \text{ where} \tag{28}$$

$$\hat{X} \triangleq \max\left\{\|\mathbf{x}(0) - \mathbf{1}x_*^T\|^2, \frac{11L_\theta^2\hat{\Theta}}{\mu_x^2} + \frac{11\|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T, \mathbf{1}\theta_*^T)\|^2}{\mu_x^2} + \frac{7n\sigma_x^2}{9(1+M_x)L_x^2}\right\}. \tag{29}$$

We will give the proof of (28) in appendix Appendix E. Lemma 3.3.1 indicates that although the problem we consider is unconstrained, the gap between the iterates generated by algorithm CDSA and the optimal solution is uniformly bounded. It is critical for the analysis of sublinear convergence rates of $U_1(k)$ and $V_1(k)$. Then based on this lemma, we will provide uniform upper bounds for the expected optimization error and expected consensus error.

Lemma 3.3.2 *Let Assumption 2.2.1~2.2.3 hold. Consider algorithm 1 with stepsize policy (25). We then have $U_1(k) \leq \frac{\hat{X}}{n}, V_1(k) \leq 4\hat{X}$.*

Proof By recalling (28) and using Cauchy-Schiwaz inequality, we obtain that

$$\begin{aligned}
U_1(k) &= \mathbb{E}[\|\bar{x}(k) - x_*\|^2] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n x_i(k) - \frac{1}{n}\sum_{i=1}^n x_*\right\|^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n (x_i(k) - x_*)\right\|^2\right] \leq \frac{1}{n^2} \times n\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] \leq \frac{\hat{X}}{n}, \\
V_1(k) &= \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|^2] = \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T + \mathbf{1}x_*^T - \mathbf{1}\bar{x}(k)^T\|^2] \\
&\leq 2\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] + 2\mathbb{E}[\|\mathbf{1}(x_* - \bar{x}(k)\|^2)] \\
&\leq 2\hat{X} + 2n \times \frac{\hat{X}}{n} \leq 4\hat{X}.
\end{aligned}$$

□

4. Main Results

In this section, we will make full use of previous results and then give a precise convergence rate analysis of algorithm 1. The elaboration will be divided into three parts. Firstly, we respectively establish the $O(\frac{1}{k})$ and $O(\frac{1}{k^2})$ convergence rate of $U_1(k) = \mathbb{E}[\|\bar{x}(k) - x_*\|^2]$ and $V_1(k) = \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|^2]$ based on two supporting lemmas in section 3.2. Secondly, we show that the convergence rate, measured by the mean-squared error of the decision variables, is as follows.

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] \leq \frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} + \frac{O\left(\frac{1}{\sqrt{n}(1-\rho_w)}\right)}{(k+K)^{1.5}} + \frac{O\left(\frac{1}{(1-\rho_w)^2}\right)}{(k+K)^2},$$

where the first term is only concerned with the stochastic gradient descent method which is network independent, while the higher-order depends on $(1-\rho_w)$. Finally, we characterize the transient time needed for CDSA to approach the asymptotic convergence rate is $O\left(\frac{n}{(1-\rho_w)^2}\right)$.

4.1. Sublinear Convergence

We first prove that the expected consensus error $V_1(k) = \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\|^2]$ decays with rate $V_1(k) = O(\frac{1}{k^2})$.

Lemma 4.1.1 *Let Assumption 2.2.1∼2.2.3 hold. Consider algorithm 1 with stepsize (25). Recall the definitions of K. Define*

$$\nabla \boldsymbol{H}(\boldsymbol{\theta}) \triangleq [\nabla h_1(\theta_1), \nabla h_2(\theta_2), \cdots \nabla h_n(\theta_n)] \in \mathbb{R}^{n\times q}, \tag{30}$$

$$K_1 \triangleq \Big\lceil \max\Big\{2K, \frac{16}{1-\rho_w^2}\Big\}\Big\rceil. \tag{31}$$

We then obtain from [30, Lemma 10] that for any $k \geq K_1 - K$,

$$V_2(k) \leq \frac{\hat{V}_2}{(k+K)^2} \text{ with } \hat{V}_2 \triangleq \max\Big\{K_1^2\hat{\Theta}, \frac{8\beta^2\rho_w^2 c_4^{'}}{\mu_\theta^2(1-\rho_w^2)}\Big\}, \tag{32}$$

$$\text{where } c_4^{'} \triangleq 2\bigg(\frac{3}{1-\rho_w^2} + M_\theta\bigg)\Big(L_\theta^2\hat{\Theta} + \|\nabla \boldsymbol{H}(\mathbf{1}\theta_*^T)\|\Big) + n\sigma_\theta^2. \tag{33}$$

Furthermore, we achieve that

$$V_1(k) \leq \frac{\hat{V}_1}{(k+K)^2} \text{ with } \hat{V}_1 \triangleq \max\left\{4K_1^2\hat{X}, \frac{8\beta^2\rho_w^2 c_4}{\mu_x^2(1-\rho_w^2)}\right\}, \tag{34}$$

$$\text{where } c_4 \triangleq 3\bigg(\frac{3}{1-\rho_w^2} + M_x\bigg)\Big(L_x^2\hat{X} + L_\theta^2\hat{\Theta} + \|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T, \mathbf{1}\theta_*^T)\|^2\Big) + n\sigma_x^2. \tag{35}$$

Proof We now prove (34). From Lemma 3.2.2 and 3.3.1 it follows that

$$V_1(k+1) \leq \frac{3+\rho_w^2}{4}V_1(k) + \alpha_k^2\rho_w^2 c_4, \quad \forall k \geq 0. \tag{36}$$

We use induction method to show that (34) holds for any $k \geq K_1 - K$. Recall from Lemma 3.3.2 that $V_1(k) \leq 4\hat{X}$. Then for $k = K_1 - K$, $V_1(k) \leq \frac{4K_1^2\hat{X}}{K_1^2} = \frac{4K_1^2\hat{X}}{(k+K)^2} \leq \frac{\hat{V}_1}{(k+K)^2}$ by the definition of $\hat{V}_1$ in (34). Suppose that (34) holds for $k = \tilde{k}$. It suffices to show that (34) holds for $k = \tilde{k}+1$.

Note from (31) that $\tilde{k} + K \geq \frac{16}{1-\rho_w^2}$ for any $\tilde{k} \geq K_1 - K$. We then have

$$\begin{aligned}\left(\frac{\tilde{k}+K}{\tilde{k}+K+1}\right)^2 - \frac{3+\rho_w^2}{4} &= 1 - \frac{2}{\tilde{k}+K+1} + \frac{1}{(\tilde{k}+K+1)^2} - \frac{3+\rho_w^2}{4} \\ &\geq \frac{1-\rho_w^2}{4} - \frac{2}{\tilde{k}+K} \geq \frac{1-\rho_w^2}{8}.\end{aligned}$$

Divide both sides of above inequality by $\frac{\beta^2\rho_w^2 c_4}{\mu_x^2}$. Recalling the definition of $\hat{V}_1$ in (34), we have

$$\frac{\beta^2\rho_w^2 c_4}{\mu_x^2}\left(\left(\frac{\tilde{k}+K}{\tilde{k}+K+1}\right)^2 - \frac{3+\rho_w^2}{4}\right)^{-1} \leq \frac{8\beta^2\rho_w^2 c_4}{\mu_x^2(1-\rho_w^2)} \leq \hat{V}_1. \tag{37}$$

This implies that

$$\frac{3+\rho_w^2}{4}\frac{\hat{V}_1}{(\tilde{k}+K)^2} + \frac{\beta^2\rho_w^2 c_4}{\mu_x^2}\frac{1}{(\tilde{k}+K)^2} \leq \frac{\hat{V}_1}{(\tilde{k}+K+1)^2}. \tag{38}$$

Then by using (36) and the definition of α_k in (25), we derive that $V_1(\tilde{k}+1) \leq \frac{\hat{V}_1}{(\tilde{k}+K+1)^2}$, i.e., (34) holds for $k = \tilde{k}+1$. Then the lemma is proved. □

In light of Lemma 4.1.1 and other auxiliary results, we establish the $O(\frac{1}{k})$ convergence rate of expected optimization error $U_1(k) = \mathbb{E}[\|\bar{x}(k) - x_*\|^2]$ in the following lemma.

Lemma 4.1.2 *Let Assumption 2.2.1~2.2.3 hold. Consider algorithm 1 with stepsize (25), where $\beta > 2$. We then have*

$$U_1(k) \leq \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2(k+K)} + \frac{(K_1+K)^{1.5\beta}}{(k+K)^{1.5\beta}}\frac{\hat{X}}{n} + \left[\frac{3\beta^2(1.5\beta-1)c_5}{(1.5\beta-2)n\mu_x^2} + \frac{12\beta L_x^2\hat{V}_1}{(1.5\beta-2)n\mu_x^2} + \frac{12\beta L_\theta^2\hat{V}_2}{(1.5\beta-2)n\mu_x^2}\right]\cdot\frac{1}{(k+K)^2}$$

for any $k \geq K_1 - K$, where

$$c_5 \triangleq \frac{3M_x L_x^2}{n}\hat{X} + \frac{3M_x L_\theta^2}{n}\hat{\Theta} + \bar{M}, \tag{39}$$

$\hat{X}, K_1, \hat{V}_2, \hat{V}_1, \bar{M}$ are defined by (29) (31) (32) (34) (17) respectively.

Proof Since $\alpha_k = \frac{\beta}{\mu_x(k+K)}$ by (25), recalling the definition of K and K_1 in (26) and (31), we can see that $\alpha_k \leq \frac{\beta}{\mu K_1} \leq \frac{\beta}{2\mu K} \leq \frac{\mu_x}{6(1+M_x)L_x^2} \leq \min\{\frac{1}{3\mu_x}, \frac{1}{L_x}\}$. Then Lemma 3.2.1(B) holds. Together with 3.3.1 it follows that for any $k \geq K_1 - K$,

$$U_1(k+1) \leq (1 - \frac{3}{2}\alpha_k\mu_x)U_1(k) + \frac{6\alpha_k L_x^2}{n\mu_x}V_1(k) + \frac{6\alpha_k L_\theta^2}{n\mu_x}V_2(k) + \frac{\alpha_k^2 c_5}{n}. \tag{40}$$

Recalling the definition of $\alpha_k = \frac{\beta}{\mu_x(k+K)}$, we have

$$U_1(k+1) \leq (1 - \frac{3\beta}{2(k+K)})U_1(k) + \frac{6\beta L_x^2 V_1(k)}{n\mu_x^2(k+K)} + \frac{6\beta L_\theta^2 V_2(k)}{n\mu_x^2(k+K)} + \frac{\beta^2 c_5}{n\mu_x^2}\cdot\frac{1}{(k+K)^2}. \tag{41}$$

Thus

$$\begin{aligned} U_1(k) &\leq \prod_{t=K_1+K}^{k+K-1}(1-\frac{3\beta}{2t})U_1(K_1) \\ &+ \sum_{t=K_1+K}^{k+K-1}\prod_{j=t+1}^{k+K-1}(1-\frac{3\beta}{2j})\left(\frac{6\beta L_x^2}{n\mu_x^2}\cdot\frac{V_1(t-K)}{t} + \frac{6\beta L_\theta^2}{n\mu_x^2}\cdot\frac{V_2(t-K)}{t} + \frac{\beta^2 c_5}{n\mu_x^2}\cdot\frac{1}{t^2}\right). \end{aligned}$$

Recall from [30, lemma 11] that for any $\forall 1 < j < k$, $j \in \mathbb{N}$ and $1 < \gamma \leq j/2$, $\prod_{t=j}^{k-1}(1-\frac{\gamma}{t}) \leq \frac{j^\gamma}{k^\gamma}$. Then we achieve

$$\begin{aligned} U_1(k) &\leq \frac{(K_1+K)^{1.5\beta}}{(k+K)^{1.5\beta}}U_1(K_1) \\ &+ \sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}}{(k+K)^{1.5\beta}}\left(\frac{6\beta L_x^2}{n\mu_x^2}\cdot\frac{V_1(t-K)}{t} + \frac{6\beta L_\theta^2}{n\mu_x^2}\cdot\frac{V_2(t-K)}{t} + \frac{\beta^2 c_5}{n\mu_x^2}\cdot\frac{1}{t^2}\right) \\ &= \frac{(K_1+K)^{1.5\beta}}{(k+K)^{1.5\beta}}U_1(K_1) + \frac{6\beta L_\theta^2}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}V_2(t-K)}{t} + \\ &\frac{6\beta L_x^2}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}V_1(t-K)}{t} + \frac{\beta^2 c_5}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}}{t^2}. \end{aligned}$$

In light of Lemma 4.1.1, we have $V_1(k-K) \leq \frac{\hat{V}_1}{k^2}$ and $V_2(k-K) \leq \frac{\hat{V}_2}{k^2}$ for any $k \geq K_1 - K$. Hence

$$\begin{aligned} U_1(k) \leq &\frac{\beta^2 c_5}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}}{t^2} + \frac{(K_1+K)^{1.5\beta}}{(k+K)^{1.5\beta}}U_1(K_1) \\ &+ \frac{6\beta L_x^2\hat{V}_1}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}}{t^3} + \frac{6\beta L_\theta^2\hat{V}_2}{n\mu_x^2(k+K)^{1.5\beta}}\sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{1.5\beta}}{t^3}. \end{aligned} \tag{42}$$

By the proof in [30, lemma 12], when $b > a \geq K_1$, we have

$$\sum_{t=a}^{b}\frac{(t+1)^{1.5\beta}}{t^2} \leq \frac{b^{1.5\beta-1}}{1.5\beta-1} + \frac{3(1.5\beta-1)b^{1.5\beta-2}}{1.5\beta-2},\ \sum_{t=a}^{b}\frac{(t+1)^{1.5\beta}}{t^3} \leq \frac{2b^{1.5\beta-2}}{1.5\beta-2}. \tag{43}$$

Thus

$$\begin{aligned} U_1(k) \leq & \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2(k+K)} + \frac{3\beta^2(1.5\beta-1)c_5}{(1.5\beta-2)n\mu_x^2}\cdot\frac{1}{(k+K)^2} + \frac{(K_1+K)^{1.5\beta}}{(k+K)^{1.5\beta}}U_1(K_1) \\ & + \frac{12\beta L_x^2\hat{V}_1}{(1.5\beta-2)n\mu_x^2}\cdot\frac{1}{(k+K)^2} + \frac{12\beta L_\theta^2\hat{V}_2}{(1.5\beta-2)n\mu_x^2}\cdot\frac{1}{(k+K)^2}. \end{aligned} \tag{44}$$

Recalling Lemma 3.3.2 yields the desired result. □

4.2. Rate Estimate

In this subsection, we will discuss the factors that affect the convergence rate of the algorithm, especially the network size n, the spectral gap $(1-\rho_w)$, the summation of initial optimization errors $\sum_{i=1}^n \|x_i(0)-x_*\|^2$ and consensus errors $\sum_{i=1}^n \|\theta_i(0)-\theta_*\|^2$, and the heterogenous of computational functions and learning functions characterized by $\sum_{i=1}^n \|\nabla_x f_i(x_*,\theta_*)\|^2$ and $\sum_{i=1}^n \|\nabla h_i(\theta_*)\|^2$. Firstly, we bound the constants appearing in Lemmas 4.1.1 and 4.1.2 by the aforementioned factors. We then utilize them to simplify the sublinear rate of the expected optimization error, based on which, we can improve the convergence rate and derive the main result for Algorithm 1.

Lemma 4.2.1 *Denote* $A_1 \triangleq \sum_{i=1}^n \|x_i(0)-x_*\|^2$, $B_1 \triangleq \sum_{i=1}^n \|\nabla_x f_i(x_*;\theta_*)\|^2$, $A_2 \triangleq \sum_{i=1}^n \|\theta_i(0)-\theta_*\|^2$, *and* $B_2 \triangleq \sum_{i=1}^n \|\nabla h_i(\theta_*)\|^2$. *Then the orders of constants'* $\hat{X}$ *(29),* $\hat{\Theta}$ *(27),* $\hat{V}_1$ *(34),* $\hat{V}_2$ *(32),* c_4 *(35) and* c_5 *(39) are as follow.*

$$\begin{aligned} \hat{X} &= O(A_1+A_2+B_1+B_2+n), \quad \hat{\Theta} = O(A_2+B_2+n), \\ \hat{V}_1 &= O\left(\frac{A_1+A_2+B_1+B_2+n}{(1-\rho_w)^2}\right), \quad \hat{V}_2 = O\left(\frac{A_2+B_2+n}{(1-\rho_w)^2}\right), \\ c_4 &= O\left(\frac{A_1+A_2+B_1+B_2+n}{1-\rho_w}\right), \quad c_5 = O\left(\frac{A_1+A_2+B_1+B_2+n}{n}\right). \end{aligned}$$

Proof The upper bound of $\hat{\Theta}$ and $\hat{V}_2$ deal only with unknown parameter θ, which can be inherited directly from [30, lemma 13]. As for $\hat{X}$, recalling (29) we have

$$\begin{aligned} \hat{X} \leq & \|\mathbf{x}(0)-\mathbf{1}x_*^T\|^2 + \frac{11L_\theta^2\hat{\Theta}}{\mu_x^2} + \frac{11\|\nabla_x F(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2}{\mu_x^2} + \frac{7n\sigma_x^2}{9(1+M_x)L_x^2} \\ = & O(A_1+A_2+B_1+B_2+n). \end{aligned} \tag{45}$$

From the definition of c_4 in (35), it follows that

$$c_4 = 3\left(\frac{3}{1-\rho_w^2}+M_x\right)(L_x^2\hat{X}+L_\theta^2\hat{\Theta}+\|\nabla_x F(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2) + n\sigma_x^2 = O\left(\frac{A_1+A_2+B_1+B_2+n}{1-\rho_w}\right) \tag{46}$$

Note from (31) and (26) that $K_1 = O\left(\frac{1}{1-\rho_w}\right)$. Then by (34) , we obtain

$$\hat{V}_1 = \max\left\{4K_1^2\hat{X}, \frac{8\beta^2\rho_w^2 c_4}{\mu_x^2(1-\rho_w^2)}\right\} = O\left(\frac{A_1+A_2+B_1+B_2+n}{(1-\rho_w)^2}\right). \tag{47}$$

In light of equation (39), we can achieve

$$c_5 = \frac{3M_xL_x^2}{n}\hat{X} + \frac{3M_xL_\theta^2}{n}\hat{\Theta} + \bar{M}_x = O\left(\frac{A_1+A_2+B_1+B_2+n}{n}\right). \tag{48}$$

□

The simplification of these constants makes it convenient for later analysis. In light of relation (34), since $\hat{V}_1$ is the only constant, the convergence result of expected consensus error $V_1(k)$ can be easily obtained. While the expected optimization error $U_1(k)$ needs to be reformulated more concisely.

Corollary 4.1 *Let Assumption 2.2.1~2.2.3 hold. Consider algorithm 1 with stepsize policy (25), where $\beta > 2$. Then we obtain from [30, Corollary 1] that*

$$U_2(k) \leq \frac{\beta^2 c_5'}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{(k+K)} + \frac{c_6'}{(k+K)^2}, \quad \forall k \geq K_1 - K,$$

where $c_5' \triangleq \frac{2M_\theta L_\theta^2}{n}\hat{\Theta} + \frac{2M_\theta \sum_{i=1}^n \|\nabla h_i(\theta_*)\|^2}{n} + \sigma_\theta^2$, $c_6' = O\left(\frac{A_2+B_2+n}{n(1-\rho_w)^2}\right)$. Based on which, we further have that for any $k \geq K_1 - K$,

$$U_1(k) \leq \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{(k+K)} + \frac{c_6}{(k+K)^2}, \tag{49}$$

where c_5 is defined in (39), and $c_6 = O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2}\right)$.

Proof In light of Lemma 4.1.2 and Lemma 4.2.1, we can obtain that

$$\begin{aligned}
U_1(k) &\leq \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2(k+K)} + \frac{(K_1+K)^{1.5\beta-2}}{(k+K)^{1.5\beta-2}}\frac{\hat{X}}{n} \cdot \frac{1}{(k+K)^2} \\
&\quad + \left[\frac{3\beta^2(1.5\beta-1)c_5}{(1.5\beta-2)n\mu_x^2} + \frac{12\beta L_x^2 \hat{V}_1}{(1.5\beta-1)n\mu_x^2} + \frac{12\beta L_\theta^2 \hat{V}_2}{(1.5\beta-1)n\mu_x^2}\right] \cdot \frac{1}{(k+K)^2} \\
&= \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{(k+K)} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n}\right)\frac{1}{(k+K)^2} \\
&\quad + \left[O\left(\frac{A_1+A_2+B_1+B_2+n}{n^2}\right) + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2}\right) + O\left(\frac{A_2+B_2+n}{n(1-\rho_w)^2}\right)\right]\frac{1}{(k+K)^2} \\
&\leq \frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{(k+K)} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2}\right)\frac{1}{(k+K)^2}.
\end{aligned}$$

□

Based on this corollary, together with Lemma 3.2.1, we further elaborate the convergence result of Algorithm 1. Especially, we give an upper bound of $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|x_i(k) - x_*\|^2\right]$ and formulate it in a way to make an intuitive comparison with the centralized algorithm.

Theorem 4.1 *Let Assumption 2.2.1~2.2.3 hold. Consider algorithm 1 with stepsize policy (25), where $\beta > 2$. Then for any $k \geq K_1 - K$, we have*

$$\begin{aligned}
\frac{1}{n}\sum\nolimits_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] &\leq \frac{\beta^2 \bar{M}}{(2\beta-1)n\mu_x^2(k+K)} \\
&\quad + O\left(\frac{A_1+A_2+B_1+B_2+n}{n\sqrt{n}(1-\rho_w)}\right)\frac{1}{(k+K)^{1.5}} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2}\right)\frac{1}{(k+K)^2},
\end{aligned} \tag{50}$$

where $\bar{M}$ is defined in (17).

Proof For $k \geq K_1 - K$, by recalling Lemma 3.2.1(**A**) and the definition of $U_1(k), V_1(k)$ and $U_2(k), V_2(k)$ in (19) and

(20), we have

$$
\begin{aligned}
U_1(k+1) &\le (1-\alpha_k\mu_x)^2 U_1(k) + \frac{\alpha_k^2 L_x^2}{n} V_1(k) + \frac{\alpha_k^2 L_\theta^2}{n} V_2(k) + \frac{2L_x L_\theta \alpha_k^2}{n}\sqrt{V_1(k)V_2(k)} \\
&\quad + \frac{2\alpha_k L_x}{\sqrt{n}}\sqrt{U_1(k)V_1(k)} + \frac{2\alpha_k L_\theta}{\sqrt{n}}\sqrt{U_1(k)V_2(k)} \\
&\quad + \alpha_k^2\left(\frac{3M_x L_x^2}{n^2}(nU_1(k)+V_1(k)) + \frac{3M_x L_\theta^2}{n^2}(nU_2(k)+V_2(k)) + \frac{\bar{M}}{n}\right) \\
&= (1-2\alpha_k\mu_x)U_1(k) + \alpha_k^2\left(\mu_x^2 + \frac{3M_x L_x^2}{n}\right)U_1(k) + \alpha_k^2\cdot\frac{3M_x L_\theta^2}{n}U_2(k) \\
&\quad + \frac{\alpha_k^2 L_x^2}{n}\left(1+\frac{3M_x}{n}\right)V_1(k) + \frac{\alpha_k^2 L_\theta^2}{n}\left(1+\frac{3M_x}{n}\right)V_2(k) + \frac{2L_x L_\theta\alpha_k^2}{n}\sqrt{V_1(k)V_2(k)} \\
&\quad + \frac{2\alpha_k L_x}{\sqrt{n}}\sqrt{U_1(k)V_1(k)} + \frac{2\alpha_k L_\theta}{\sqrt{n}}\sqrt{U_1(k)V_2(k)} + \frac{\alpha_k^2\bar{M}}{n},
\end{aligned}
$$

where the first inequality follows the Cauchy-Schwarz inequality in the probabilistic form and the fact that

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] &= \mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{1}\bar{x}^T + \mathbf{1}\bar{x}^T - \mathbf{1}x_*^T\|^2\right] \\
&\le 2\mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{1}\bar{x}^T\|^2\right] + 2\mathbb{E}\left[\|\mathbf{1}\bar{x}^T - \mathbf{1}x_*^T\|^2\right] \\
&= 2\mathbb{E}\left[\|\mathbf{x}(k) - \mathbf{1}\bar{x}^T\|^2\right] + 2n\mathbb{E}\left[\|\bar{x} - x_*\|^2\right] = 2V_1(k) + 2nU_1(k).
\end{aligned} \tag{51}
$$

Thus, due to $\alpha_k = \frac{\beta}{\mu_x(k+K)}$, we have

$$
\begin{aligned}
U_1(k+1) &\le \left(1 - \frac{2\beta}{k+K}\right)U_1(k) + \frac{\beta^2 U_1(k)}{(k+K)^2}\left(1 + \frac{3M_x L_x^2}{n\mu_x^2}\right) + \frac{3M_x L_\theta^2\beta^2 U_2(k)}{n\mu_x^2(k+K)^2} \\
&\quad + \frac{\beta^2 L_x^2}{n\mu_x^2}\left(1+\frac{3M_x}{n}\right)\frac{V_1(k)}{(k+K)^2} + \frac{\beta^2 L_\theta^2}{n\mu_x^2}\left(1+\frac{3M_x}{n}\right)\frac{V_2(k)}{(k+K)^2} \\
&\quad + \frac{2L_x L_\theta\beta^2}{n\mu_x^2}\frac{\sqrt{V_1(k)V_2(k)}}{(k+K)^2} + \frac{2\beta L_x}{\sqrt{n}\mu_x}\frac{\sqrt{U_1(k)V_1(k)}}{k+K} + \frac{2\beta L_\theta}{\sqrt{n}\mu_x}\frac{\sqrt{U_1(k)V_2(k)}}{k+K} + \frac{\beta^2\bar{M}}{n\mu_x^2}\frac{1}{(k+K)^2}
\end{aligned}
$$

Denote by $c_7 = 1 + \frac{3M_x L_x^2}{n\mu_x^2}$ and $c_8 = 1 + \frac{3M_x}{n}$. Then in light of [30, Lemma 11], we obtain that

$$
\begin{aligned}
U_1(k) &\le \prod_{t=K_1+K}^{k+K-1}\left(1-\frac{2\beta}{t}\right)U_1(K_1) + \sum_{t=K_1+K}^{k+K-1}\left(\prod_{i=t+1}^{k+K-1}\left(1-\frac{2\beta}{i}\right)\right)\Bigg[\frac{\beta^2\bar{M}}{n\mu_x^2 t^2} \\
&\quad + \frac{3M_x L_\theta^2\beta^2}{n\mu_x^2}\frac{U_2(t-K)}{t^2} + \frac{\beta^2 L_x^2 c_8}{n\mu_x^2}\frac{V_1(t-K)}{t^2} + \frac{\beta^2 L_\theta^2 c_8}{n\mu_x^2}\frac{V_2(t-K)}{t^2} + \frac{2L_x L_\theta\beta^2}{n\mu_x^2}\frac{\sqrt{V_1(t-K)V_2(t-K)}}{t^2} \\
&\quad + \frac{\beta^2 c_7 U_1(t-K)}{t^2} + \frac{2\beta L_x}{\sqrt{n}\mu_x}\frac{\sqrt{U_1(t-K)V_1(t-K)}}{t} + \frac{2\beta L_\theta}{\sqrt{n}\mu_x}\frac{\sqrt{U_1(t-K)V_2(t-K)}}{t}\Bigg] \\
&\le \frac{(K_1+K)^{2\beta}}{(k+K)^{2\beta}}U_1(K_1) + \sum_{t=K_1+K}^{k+K-1}\frac{(t+1)^{2\beta}}{(k+K)^{2\beta}}\Bigg[\frac{\beta^2\bar{M}}{n\mu_x^2 t^2} + \frac{\beta^2 c_7 U_1(t-K)}{t^2} \\
&\quad + \frac{\beta^2 L_x^2 c_8}{n\mu_x^2}\frac{V_1(t-K)}{t^2} + \frac{\beta^2 L_\theta^2 c_8}{n\mu_x^2}\frac{V_2(t-K)}{t^2} + \frac{2L_x L_\theta\beta^2}{n\mu_x^2}\frac{\sqrt{V_1(t-K)V_2(t-K)}}{t^2} \\
&\quad + \frac{3M_x L_\theta^2\beta^2 U_2(t-K)}{n\mu_x^2 t^2} + \frac{2\beta L_x\sqrt{U_1(t-K)V_1(t-K)}}{\sqrt{n}\mu_x t} + \frac{2\beta L_\theta\sqrt{U_1(t-K)V_2(t-K)}}{\sqrt{n}\mu_x t}\Bigg].
\end{aligned} \tag{52}
$$

According to Corollary 4.1 and Lemma 4.1.1, we have

$$
\begin{aligned}
U_1(k) \leq & \frac{(K_1+K)^{2\beta}}{(k+K)^{2\beta}} U_1(K_1) + \frac{1}{(k+K)^{2\beta}} \cdot \frac{\beta^2 \bar{M}}{n\mu_x^2} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \\
& + \frac{\beta^2 c_7}{(k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \left[\frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{t} + \frac{c_6}{t^2} \right] \\
& + \frac{3M_x L_\theta^2 \beta^2}{n\mu_x^2 (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \left[\frac{\beta^2 c_5^{'}}{(1.5\beta-1)n\mu_\theta^2} \cdot \frac{1}{t} + \frac{c_6^{'}}{t^2} \right] \\
& + \frac{\beta^2 L_x^2 c_8}{n\mu_x^2 (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \cdot \frac{\hat{V}_1}{t^2} \\
& + \frac{\beta^2 L_\theta^2 c_8}{n\mu_x^2 (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \cdot \frac{\hat{V}_2}{t^2} + \frac{2L_x L_\theta \beta^2}{n\mu_x^2 (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2} \frac{\sqrt{\hat{V}_1 \hat{V}_2}}{t^2} \\
& + \frac{2\beta L_x}{\sqrt{n}\mu_x (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t} \sqrt{\frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{t} + \frac{c_6}{t^2}} \sqrt{\frac{\hat{V}_1}{t^2}} \\
& + \frac{2\beta L_\theta}{\sqrt{n}\mu_x (k+K)^{2\beta}} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t} \sqrt{\frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{t} + \frac{c_6}{t^2}} \sqrt{\frac{\hat{V}_2}{t^2}}.
\end{aligned}
$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we can achieve

$$
\sqrt{\frac{\beta^2 c_5}{(1.5\beta-1)n\mu_x^2} \cdot \frac{1}{t} + \frac{c_6}{t^2}} \cdot \sqrt{\frac{\hat{V}_1}{t^2}} \leq \beta \sqrt{\frac{c_5 \hat{V}_1}{(1.5\beta-1)n\mu_x^2}} \cdot \frac{1}{t^{1.5}} + \frac{\sqrt{c_6 \hat{V}_1}}{t^2}, \tag{53}
$$

then

$$
\begin{aligned}
U_1(k) \leq & \frac{\beta^2 \bar{M} \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^2}}{(k+K)^{2\beta} n\mu_x^2} + \frac{(K_1+K)^{2\beta} U_1(K_1)}{(k+K)^{2\beta}} + \frac{2\beta^2 \sqrt{c_5}(L_x \sqrt{\hat{V}_1} + L_\theta \sqrt{\hat{V}_2}) \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^{2.5}}}{\sqrt{1.5\beta-1} \times n\mu_x^2 (k+K)^{2\beta}} \\
& + \frac{1}{(k+K)^{2\beta}} \left[\frac{\beta^4 c_5 c_7}{(1.5\beta-1)n\mu_x^2} + \frac{3M_x \beta^4 L_\theta^2 c_5^{'}}{(1.5\beta-1)n^2 \mu_x^2 \mu_\theta^2} + \frac{2\beta(L_x \sqrt{c_6 \hat{V}_1} + L_\theta \sqrt{c_6^{'} \hat{V}_2})}{\sqrt{n}\mu_x} \right] \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^3} \\
& + \frac{1}{(k+K)^{2\beta}} \left[\beta^2 c_6 c_7 + \frac{3M_x \beta^2 L_\theta^2 c_6^{'}}{n\mu_x^2} + \frac{\beta^2 (L_x^2 \hat{V}_1 + L_\theta^2 \hat{V}_2) c_8}{n\mu_x^2} + \frac{2L_x L_\theta \beta^2 \sqrt{\hat{V}_1 \hat{V}_2}}{n\mu_x^2} \right] \sum_{t=K_1+K}^{k+K-1} \frac{(t+1)^{2\beta}}{t^4}.
\end{aligned}
$$

Recall (43) and note that

$$
\begin{aligned}
\sum_{t=a}^{b} \frac{(t+1)^{2\beta}}{t^{2.5}} \leq \sum_{t=a}^{b} \frac{2(t+1)^{2\beta}}{(t+1)^{2.5}} \leq \int_{a+1}^{b+1} 2t^{2\beta-2.5} dt \leq \frac{2(b+1)^{2\beta-1.5}}{2\beta-1.5}, \\
\sum_{t=a}^{b} \frac{(t+1)^{2\beta}}{t^4} \leq \sum_{t=a}^{b} \frac{2(t+1)^{2\beta}}{(t+1)^4} \leq \int_{a+1}^{b+1} 2t^{2\beta-4} dt \leq \frac{2(b+1)^{2\beta-3}}{2\beta-4}, \quad \forall a \geq 16.
\end{aligned} \tag{54}
$$

Then by noticing that $c_7 = c_8 = O(1)$ and using Lemma 4.2.1, we have

$$
\begin{aligned}
U_1(k) \leq & \frac{\beta^2 \bar{M}}{(2\beta-1)n\mu_x^2 (k+K)} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n\sqrt{n}(1-\rho_w)} \right) \frac{1}{(k+K)^{1.5}} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2} \right) \frac{1}{(k+K)^2} \\
& + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2} \right) \frac{1}{(k+K)^3} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^{2\beta}} \right) \frac{1}{(k+K)^{2\beta}} \\
= & \frac{\beta^2 \bar{M}}{(2\beta-1)n\mu_x^2 (k+K)} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n\sqrt{n}(1-\rho_w)} \right) \frac{1}{(k+K)^{1.5}} + O\left(\frac{A_1+A_2+B_1+B_2+n}{n(1-\rho_w)^2} \right) \frac{1}{(k+K)^2}.
\end{aligned}
$$

By recalling (51), we have $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2] \le 2U_1(k)+\frac{2V_1(k)}{n}$. This together with (34) and the estimate of $\hat{V}_1$ in Lemma 4.2.1 prove the result. □

In light of relation (50), by recalling the definitions of A_1, A_2, B_1, B_2 in Lemma 4.2.1, we can see that the convergence rate is proportional to initial errors for both computational problem $\sum_{i=1}^{n}\|x_i(0)-x_*\|^2$ and parameter learning problem $\sum_{i=1}^{n}\|\theta_i(0)-\theta_*\|^2$. It is worth noting that the heterogeneity of agents' individual cost functions, measured by $B_1=\sum_{i=1}^{n}\|\nabla_x f_i(x_*;\theta_*)\|^2$, $B_2=\sum_{i=1}^{n}\|\nabla h_i(\theta_*)\|^2$, also influence the convergence rate in a similar way. Though θ_*, x_* are respectively the optimal solutions to $\min_\theta \frac{1}{n}\sum_{i=1}^{n} h_i(\theta)$ and $\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x;\theta^*)$, they are usually not the optimal solution to each local function $h_i(\theta), f_i(x,\theta)$. Therefore, the bigger the difference between the local costs, the slower the convergence rate of the algorithm.

Remark 2 *Here we give some comments regerading the influence of the network size n and the spectral gap* $(1-\rho_w)$ *on the convergence rate. Since* A_1, A_2, B_1 *and* B_2 *are all* $O(n)$*, we can simplify the relation (50) as follow.*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2] \le \frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} + \frac{O\left(\frac{1}{\sqrt{n}(1-\rho_w)}\right)}{(k+K)^{1.5}} + \frac{O\left(\frac{1}{(1-\rho_w)^2}\right)}{(k+K)^2}. \tag{55}$$

It is noticed that the algorithm converges faster for better network connectivity (i.e., smaller ρ_w*). For example, a fully connected graph is the most efficient connection topology since* $\rho_w = 0$*. In contrast, it holds* $1-\rho_w \to 0$ *as* $n \to \infty$ *for the cycle graph, which indicates that the algorithm will converge very slowly for large-scale cycle graphs. The following table taken from [35, Chapter 4] characterizes the relation between network size n and the spectral gap. Considering plugging the order concerning n from the table into relation (55), we may obtain the quantitive influence of the network size on the convergence rate.*

Table 2: Relation between the network size n and the spectral gap $1-\rho_w$

Network Topology	Spectral Gap$(1-\rho_w)$	Network Topology	Spectral Gap$(1-\rho_w)$
Path Graph	$O(\frac{1}{n^2})$	2D-mesh Graph	$O(\frac{1}{n})$
Cycle Graph	$O(\frac{1}{n^2})$	Complete Graph	1

There are other factors such as the strong convexity and Lipschitz smoothness parameters, as well as the variance of the stochastic gradient, all of which can also affect the convergence rate. We will not include a quantitative analysis of these factors since the big O constant in the convergence rate is already quite complex and we often use the relation like $\mu_x \le L_x$ for simplicity. While some intuitive property can be naturally obtained from (55): the larger convexity and Lipschitz smoothness parameters can lead to the faster rate; the higher variance of stochastic gradient descent leads to a lower convergence rate since term $\bar{M}$ defined by (17) gets bigger.

4.3. Transient Time

In this subsection, we will establish the transient iteration needed for the CDSA algorithm to reach its dominant rate.

Firstly, we recall the convergence rate from [30, Theorem 2] for the centralized stochastic gradient descent,

$$\mathbb{E}\left[\|x(k)-x_*\|^2\right] \le \frac{\beta^2\bar{M}}{(2\beta-1)n\mu^2 k} + O\left(\frac{1}{n}\right)\frac{1}{k^2}. \tag{56}$$

Comparing it to (55), we may conclude that our distributed algorithm converges to the optimal solution at a comparable rate to the centralized algorithm, since they are both of the same order $O(\frac{1}{k})$. Besides, our work demonstrates that the network connectivity ρ_w does not influence the term $O(\frac{1}{k})$, it only appears in higher-order terms $O(\frac{1}{k^{1.5}})$ and $O(\frac{1}{k^2})$. Though our distributed algorithm asymptotically reaches the same order of convergence rate as that of the centralized algorithm, it's unclear how many iterations it takes to reach the dominate order $O(\frac{1}{k})$ since there are two extra error terms $O(\frac{1}{k^{1.5}})$ and $O(\frac{1}{k^2})$ induced by averaging consensus. We refer to the number of iterations before distributed stochastic approximation method reaches its dominant rate as **transient iterations**, i.e., when iteration k is relatively

small, the terms other than n and k still dominate the convergence rate[36, Section 2]. The next theorem state the iterations needed for Algorithm 1 to reach its dominant rate.

Theorem 4.2 *Let Assumption 2.2.1~2.2.3 hold, and set stepsize as (25), where $\beta > 2$. It takes $K_T = O(\frac{n}{(1-\rho_w)^2})$ iteration counts for algorithm 1 to reach the asymptotic rate of convergence, i.e. when $k \geq K_T$, we have $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] \leq \frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2 k}O(1)$.*

Proof Recalling relation in eq. (55), we see that for any $k \geq O\left(\frac{n}{(1-\rho_w)^2}\right)$,

$$\frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} \geq O(\frac{1}{\sqrt{n}(1-\rho_w)})\frac{1}{(k+K)^{1.5}},$$
$$\frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} \geq O\left(\frac{1}{(1-\rho_w)^2}\right)\frac{1}{(k+K)^2}.$$

□

5. Experiments

In this section, we will provide numerical examples to verify our theoretical findings, and carry out experiments by Bluefog[1]. It is a python library that can be connected to the NVIDIA Collective Communications Library (NCCL) for multi-GPU computing or Message Passing Interface (MPI) library for multi-CPU computing[37], i.e., each agent in our distributed experiment scenario is CPU.

5.1. Ridge Regression

Consider the following ridge-distributed regression problem with an unknown regularization parameter θ_*,

$$C_x(\theta_*) : \min_{x\in\mathbb{R}^p} \sum\nolimits_{i=1}^n \mathbb{E}_{u_i,v_i}\left[\left(u_i^T x - v_i\right)^2 + \theta_*\|x\|^2\right],$$

where θ_* can be obtained by the distributed learning problem below,

$$\mathcal{L}_\theta : \theta_* = \operatorname{argmin} \sum\nolimits_{i=1}^n (\theta - \alpha_i)^2 .$$

Specially, for agent $i \in \mathcal{N} \triangleq \{1, \cdots, n\}$, its local objective functions are specified as

$$f_i(x;\theta) = \min_x \mathbb{E}_{u_i,v_i}\left[\left(u_i^T x - v_i\right)^2 + \theta\|x\|^2\right], h_i(\theta) = \min_\theta(\theta - \alpha_i)^2.$$

Here (u_i, v_i) are data sample collected by each agent i, where $u_i \in \mathbb{R}^p$ are the sample features, while $v_i \in \mathbb{R}$ represent the observed outputs.

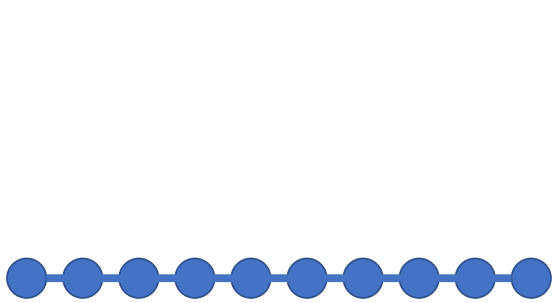
Parameter settings. Set $p = 5$ and suppose that for all $i \in \mathcal{N}$, each component of $u_i \in \mathbb{R}^p$ is an independent identical distribution in $U(-0.5, 0.5)$, and v_i is drawn according to $v_i = u_i^T\widetilde{x_i} + \epsilon_i$, where ϵ_i is an gaussian random variable specified by $N(0, 0.01)$, and $\widetilde{x_i} = (1\ 3\ 5\ 4\ 9)$ is a predefined parameter. Set $\alpha_i = 0.01 \times i$. It can be easily calculated that the optimal solutions are $\theta_* = 0.005(n+1)$, and $x_* = \left[\sum_{i=1}^n \mathbb{E}_{u_i}(u_i u_i^T) + n\theta_*\mathbf{I}\right]\mathbb{E}_{u_i}(u_i u_i^T) = \frac{1}{12}(\frac{1}{12} + \theta_*)^{-1}\frac{1}{n}\sum_{i=1}^n \widetilde{x_i}$.

We compare the performance of Algorithm 1 under the path graph and complete graph topology with different network size n. In light of the results in table 2 of the path graph and complete graph, convergence rate estimation can be reformulated.

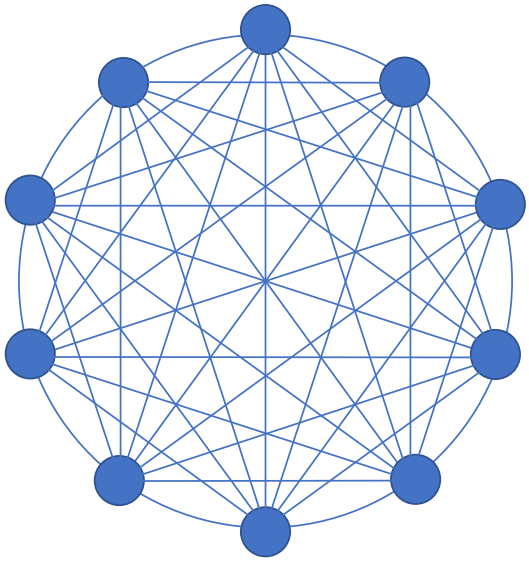
$$Path : \frac{1}{n}\sum\nolimits_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] \leq \frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} + \frac{O(n^{3/2})}{(k+K)^{1.5}} + \frac{O(n^2)}{(k+K)^2}, \tag{57}$$
$$Complete : \frac{1}{n}\sum\nolimits_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] \leq \frac{\beta^2\bar{M}}{(2\beta-1)n\mu_x^2(k+K)} + \frac{O(1/\sqrt{n})}{(k+K)^{1.5}} + \frac{1}{(k+K)^2}. \tag{58}$$

[1] https://github.com/Bluefog-Lib/bluefog

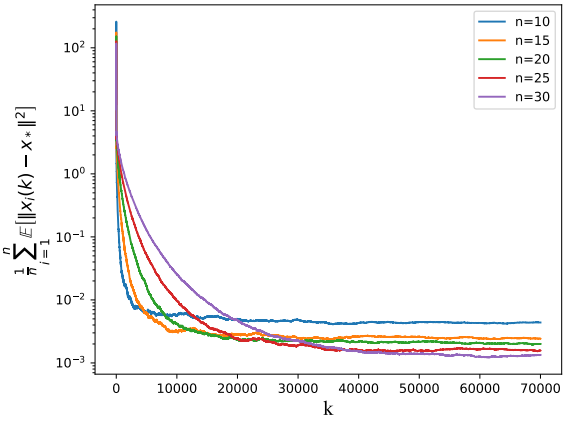
We run Algorithm 1, where the initial values are set as $(x_i(0), \theta_i(0)) = (\mathbf{0}_5, 1)\forall i$, and the weighted adjacency matrix of the communication network is built according to the Metropolis-Hastings rule [12]. According to (25), we choose the stepsizes as $\alpha_k = \gamma_k = \frac{20}{k+20}$ for any $k \geq 0$.



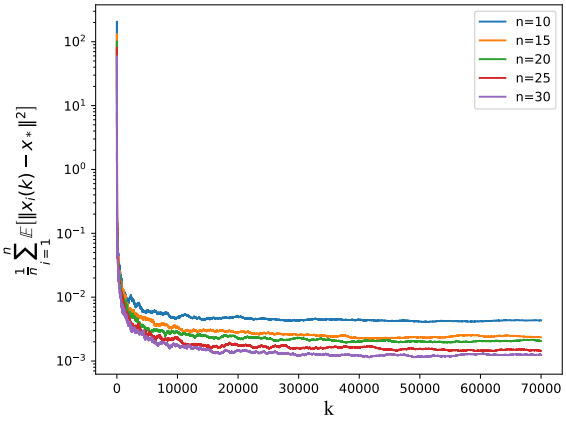
(a.1)n=10 path graph topology



(b.1)n=10 complete graph topology



(a.2) The performance of path graph



(b.2) The performance of complete graph

Figure 2: The performance of CDSA between path graph and complete graph topology. The results are averaged over 200 Monte Carlo sampling.

We demonstrate the empirical results in Fig. 2, where the empirical mean-squared error $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k) - x_*\|^2]$ is calculated by averaging through 200 sample paths. We can see from the Subfigure (a.2) that for the path graph, when the iterate k is small, the larger network size n will lead to the higher mean-squared error $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k) - x_*\|^2]$. However, with the increase of k, we observe a phase transition that a larger network size n will lead to a smaller mean-squared error $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k) - x_*\|^2]$ (namely faster convergence rate). This phenomenon matches the theoretical result (57): when k is small, the main factor influencing the convergence rate is the second and third term concerning the network size n via the distributed consensus protocol, while when k is large, the first term inherited from centralized stochastic gradient descent dominates the convergence rate.

Compared it to the empirical performance of the complete graph shown in subfigure (b.2), we can find that from the beginning to the end, a larger network n generates smaller errors, which also matches (58).

5.2. Logistic Regression

We further consider convex but not strongly convex problem, and use logistic regression to demonstrate that our algorithm can also leads to asymptotic convergence.

Consider the binary classification via logistic regression with unknown regularization parameter θ_*,

$$C_\eta(\theta_*): \ \min_\eta \sum_{i=1}^n \sum_{j=1}^{m_i} \ln\left(1+e^{-\eta^T x_{ij} l_{ij}}\right) + \frac{\theta_*}{2}\|\eta\|^2,$$

where θ_* can be obtained by a distributed parameter learning problem as follow,

$$\mathcal{L}_\theta: \ \theta_* = \operatorname{argmin} \sum_{i=1}^n (\theta-\alpha_i)^2 .$$

As for agent i, its its own local computational problem and parameter learning problem are as follows.

$$f_i(\eta;\theta) = \min_\eta \sum_{j=1}^{m_i} \ln\left(1+e^{-\eta^T x_{ij} l_{ij}}\right) + \frac{\theta_*}{2n}\|\eta\|^2, \ h_i(\theta) = \min_\theta (\theta-\alpha_i)^2.$$

In this scenario, we set $\alpha_i = 0.01 \times i$ and let each agent $i \in \mathcal{N}$ possess dataset $\mathcal{D}_i \triangleq \{(x_{ij}, l_{ij}) : j = 1, \cdots m_i\}$, where x_{ij} represents a three-dimensional sample feature where the first dimension is 1 and the other two dimension are selected from $N((1,0)^T, \boldsymbol{I})$ or $N((0,1)^T, \boldsymbol{I})$, while l_{ij} is the related sample label 1 or -1 respectively. Suppose that every agent holds a number of positive samples and negative samples which only accessible to itself.

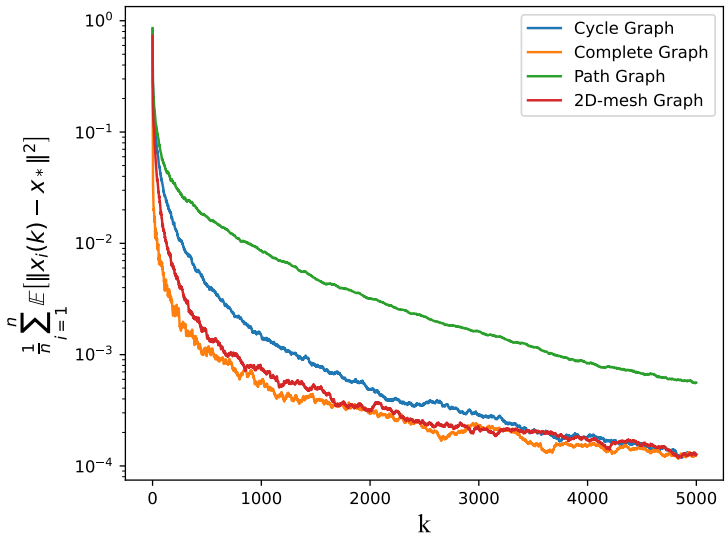


Figure 3: The performance of CDSA of 25 agents under four topologies in table 2 for binary classification via logistic regression. The results are averaged over 200 Monte Carlo sampling.

We now compare the empirical performance of Algorithm 1 under four classes of graph topologies, path graph, cycle graph, 2D-mesh graph, and complete graph. We set $n = 25$ and run Algorithm 1 with initial values $(\eta_i(0), \theta_i(0)) = \mathbf{0}_4$ for all $i \in \mathcal{N}$, where the stepsize and weighted adjacency matrix are set the same as Ridge Regression. The empirical results are shown in fig. 3, which shows that the complete graph has best performance, 2D-mesh graph has the second-best performance, while the path graph displays the worst performance. These empirical findings match that listed in table 2, where the 2D-mesh graph has a larger spectral gap than the path graph and cycle path, hence leads to a lower mean-squared error.

6. Conclusions

In this work, we consider the distributed optimization problem $\min_x \frac{1}{n}\sum_{i=1}^n f_i(x;\theta^*)$ with the unknown parameter θ^* collaboratively solved by a distributed parameter learning problem $\min_\theta \frac{1}{n}\sum_{i=1}^n h_i(\theta)$. Each agent only has access to its local computational problem $f_i(x,\theta)$ and its parameter learning problem $h_i(\theta)$. We propose a coupled distributed stochastic approximation algorithm for resolving this special distributed optimization, where agents can exchange information about decision variables x and learning parameter θ with neighbors over a connected network. We quantitatively characterize the factors that influence the rate of convergence, and validates that the algorithm asymptotically achieves the optimal network-independent convergence rate compared to the centralized algorithm scheme. In addition, we analyze the transient time K_T, and show that when the iterate $k \geq K_T$, the dominate factor influencing the convergence rate is related to stochastic gradient descent, while for small $k < K_T$, the main factor influencing the convergence rate originates from the distributed average consensus method. Future work will consider more general problems under weakened assumptions. It is of interests to explore the accelerated algorithm to obtain a faster convergence rate.

References

[1] G. Binetti, A. Davoudi, D. Naso, B. Turchiano, F. L. Lewis, A distributed auction-based algorithm for the nonconvex economic dispatch problem, IEEE Transactions on Industrial Informatics 10 (2) (2014) 1124–1132. doi:10.1109/TII.2013.2287807.

[2] P. Yi, Y. Hong, F. Liu, Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems, Automatica 74 (2016) 259–269. doi:https://doi.org/10.1016/j.automatica.2016.08.007.

[3] A. Cortés, S. MartÍnez, A projection-based decomposition algorithm for distributed fast computation of control in microgrids, SIAM Journal on Control and Optimization 56 (2) (2018) 583–609. doi:10.1137/15M103889X.
URL https://doi.org/10.1137/15M103889X

[4] S. Sahyoun, S. M. Djouadi, K. Tomsovic, S. Lenhart, Optimal Distributed Control for Continuum Power Systems, pp. 416–422. doi:10.1137/1.9781611974072.57.
URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974072.57

[5] L.-N. Liu, G.-H. Yang, Distributed optimal economic environmental dispatch for microgrids over time-varying directed communication graph, IEEE Transactions on Network Science and Engineering 8 (2) (2021) 1913–1924. doi:10.1109/TNSE.2021.3076526.

[6] V. Krishnan, S. Martínez, Distributed control for spatial self-organization of multi-agent swarms, SIAM Journal on Control and Optimization 56 (5) (2018) 3642–3667. doi:10.1137/16M1080926.
URL https://doi.org/10.1137/16M1080926

[7] T. Skibik, M. M. Nicotra, Analysis of time-distributed model predictive control when using a regularized primal–dual gradient optimizer, IEEE Control Systems Letters 7 (2022) 235–240. doi:10.1109/LCSYS.2022.3186631.

[8] Y.-L. Yang Tao, Xu Lei, Event-triggered distributed optimization algorithms, Acta Automatica Sinica 48 (1) (2022) 133–143. doi:10.16383/j.aas.c200838.

[9] A. Nedic, Distributed gradient methods for convex machine learning problems in networks: Distributed optimization, IEEE Signal Processing Magazine 37 (3) (2020) 92–101. doi:10.1109/MSP.2020.2975210.

[10] S. A. Alghunaim, A. H. Sayed, Distributed coupled multiagent stochastic optimization, IEEE Transactions on Automatic Control 65 (1) (2020) 175–190. doi:10.1109/TAC.2019.2906495.

[11] B. Touri, B. Gharesifard, A unified framework for continuous-time unconstrained distributed optimization, SIAM Journal on Control and Optimization 61 (4) (2023) 2004–2020. doi:10.1137/21M1442711.
URL https://doi.org/10.1137/21M1442711

[12] G. Notarstefano, I. Notarnicola, A. Camisa, Distributed optimization for smart cyber-physical networks, Foundations and Trends in Systems and Control 7 (3) (2020) 253–383. doi:10.1561/2600000020.

[13] X. Meng, Q. Liu, A consensus algorithm based on multi-agent system with state noise and gradient disturbance for distributed convex optimization, Neuro computing 519 (2023) 148–157. doi:https://doi.org/10.1016/j.neucom.2022.11.051.

[14] N. S. Aybat, E. Y. Hamedani, A distributed admm-like method for resource sharing over time-varying networks, SIAM Journal on Optimization 29 (4) (2019) 3036–3068. doi:10.1137/17M1151973.
URL https://doi.org/10.1137/17M1151973

[15] L. Carlone, V. Srivastava, F. Bullo, G. C. Calafiore, Distributed random convex programming via constraints consensus, SIAM Journal on Control and Optimization 52 (1) (2014) 629–662. doi:10.1137/120885796.
URL https://doi.org/10.1137/120885796

[16] N. S. Aybat, H. Ahmadi, U. V. Shanbhag, On the analysis of inexact augmented lagrangian schemes for misspecified conic convex programs, IEEE Transactions on Automatic Control 67 (8) (2021) 3981–3996.

[17] D. Bertsimas, D. B. Brown, C. Caramanis, Theory and applications of robust optimization, SIAM review 53 (3) (2011) 464–501.

[18] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, Robust optimization, Princeton university press, 2009.

[19] D. Bertsimas, V. Gupta, N. Kallus, Data-driven robust optimization, Mathematical Programming 167 (2018) 235–292.

[20] C. Jie, L. Prashanth, M. Fu, S. Marcus, C. Szepesvári, Stochastic optimization in a cumulative prospect theory framework, IEEE Transactions on Automatic Control 63 (9) (2018) 2867–2882.

[21] A. Shapiro, D. Dentcheva, A. Ruszczynski, Lectures on stochastic programming: modeling and theory, SIAM, 2021.

[22] C. Wilson, V. V. Veeravalli, A. Nedić, Adaptive sequential stochastic optimization, IEEE Transactions on Automatic Control 64 (2) (2018) 496–509.
[23] H. Jiang, U. V. Shanbhag, On the solution of stochastic optimization and variational problems in imperfect information regimes, SIAM Journal on Optimization 26 (4) (2016) 2394–2429.
[24] N. Ho-Nguyen, F. Kılınç-Karzan, Exploiting problem structure in optimization under uncertainty via online convex optimization, Mathematical Programming 177 (1-2) (2018) 113–147. doi:10.1007/s10107-018-1262-8.
[25] H. Ahmadi, U. V. Shanbhag, On the resolution of misspecified convex optimization and monotone variational inequality problems, Computational Optimization and Applications 77 (1) (2020) 125–161.
[26] N. Liu, L. Guo, Stochastic adaptive linear quadratic differential games, arXiv preprint arXiv:2204.08869 (2022).
[27] A. Kannan, A. Nedić, U. V. Shanbhag, Distributed stochastic optimization under imperfect information, in: 2015 54th IEEE Conference on Decision and Control (CDC), IEEE, 2015, pp. 400–405.
[28] I. Notarnicola, A. Simonetto, F. Farina, G. Notarstefano, Distributed personalized gradient tracking with convex parametric models, IEEE Transactions on Automatic Control 68 (1) (2023) 588–595. doi:10.1109/TAC.2022.3147007.
[29] J. Du, Y. Liu, Y. Zhi, H. Gao, Computational convergence rate analysis of distributed optimization algorithm, in: International Conference on Guidance, Navigation and Control, Springer, 2022, pp. 5288–5299.
[30] S. Pu, A. Olshevsky, I. C. Paschalidis, A sharp estimate on the transient time of distributed stochastic gradient descent, IEEE Transactions on Automatic Control 67 (11) (2021) 5900–5915.
[31] S. Liang, L. Wang, G. Yin, Distributed quasi-monotone subgradient algorithm for nonsmooth convex optimization over directed graphs, Automatica 101 (2019) 175–181.
[32] L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Review 60 (2) (2018) 223–311. doi:10.1137/16M1080173.
URL https://doi.org/10.1137/16M1080173
[33] L. Xiao, S. Boyd, Fast linear iterations for distributed averaging, Systems & Control Letters 53 (1) (2004) 65–78.
[34] G. Qu, N. Li, Harnessing smoothness to accelerate distributed optimization, IEEE Transactions on Control of Network Systems 5 (3) (2017) 1245–1260.
[35] F. Bullo, Lectures on Network Systems, 1.6 Edition, Kindle Direct Publishing, 2022.
URL http://motion.me.ucsb.edu/book-lns
[36] B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, W. Yin, Exponential graph is provably efficient for decentralized deep training, Advances in Neural Information Processing Systems 34 (2021) 13975–13987.
[37] B. Ying, K. Yuan, H. Hu, Y. Chen, W. Yin, Bluefog: Make decentralized algorithms practical for optimization and deep learning, arXiv preprint arXiv:2111.04287 (2021).

Appendix A. Proof of Lemma 3.1.1

Proof By using Assumption 2.2.2, we obtain that

$$\begin{aligned}
&\mathbb{E}[\|\bar{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2|\mathcal{F}(k)] \\
&= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i(x_i(k),\theta_i(k),\xi_i(k)) - \frac{1}{n}\sum_{i=1}^n \nabla_x f_i(x_i(k),\theta_i(k))\right\|^2|\mathcal{F}(k)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|g_i(x_i(k),\theta_i(k),\xi_i(k)) - \nabla_x f_i(x_i(k),\theta_i(k))\|^2|\mathcal{F}(k)\right] \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \left(\sigma_x^2 + M_x\|\nabla_x f_i(x_i(k),\theta_i(k))\|^2\right) \leq \frac{\sigma_x^2}{n} + \frac{M_x\sum_{i=1}^n \|\nabla_x f_i(x_i(k),\theta_i(k))\|^2}{n^2},
\end{aligned} \tag{A.1}$$

where the second equality use the fact that $\xi_i, \forall i$ are independent random variables. By recalling assumption 2.2.1, we achieve

$$\begin{aligned}
\|\nabla_x f_i(x_i(k),\theta_i(k))\|^2 &= \|\nabla_x f_i(x_i(k),\theta_i(k)) - \nabla_x f_i(x_*,\theta_i(k)) \\
&\quad + \nabla_x f_i(x_*,\theta_i(k)) - \nabla_x f_i(x_*,\theta_*) + \nabla_x f_i(x_*,\theta_*)\|^2 \\
&\leq 3\|\nabla_x f_i(x_i(k),\theta_i(k)) - \nabla_x f_i(x_*,\theta_i(k))\|^2 + 3\|\nabla_x f_i(x_*,\theta_i(k)) \\
&\quad - \nabla_x f_i(x_*,\theta_*)\|^2 + 3\|\nabla_x f_i(x_*,\theta_*)\|^2 \\
&\leq 3L_x^2\|x_i(k) - x_*\|^2 + 3L_\theta^2\|\theta_i(k) - \theta_*\|^2 + 3\|\nabla_x f_i(x_*,\theta_*)\|^2.
\end{aligned} \tag{A.2}$$

Combining (A.2) and (A.1) yields the result (16). □

Appendix B. Proof of Lemma 3.1.2

Proof By recalling the definition of $\bar{x}(k), \bar{\theta}(k)$ and $\bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))$ in (13) and (15), using Assumption 2.2.1, we have

$$\begin{aligned}
&\|\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\| \\
=&\|\frac{1}{n}\sum\nolimits_{i=1}^{n} \nabla_x f_i(\bar{x}(k), \bar{\theta}(k)) - \frac{1}{n}\sum\nolimits_{i=1}^{n} \nabla_x f_i(x_i(k), \theta_i(k))\| \\
\leq&\frac{1}{n}\sum\nolimits_{i=1}^{n} \|\nabla_x f_i(\bar{x}(k), \bar{\theta}(k)) - \nabla_x f_i(x_i(k), \theta_i(k))\| \\
=&\frac{1}{n}\sum\nolimits_{i=1}^{n} \|\nabla_x f_i(\bar{x}(k), \bar{\theta}(k)) - \nabla_x f_i(x_i(k), \bar{\theta}(k)) + \nabla_x f_i(x_i(k), \bar{\theta}(k)) - \nabla_x f_i(x_i(k), \theta_i(k))\| \\
\leq&\frac{1}{n}\sum_{i=1}^{n} \Big[\|\nabla_x f_i(\bar{x}(k), \bar{\theta}(k)) - \nabla_x f_i(x_i(k), \bar{\theta}(k))\| + \|\nabla_x f_i(x_i(k), \bar{\theta}(k)) - \nabla_x f_i(x_i(k), \theta_i(k))\|\Big] \\
\leq&\frac{1}{n}\Big(L_x \sum\nolimits_{i=1}^{n} \|\bar{x}(k) - x_i(k)\| + L_\theta \sum\nolimits_{i=1}^{n} \|\bar{\theta}(k) - \theta_i(k)\|\Big) \\
\leq&\frac{L_x}{\sqrt{n}}\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)^T\| + \frac{L_\theta}{\sqrt{n}}\|\boldsymbol{\theta}(k) - \mathbf{1}\bar{\theta}(k)^T\|,
\end{aligned}$$

where the last relation follows from Cauchy-Schwarz inequality. □

Appendix C. Proof of Lemma 3.2.1

Proof According to the definitions of $\bar{x}(k)$ and $\bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k))$ in (13) and (14), together with $\sum_{i=1}^{n} w_{ij} = 1$ form Assumption 2.2.3, we have

$$\begin{aligned}
\bar{x}(k+1) &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\Big(\sum\nolimits_{j=1}^{n} w_{ij}\Big(x_j(k) - \alpha_k g_j(x_j(k), \theta_j(k), \xi_j(k))\Big)\Big) \\
&= \frac{1}{n}\sum_{j=1}^{n} x_j(k) - \alpha_k \cdot \frac{1}{n}\sum_{j=1}^{n} g_j(x_j(k), \theta_j(k), \xi_j(k)) = \bar{x}(k) - \alpha_k \bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k)).
\end{aligned} \tag{C.1}$$

Thus,

$$\begin{aligned}
&\|\bar{x}(k+1) - x_*\|^2 = \|\bar{x}(k) - \alpha_k \bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k)) - x_*\|^2 \\
&= \|\bar{x}(k) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - x_* + \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - \alpha_k \bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k))\|^2 \\
&= \|\bar{x}(k) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - x_*\|^2 + \alpha_k^2 \|\bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - \bar{g}(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k))\|^2 \\
&+ 2\alpha_k (\bar{x}(k) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - x_*)^T \left(\bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - \bar{g}\left(\mathbf{x}(k), \boldsymbol{\theta}(k), \boldsymbol{\xi}(k)\right)\right).
\end{aligned}$$

In light of Assumption 2.2.2 and Lemma 3.1.1, by taking conditional expectation on both sides of above equation, we have

$$\begin{aligned}
\mathbb{E}[\|\bar{x}(k+1) - x_*\|^2 | \mathcal{F}(k)]\| \leq& \|\bar{x}(k) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - x_*\|^2 \\
&+ \alpha_k^2 \left(\frac{3M_x L_x^2}{n^2}\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2 + \frac{3M_x L_\theta^2}{n^2}\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2 + \frac{\bar{M}}{n}\right).
\end{aligned} \tag{C.2}$$

Next, we bound the first term on the right side of (C.2).

$$\begin{aligned}
&\|\bar{x}(k) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k)) - x_*\|^2 \\
&= \|\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k); \bar{\theta}(k)) - x_* + \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \alpha_k \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\|^2 \\
&= \|\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - x_*\|^2 + \alpha_k^2 \|\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\|^2 \\
&+ 2\alpha_k (\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - x_*)^T (\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))) \\
&\leq \underbrace{\|\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - x_*\|^2}_{\text{Term 1}} + \underbrace{\alpha_k^2 \|\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\|^2}_{\text{Term 2}} \\
&+ \underbrace{2\alpha_k \|\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - x_*\| \times \|\nabla_x f(\bar{x}(k), \bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k), \boldsymbol{\theta}(k))\|}_{\text{Term 3}}.
\end{aligned} \tag{C.3}$$

As for Term 1, by leveraging the fact that $\alpha_k \leq \frac{1}{L}$, Lemma 2.2.1 indicates

$$\|\bar{x}(k) - \alpha_k \nabla_x f(\bar{x}(k), \bar{\theta}(k)) - x_*\|^2 \leq (1 - \alpha_k \mu_x)^2 \|\bar{x}(k) - x_*\|^2. \tag{C.4}$$

By Lemma 3.1.2, Term 2 can be bounded as follow

$$\alpha_k^2\|\nabla_x f(\bar{x}(k),\bar{\theta}(k)) - \bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \le \left(\frac{\alpha_k L_x\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)^T\|}{\sqrt{n}} + \frac{\alpha_k L_\theta\|\boldsymbol{\theta}(k)-\mathbf{1}\bar{\theta}(k)^T\|}{\sqrt{n}}\right)^2$$
$$\le \frac{\alpha_k^2 L_x^2\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)^T\|^2}{n} + \frac{\alpha_k^2 L_\theta^2\|\boldsymbol{\theta}(k)-\mathbf{1}\bar{\theta}(k)^T\|^2}{n} + \frac{2L_xL_\theta\alpha_k^2}{n}\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)^T\|\times\|\boldsymbol{\theta}(k)-\mathbf{1}\bar{\theta}(k)^T\|. \tag{C.5}$$

Finally, Term 3 can be bounded by invoking the same transformation approches used in Term 1 and Term 2:

$$\text{Term 3} \le 2\alpha_k(1-\alpha_k\mu_x)\|\bar{x}(k)-x_*\|\left(\frac{L_x}{\sqrt{n}}\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)^T\|\frac{L_\theta}{\sqrt{n}}\|\boldsymbol{\theta}(k)-\mathbf{1}\bar{\theta}(k)^T\|\right)$$
$$\le \frac{2\alpha_k L_x(1-\alpha_k\mu_x)\|\bar{x}(k)-x_*\|\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)^T\|}{\sqrt{n}} + \frac{2\alpha_k L_\theta(1-\alpha_k\mu_x)\|\bar{x}(k)-x_*\|\|\boldsymbol{\theta}(k)-\mathbf{1}\bar{\theta}(k)^T\|}{\sqrt{n}}. \tag{C.6}$$

In light of relation (C.3)~(C.6), taking full expectation on both side of relation (C.2) yields the result (**A**). Furthermore by using mean value inequality $2ab \le a^2+b^2$, we rearrange (21) and obtain that

$$\begin{aligned}
U_1(k+1) &\le (1-\alpha_k\mu_x)^2U_1(k) + \frac{\alpha_k^2L_x^2}{n}V_1(k) + \frac{\alpha_k^2L_\theta^2}{n}V_2(k) + \frac{\alpha_k^2L_x^2}{n}V_1(k) + \frac{\alpha_k^2L_\theta^2}{n}V_2(k)\\
&\quad + (1-\alpha_k\mu_x)^2c_1U_1(k) + \frac{\alpha_k^2L_x^2}{n}\cdot\frac{1}{c_1}V_1(k) + (1-\alpha_k\mu_x)^2c_2U_1(k) + \frac{\alpha_k^2L_\theta^2}{n}\cdot\frac{1}{c_2}V_2(k)\\
&\quad + \alpha_k^2\left(\frac{3M_xL_x^2}{n^2}\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}x_*^T\|] + \frac{3M_xL_\theta^2}{n^2}\mathbb{E}[\|\boldsymbol{\theta}(k)-\mathbf{1}\theta_*^T\|] + \frac{\bar{M}}{n}\right)\\
&\le (1+c_1+c_2)(1-\alpha_k\mu_x)^2U_1(k) + (2+\frac{1}{c_1})\frac{\alpha_k^2L_x^2}{n}V_1(k) + (2+\frac{1}{c_2})\frac{\alpha_k^2L_\theta^2}{n}V_2(k)\\
&\quad + \alpha_k^2\left(\frac{3M_xL_x^2}{n^2}\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}x_*^T\|] + \frac{3M_xL_\theta^2}{n^2}\mathbb{E}[\|\boldsymbol{\theta}(k)-\mathbf{1}\theta_*^T\|] + \frac{\bar{M}}{n}\right),
\end{aligned} \tag{C.7}$$

where $c_1, c_2 > 0$. Take $c_1 = c_2 = \frac{3}{16}\alpha_k\mu_x$, then $c_1+c_2 = \frac{3}{8}\alpha_k\mu_x$. Noticing that $\alpha_k \le \frac{1}{3\mu_x}$, i.e. $\alpha_k\mu_x \le \frac{1}{3}$, we have

$$\begin{aligned}
(1+c_1+c_2)(1-\alpha_k\mu_x)^2 &= 1 - \frac{13}{8}\alpha_k\mu_x + \frac{1}{4}\alpha_k^2\mu_x^2 + \frac{3}{8}\alpha_k^3\mu_x^3\\
&\le 1 - \frac{13}{8}\alpha_k\mu_x + \frac{1}{12}\alpha_k\mu_x + \frac{3}{8}\times\frac{1}{9}\alpha_k\mu_x = 1 - \frac{3}{2}\alpha_k\mu_x,
\end{aligned} \tag{C.8}$$

and $(2+\frac{1}{c_m})\alpha_k \le \frac{6}{\mu_x}, m = 1, 2$. Plug them into (C.7) yeilds the result **B**. □

Appendix D. Proof of Lemma 3.2.2

Proof Recalling the definition of $\mathbf{g}(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\xi})$ in (8) and relation $\bar{x}(k+1) = \bar{x}(k) - \alpha_k\bar{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k))$ in eq. (C.1), and using (10), we have

$$\begin{aligned}
\mathbf{x}(k+1) - \mathbf{1}\bar{x}(k+1) &= W\left(\mathbf{x}(k) - \alpha_k\mathbf{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\right) - \mathbf{1}\left(\bar{x}(k) - \alpha_k\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\right)\\
&= (W - \frac{\mathbf{1}\mathbf{1}^T}{n})\left[(\mathbf{x}(k)-\mathbf{1}\bar{x}(k)) - \alpha_k(\mathbf{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right))\right].
\end{aligned} \tag{D.1}$$

Thus by Lemma 2.2.2, we obtain

$$\begin{aligned}
&\|\mathbf{x}(k+1)-\mathbf{1}\bar{x}(k+1)\|^2\\
&\quad\le \rho_w^2\|(\mathbf{x}(k)-\mathbf{1}\bar{x}(k)) - \alpha_k(\mathbf{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\|^2\\
&\quad= \rho_w^2[\|\mathbf{x}(k)-\mathbf{1}\bar{x}(k)\|^2 + \underbrace{\alpha_k^2\|\mathbf{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\|^2}_{\text{Term 4}}\\
&\qquad\underbrace{-2\alpha_k(\mathbf{x}(k)-\mathbf{1}\bar{x}(k))^T(\mathbf{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)]}_{\text{Term 5}}
\end{aligned} \tag{D.2}$$

In the following, we will separately consider Term 4 and Term 5. Note that

$$\|I - \mathbf{1}\mathbf{1}^T/n\| \le 1. \tag{D.3}$$

The by using Assumptions 2.2.2 (a) and 2.2.2 (b), we derive

$$\begin{aligned}
&\mathbb{E}\left[\alpha_k^2 \ \|\mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\|^2|\mathcal{F}(k)\right] \\
&\quad = \alpha_k^2\mathbb{E}\Big[\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) \\
&\quad +\mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k)) + \mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k)),\boldsymbol{\xi}(k) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\|^2|\mathcal{F}(k)\Big] \\
&\quad = \alpha_k^2\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 + \alpha_k^2\mathbb{E}[\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) \\
&\qquad - \mathbf{g}((\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \mathbf{1}(\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \bar{g}(\mathbf{x}(k),\boldsymbol{\theta}(k)))\|^2|\mathcal{F}(k)] \\
&\quad \overset{(D.3)}{\leq} \alpha_k^2\Big[\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\qquad +\mathbb{E}[\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k))\|^2|\mathcal{F}(k)]\Big] \\
&\quad \leq \alpha_k^2\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 + \alpha_k^2 n\sigma_x^2 + \alpha_k^2 M_x\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2. \qquad \text{(D.4)}
\end{aligned}$$

Recalling Assumption 2.2.2 (a), we obtain that

$$\begin{aligned}
&\mathbb{E}[-2\alpha_k(\mathbf{x}(k) - \mathbf{1}\bar{x}(k))^T\left(\mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \mathbf{1}\bar{g}\left(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)\right)\right)|\mathcal{F}(k)] \\
&\quad = -2\alpha_k(\mathbf{x}(k) - \mathbf{1}\bar{x}(k))^T\left(\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\right), \qquad \text{(D.5)}
\end{aligned}$$

In light of (D.2),(D.4) and (D.5), we have

$$\begin{aligned}
&\frac{1}{\rho_w^2}\mathbb{E}[\|\mathbf{x}(k+1) - \mathbf{1}\bar{x}(k+1)\|^2|\mathcal{F}(k)] \\
&\quad \leq \|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + \alpha_k^2\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\qquad + \alpha_k^2 n\sigma_x^2 + \alpha_k^2 M_x\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\qquad - 2\alpha_k(\mathbf{x}(k) - \mathbf{1}\bar{x}(k))^T(\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))) \\
&\quad \leq \|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + \alpha_k^2\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\qquad + \alpha_k^2 n\sigma_x^2 + \alpha_k^2 M_x\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\qquad + c_3\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + \frac{1}{c_3}\alpha_k^2\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \mathbf{1}\bar{\nabla}_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 \\
&\quad \leq (1+c_3)\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + \alpha_k^2 n\sigma_x^2 + \alpha_k^2\left(1 + M_x + \frac{1}{c_3}\right)\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2, \qquad \text{(D.6)}
\end{aligned}$$

where $c_3 > 0$ is arbitrary, and the last inequality also uses the porperty in (D.3).

We then consider the upper bound of $\|\nabla_x F(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2$ as follow,

$$\begin{aligned}
\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k))\|^2 &= \|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\boldsymbol{\theta}(k)) \\
&\qquad + \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\boldsymbol{\theta}(k)) - \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T) + \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2 \\
&\leq 3\|\nabla_x \boldsymbol{F}(\mathbf{x}(k),\boldsymbol{\theta}(k)) - \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\boldsymbol{\theta}(k))\|^2 \\
&\quad + 3\|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\boldsymbol{\theta}(k)) - \nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2 + 3\|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2 \\
&\leq 3L_x^2\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + 3L_\theta^2\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2 + 3\|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2. \qquad \text{(D.7)}
\end{aligned}$$

Let $c_3 = \frac{1-\rho_w^2}{2}$. Combining (D.6) and (D.7), we obtain that

$$\begin{aligned}
&\frac{1}{\rho_w^2}\mathbb{E}[\|\mathbf{x}(k+1) - \mathbf{1}\bar{x}(k+1)\|^2|\mathcal{F}(k)] \\
&\qquad \leq \frac{3-\rho_w^2}{2}\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 + 3\alpha_k^2\left(\frac{3}{1-\rho_w^2} + M_x\right)\Big(L_x^2\|\mathbf{x}(k) - \mathbf{1}\bar{x}(k)\|^2 \\
&\qquad\quad +L_\theta^2\|\boldsymbol{\theta}(k) - \mathbf{1}\theta_*^T\|^2 + \|\nabla_x \boldsymbol{F}(\mathbf{1}x_*^T,\mathbf{1}\theta_*^T)\|^2\Big) + \alpha_k^2 n\sigma_x^2. \qquad \text{(D.8)}
\end{aligned}$$

Note that $\rho_w^2(\frac{3-\rho_w^2}{2}) \leq \frac{3+\rho_w^2}{4}$ by $\rho_w \in (0,1)$. Then by taking full expectation on both sides of (D.8) and multiplying ρ_w^2 leads to the result (24). □

Appendix E. Proof of Lemma 3.3.1

Proof For any $k \geq 0$, in order to bound $\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2]$, we firstly consider bounding $\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k),\theta_i(k),\xi_i(k)) -$

$x_*\|^2]$ for all $i \in \mathcal{N}$. By using Assumption 2.2.1 (i) and Assumption 2.2.2 (c), we have

$$\begin{aligned}
\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k), \theta_i(k), \xi_i(k)) - x_*\|^2 | \mathcal{F}(k)] &= \|x_i(k) - x_* - \alpha_k \nabla_x f_i(x_i(k), \theta_i(k))\|^2 \\
&\quad + \alpha_k^2 \mathbb{E}\left[\|\nabla_x f_i(x_i(k), \theta_i(k)) - g_i(x_i(k), \theta_i(k), \xi_i(k))\|^2 | \mathcal{F}(k)\right] \\
&\leq \|x_i(k) - x_*\|^2 - 2\alpha_k \nabla_x f_i(x_i(k), \theta_i(k))^T (x_i(k) - x_*) \\
&\quad + \alpha_k^2 \|\nabla_x f_i(x_i(k), \theta_i(k))\|^2 + \alpha_k^2 (\sigma_x^2 + M_x \|\nabla_x f_i(x_i(k), \theta_i(k))\|^2) \\
&\leq \|x_i(k) - x_*\|^2 - 2\alpha_k \mu_x \|x_i(k) - x_*\|^2 + 2\alpha_k \|\nabla_x f_i(x_*, \theta_i(k))\| \|x_i(k) - x_*\| \\
&\quad + \alpha_k^2 (1 + M_x) \|\nabla_x f_i(x_i(k), \theta_i(k))\|^2 + \alpha_k^2 \sigma_x^2,
\end{aligned} \tag{E.1}$$

Consider the upper bound of the term $\|\nabla_x f_i(x_i(k), \theta_i(k))\|^2$ on the right side of above inequality. Using Assumption 2.2.1 (i) and (ii), we have

$$\begin{aligned}
\|\nabla_x f_i(x_i(k), \theta_i(k))\|^2 &= \|\nabla_x f_i(x_i(k), \theta_i(k)) - \nabla_x f_i(x_*, \theta_i(k)) + \nabla_x f_i(x_*, \theta_i(k)) \\
&\quad - \nabla_x f_i(x_*, \theta_*) + \nabla_x f_i(x_*, \theta_*)\| \\
&\leq 3L_x^2 \|x_i(k) - x_*\|^2 + 3L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 3\|\nabla_x f_i(x_*, \theta_*)\|^2.
\end{aligned} \tag{E.2}$$

We can similarly obtain $\|\nabla_x f_i(x_*, \theta_i(k))\|^2 \leq 2L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 2\|\nabla_x f_i(x_*, \theta_*)\|^2$. Combining (E.2) and (E.1), it produces

$$\begin{aligned}
\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k), \theta_i(k), \xi_i(k)) - x_*\|^2 | \mathcal{F}(k)] &\leq \|x_i(k) - x_*\|^2 - 2\alpha_k \mu_x \|x_i(k) - x_*\|^2 \\
&\quad + \alpha_k^2 \sigma_x^2 + 2\alpha_k \sqrt{2L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 2\|\nabla_x f_i(x_*, \theta_*)\|^2} \|x_i(k) - x_*\| \\
&\quad + \alpha_k^2 (1 + M_x)(3L_x^2 \|x_i(k) - x_*\|^2 + 3L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 3\|\nabla_x f_i(x_*, \theta_*)\|^2) \\
&\leq (1 - 2\alpha_k \mu_x + 3\alpha_k^2 (1 + M_x) L_x^2) \|x_i(k) - x_*\|^2 \\
&\quad + 2\alpha_k \sqrt{2L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 2\|\nabla_x f_i(x_*, \theta_*)\|^2} \|x_i(k) - x_*\| \\
&\quad + \alpha_k^2 [3(1 + M_x) L_\theta^2 \|\theta_i(k) - \theta_*\|^2 + 3(1 + M_x)\|\nabla_x f_i(x_*, \theta_*)\|^2 + \sigma_x^2].
\end{aligned} \tag{E.3}$$

From the definition of K in (26), for all $k \geq 0$, we have $\alpha_k \leq \frac{\mu_x}{3(1+M_x)L_x^2}$. Recall the fact that $\mathbb{E}[\|\theta_i(k) - \theta_*\|^2] \leq \hat{\Theta}_i$ in (27). By taking full expectation on both sides of (E.3) and using $\mathbb{E}[\|x_i(k) - x_*\|] \leq \sqrt{\mathbb{E}[\|x_i(k) - x_*\|^2]}$, we have

$$\begin{aligned}
\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k), \theta_i(k), \xi_i(k)) - x_*\|^2] &\leq (1 - \alpha_k \mu_x) \mathbb{E}[\|x_i(k) - x_*\|^2] \\
&\quad + 2\alpha_k \sqrt{2L_\theta^2 \hat{\Theta}_i + 2\|\nabla_x f_i(x_*, \theta_*)\|^2} \sqrt{\mathbb{E}[\|x_i(k) - x_*\|^2]} \\
&\quad + \alpha_k \left[\frac{\mu_x L_\theta^2}{L_x^2} \hat{\Theta}_i + \frac{\mu_x}{L_x^2} \|\nabla_x f_i(x_*, \theta_*)\|^2 + \frac{\mu_x \sigma_x^2}{3(1 + M_x) L_x^2}\right] \\
&\leq \mathbb{E}[\|x_i(k) - x_*\|^2] - \alpha_k \Big[\mu_x \mathbb{E}\|x_i(k) - x_*\|^2 \\
&\quad - 2\sqrt{2L_\theta^2 \hat{\Theta}_i + 2\|\nabla_x f_i(x_*, \theta_*)\|^2} \sqrt{\mathbb{E}[\|x_i(k) - x_*\|^2]} \\
&\quad - \left(\frac{\mu_x L_\theta^2}{L_x^2} \hat{\Theta}_i + \frac{\mu_x}{L_x^2} \|\nabla_x f_i(x_*, \theta_*)\|^2 + \frac{\mu_x \sigma_x^2}{3(1 + M_x) L_x^2}\right)\Big].
\end{aligned} \tag{E.4}$$

Next, we consider the following set:

$$\begin{aligned}
\mathcal{X}_i \triangleq \Big\{ q \geq 0 : &\mu_x q - 2\sqrt{2L_\theta^2 \hat{\Theta}_i + 2\|\nabla_x f_i(x_*, \theta_*)\|^2} \sqrt{q} \\
&- \frac{\mu_x}{3L_x^2} \left(3L_\theta^2 \hat{\Theta}_i + 3\|\nabla_x f_i(x_*, \theta_*)\|^2 + \frac{\sigma_x^2}{1 + M_x}\right) \leq 0\Big\}.
\end{aligned} \tag{E.5}$$

It can be seen that $\mathcal{X}_i$ is non-empty and compact. If $\mathbb{E}[\|x_i(k) - x_*\|^2] \notin \mathcal{X}_i$, in light of (E.4) we known that $\mathbb{E}[\|x_i(k) -$

$\alpha_k g_i(x_i(k),\theta_i(k),\xi_i(k)) - x_*\|^2] \le \mathbb{E}[\|x_i(k)-x_*\|^2]$. While for $\mathbb{E}[\|x_i(k)-x_*\|^2] \in \mathcal{X}_i$, by using $\alpha_k \le \frac{\mu_x}{3(1+M_x)L_x^2}$, we derive

$$\begin{aligned}
&\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k),\theta_i(k),\xi_i(k)) - x_*\|^2] \\
&\le \max_{q\in\mathcal{X}_i}\Bigg\{q - \frac{\mu_x}{3(1+M_x)L_x^2}\Big[\mu_x q - 2\sqrt{2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*)\|^2}\,\sqrt{q} \\
&\quad - \frac{\mu_x}{3L_x^2}\left(3L_\theta^2\hat{\Theta}_i + 3\|\nabla_x f_i(x_*,\theta_*)\|^2 + \frac{\sigma_x^2}{1+M_x}\right)\Big]\Bigg\} \triangleq R_i. \qquad \text{(E.6)}
\end{aligned}$$

Based on previous arguments, we conclude that for all $k > 0$,

$$\mathbb{E}[\|x_i(k) - \alpha_k g_i(x_i(k),\theta_i(k),\xi_i(k)) - x_*\|^2] \le \max\left\{\mathbb{E}[\|x_i(k)-x_*\|^2], R_i\right\}. \qquad \text{(E.7)}$$

In light of $W\mathbf{1} = \mathbf{1}$, by noting from (10) that

$$\begin{aligned}
\|\mathbf{x}(k+1) - \mathbf{1}x_*^T\|^2 &\le \|W\|^2\|\mathbf{x}(k) - \alpha_k\mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \mathbf{1}x_*^T\|^2 \\
&\le \|\mathbf{x}(k) - \alpha_k\mathbf{g}(\mathbf{x}(k),\boldsymbol{\theta}(k),\boldsymbol{\xi}(k)) - \mathbf{1}x_*^T\|^2. \qquad \text{(E.8)}
\end{aligned}$$

This together with (E.7) produces

$$\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^T\|^2] \le \max\left\{\mathbb{E}[\|\mathbf{x}(0) - \mathbf{1}x_*^T\|^2], \sum_{i=1}^n R_i\right\} \qquad \text{(E.9)}$$

In the following, we will give an upper bound of R_i. From the definition of $\mathcal{X}_i$ in (E.5), we know that the right zero of the upward opening parabola is

$$\begin{aligned}
\sqrt{q_i} = \frac{1}{2\mu_x}\Bigg[&2\sqrt{2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*)\|^2} \\
&+ \sqrt{4(2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*)\|^2) + \frac{4\mu_x^2}{3L_x^2}\left(3L_\theta^2\hat{\Theta}_i + 3\|\nabla_x f_i(x_*,\theta_*)\|^2 + \frac{\sigma_x^2}{1+M_x}\right)}\Bigg].
\end{aligned}$$

Then by using $\mu_x \le L_x$, we achieve

$$\begin{aligned}
q_i &\le \frac{1}{4\mu_x^2}\Bigg[2\times 4(2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*)\|^2) \\
&\quad + 2\left(8L_\theta^2\hat{\Theta}_i + 8\|\nabla_x f_i(x_*,\theta_*)\|^2 + 4L_\theta^2\hat{\Theta}_i + 4\|\nabla_x f_i(x_*,\theta_*)\|^2 + \frac{4\mu_x^2\sigma_x^2}{3L_x^2(1+M_x)}\right)\Bigg] \\
&\le \frac{10L_\theta^2\hat{\Theta}_i}{\mu_x^2} + \frac{10\|\nabla_x f_i(x_*,\theta_*)\|^2}{\mu_x^2} + \frac{2\sigma_x^2}{3(1+M_x)L_x^2} \triangleq q_i^*.
\end{aligned}$$

Thus, $\mathcal{X}_i = [0, q_i] \subset [0, q_i^*]$. Hence from (E.6) it follows that

$$\begin{aligned}
R_i &\le q_i^* - \frac{\mu_x}{3(1+M_x)L_x^2}\Bigg[\mu_x q - 2\sqrt{2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*)\|^2}\,\sqrt{q} \\
&\quad - \frac{\mu_x}{3L_x^2}\left(3L_\theta^2\hat{\Theta}_i + 3\|\nabla_x f_i(x_*,\theta_*)\|^2 + \frac{\sigma_x^2}{1+M_x}\right)\Bigg]\Bigg|_{q=\frac{\sqrt{2L_\theta^2\hat{\Theta}_i+2\|\nabla_x f_i(x_*,\theta_*)\|^2}}{\mu_x}} \\
&\le \frac{10L_\theta^2\hat{\Theta}_i}{\mu_x^2} + \frac{10\|\nabla_x f_i(x_*,\theta_*)\|^2}{\mu_x^2} + \frac{2\sigma_x^2}{3(1+M_x)L_x^2} \\
&\quad + \frac{\mu_x}{3(1+M_x)L_x^2}\Bigg[\frac{\mu_x}{3L_x^2}\left(3L_\theta^2\hat{\Theta}_i + 3\|\nabla_x f_i(x_*,\theta_*)\|^2 + \frac{\sigma_x^2}{1+M_x}\right) \\
&\quad + \frac{2L_\theta^2\hat{\Theta}_i + 2\|\nabla_x f_i(x_*,\theta_*\|^2)}{\mu_x}\Bigg] \\
&\le \frac{11L_\theta^2\hat{\Theta}_i}{\mu_x^2} + \frac{11\|\nabla_x f_i(x_*,\theta_*)\|^2}{\mu_x^2} + \frac{7\sigma_x^2}{9(1+M_x)L_x^2}, \qquad \text{(E.10)}
\end{aligned}$$

where the last inequality has used $\mu_x \le L_x$.

Combing (E.10) and (E.9), the lemma holds. □