# ON THE PERFORMANCE OF DISTRIBUTED ALGORITHMS FOR NETWORK OPTIMIZATION PROBLEMS

BY

THINH THANH DOAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Associate Professor Carolyn L. Beck, Chair
Assistant Professor Subhonmesh Bose
Professor Daniel Liberzon
Professor Rayadurgam Srikant

# Abstract

This thesis considers optimization problems defined over a network of nodes, where each node knows only part of the objective functions. We are motivated by broad applications of these problems within engineering and sciences, where problems are characterized by either complex networks with a large number of nodes or massive amounts of data. Algorithms for solving these problems should be implemented in parallel between the nodes, and are based only on local computation and communication, necessitating the development of distributed algorithms.

Our interest, therefore, is to study distributed methods for solving networked optimization problems, where our focus is on distributed gradient algorithms. In particular, we move beyond the existing results to significantly enhance the performance and reduce the complexity of distributed gradient methods, while taking practical issues, such as communication delays and resource uncertainty, into account. Our goal is to bridge the gap between theory and practice, leading to significant improvement in their performance for solving real-world problems.

The remainder of this thesis is to focus on three main thrusts — first, we study the impact of communication delays, an inevitable issue in distributed systems, on the performance of distributed gradient algorithms. Our results address a notable omission in the existing literature, where the delays are often ignored. Second, we study different variants of distributed gradient algorithms, and show that under certain conditions we can improve their convergence. Finally, we study an important problem within engineering and computer science, namely, network resource allocation. For solving this problem, we propose distributed Lagrangian methods and show that our methods are robust to resource uncertainty. In addition, we design a novel algorithm, namely, the distributed gradient balancing protocol, for solving a special case of network resource allocation problems. We show that our algorithm achieves a quadratic convergence time, which improves the convergence of the existing algorithms by a factor of $n$, the size of the network.

*To the memory of my father and to my family, for their love and support.*

*"Be happy with what you have. Work hard for what you want."*

# Acknowledgments

Obtaining a Ph.D. degree in Electrical and Computer Engineering at UIUC is my ultimate goal during the last five years. Along my journey I am proud to be a member of CSL, where I have been very fortunate to interact with many great people. My Ph.D. advisor, Prof. Carolyn Beck, obviously plays a significant role in my success at Illinois. During my time working under her supervision, Carolyn showed me her constant guidance and support for my studies and career. She taught me how to conduct research, how to write a paper, and above all how to shape my career. Whenever I seek advice, she has always been there with the best solution. I am thankful and indebted for her great care and concern about my career and my family.

I thank Prof. R. Srikant, who has been my great collaborator for many important results in this thesis. Srikant has also been very helpful in shaping my research and future career. I have learned so much from our collaboration. I thank Prof. Alex Olshevsky, who helped me to gain a strong background for the work in this thesis, where the last chapter is our collaboration in my first year. I thank Prof. Subhonmesh Bose for our collaboration during my last two years, producing an important result in my research. Bose has also been a great source of advice for my future career. I thank Prof. Daniel Liberzon for giving me lots of advice and support for my time here. I also thank Prof. Subhonmesh Bose, Prof. Daniel Liberzon, and Prof. R. Srikant for serving on my committee.

I thank Prof. Choon-Yik Tang, who was my advisor of my master's degree at the University of Oklahoma. Choon-Yik has continued his support for my career. I thank Prof. Angelia Nedić for her support with many useful discussions related to my research. I thank Prof. Dinh Hoa Nguyen and Joseph Lubars for our collaborations during my last semester here.

One of the best aspects provided to UIUC students is the opportunity to learn various courses taught by excellent instructors. I am very fortunate to have experienced classes with the best teachers, for example, Prof. R. Srikant, Prof. Daniel Liberzon, Prof. Maxim Raginsky, and Prof. Karthik Chandrasekaran, who have inspired me by their care about students' learning.

I thank my friends and my research group at UIUC, especially, Trong Tong, Hoang-Vu Dang, Phuong Minh Cao, Quang Tran, and Philip E. Paré for their support and help through my studies here. Administrative staffs in ECE and CSL have made my time at UIUC much easier. Special thanks to Angie Ellis and Linda Stimson who have always been there whenever I need. I thank Jan Progen for carefully editing my thesis.

I thank NSF and the ECE department for their financial support through my assistantship and fellowship during my Ph.D. studies.

Most importantly, I thank my family for their unconditional love and unlimited support for my career success. I extremely appreciate my mom and my mother-in-law for being here to take care of my kids, so that I would have more time for my studies. I also thank my kids, Lam, Thu, and Tri for the time taken from them. They are always a big motivation to push me forward. I wish to thank my wife, Chi, for being with me and showing her constant support for my career. She has also been very patient with all of my frustrations during my studies.

I wish to show my gratitude and love to my parents for their unlimited support of my career choice. They have worked very hard and given up many things to make my success possible. This thesis is dedicated to them.

# Table of Contents

# Chapter 1

# Scope of Thesis

## 1.1 Introduction

This thesis considers optimization problems defined over a network of nodes, where each node knows only part of the set of objective functions and constraint sets that define the problem of interest. Each node—which may be a sensor, a processor, an electric power generator, a robot, or an autonomous vehicle—has computational capabilities and can communicate with other nodes that are connected to it in the network. We assume that there is no central coordinator between the nodes, requiring them to cooperatively solve the problem. Such optimization problems have received increased interest because of their broad applications within engineering and sciences; a few examples include:

1. *Machine Learning* — A common problem is to find the parameters of statistical models through minimizing empirical loss functions defined over massive amounts of data [1,2]. Due to the explosion in the size of datasets, on the order of terabytes, both the data and computation must be distributed over a network of processors. Therefore, the processors have to perform local computations over their local data, the results of which are then exchanged to arrive at the globally optimal solution.

2. *Estimation over Sensor Networks* — An application of interest is the problem of estimating the radio frequency in a wireless network of sensors [3]. The goal is to cooperatively estimate the radio-frequency power spectral density through solving a regression problem, which is defined over the total data locally measured by the sensors. In this application, the sensors are scattered across a large geographical area, therefore, they are required to share their estimates with other sensors to find the global density.

3. *Networked Resource Allocation* — This problem involves a number of sources, which can send information through a set of communication links [4]; one prominent example is the Internet. Each source has a local utility function defined over the transmission rate assigned to it. The goal is to decide the source rates that maximize the network utilities subject to the link capacity constraints, and reduce packet losses. Such a problem can be solved through considering a networked optimization problem, where each source knows only a local objective function and the link constraints associated with it.

4. *Power Networks* — An important application is solving multi-area economic dispatch (tie-line scheduling) problems in power networks [5], wherein different system operators control parts of an interconnected power network and their associated grid assets. The system operators must coordinate to solve a joint optimal power flow problem to compute the minimum cost dispatch across the entire power network [6–11].

5. *Coverage Control* — In this application, the goal is to optimally allocate a large number of sensors to an unknown environment such that the coverage area is maximized [12, 13]. The sensors coordinate their local positions with other local sensors through a wireless network to determine their optimal locations; for example, finding the centroid of the Voronoi diagram of the coverage area.

These applications are characterized by either complex networks with a large number of nodes or massive amounts of data, where centralized access to information may not be available. This necessitates the development of distributed algorithms, which can be implemented in parallel between the nodes, and are based only on local computation and communication. In addition, the computation and communication in these algorithms should be efficient enough so that the network latencies and communication failures do not offset the computational gains. Our main focus is, therefore, to study distributed algorithms for solving network optimization problems, while taking into account practical considerations.

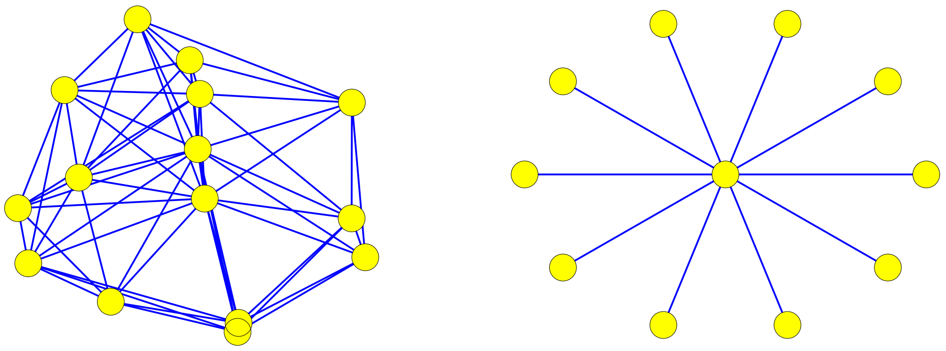

Figure 1.1: The left figure illustrates a general network with 62 edges between 15 nodes, while the right figure illustrates a star network with 10 nodes connected to a central node.

## 1.2 History of Distributed Algorithms

In solving networked optimization problems, there are a variety of distributed algorithms in the literature, which depend on communication structures[1] between the nodes. In this thesis, we focus on two types of communication networks, namely, general networks and star networks. Examples of these two network structures are given in Fig. 1.1. We provide a brief review of existing distributed algorithms associated with these two types of communication networks, which is by no means exhaustive. In addition, we only consider distributed gradient algorithms in this thesis, often referred to as distributed first-order methods because the algorithms make use of the gradients of the objective functions and no higher-order terms.

### General network architectures

In a general network, we consider a peer-to-peer architecture where each node is only connected to a small subset of the other nodes, often referred to as the node's neighbors. Such structure imposes communication constraints on the nodes, that is, each node is only allowed to interact with local nodes. The nodes, however, do not know the network topology. Distributed gradient algorithms on this network structure were originally introduced and studied in the 1980s in the context of parallel and distributed computations [14–16]. Numerous applications of network optimization have motivated a surge of

[1] In this thesis, the terms communication structures, communication topologies, and communication networks are used interchangeably.

interest in distributed gradient algorithms during the past two decades [17–27], where the focus has been on distributed consensus-based methods. In particular, the studies in [17, 18] build on the seminal work in [14], which are the first to provide rigorous analysis for the convergence and convergence rate of such methods. More recent studies [19–21] are focused on improving the convergence rate of distributed gradient algorithms, where the main goal is to obtain the same rate as that attained using standard gradient descent for solving minimization problems in centralized frameworks. We refer to the recent survey paper [28] for a summary about the impact of network topology on the convergence of distributed consensus-based gradient methods.

**Master-worker architectures**

Master-worker architectures are widely used in computer networks, especially in data centers, where there is a server (master) connected to several other processors (workers); this architecture results in a star network toplogy. Distributed algorithms are relatively simple to implement on such architectures, for example, distributed stochastic gradient descent is widely used in the context of machine learning for master-worker architectures. In particular, the master maintains a copy of the model parameter, while the workers store the data defining the objective problems. At each iteration in distributed (stochastic) gradient algorithms, the master sends its current value to the workers, who estimate their local (stochastic) gradients based on this value. The workers then send their (stochastic) gradients to the master to update the master's variable value. Distributed (stochastic) gradient descent has been recognized as an efficient method for data-intensive machine learning problems [29–36], where the focus is to speed up the algorithm through parallelizing the computation of the gradients.

## 1.3 Thesis Outline

The focus of this thesis is to study distributed gradient algorithms for solving networked optimization problems for general communication networks and star networks. In particular, we move beyond the existing results to significantly enhance the performance and reduce the complexity of distributed gradient methods, while taking practical issues, such as communication delays and resource uncertainty, into account. Our goal is to bridge the gap between theory and practice, leading to significant improvement in their performance for solving real-world problems. The main contributions of this thesis are briefly discussed in the following.

1. Chapter 2 provides an introduction for the problems studied in this thesis. We begin by formulating network optimization problems and distributed algorithms. We then review distributed consensus-based methods, which are our main focus in the subsequent chapters.

2. Our first contribution is presented in Chapter 3, where we study the impact of communication delays on the performance of distributed gradient methods. In particular, we provide an explicit formula for the rate of convergence of such methods, as a function of the network topology and delay constants. This result addresses a notable omission in the existing literature, where the delays are often ignored.

3. In Chapter 4, we consider distributed aggregated gradient methods, which have recently been shown to achieve the same convergence rate as the standard centralized gradient descent. Our main contribution is to consider the stochastic variant of such methods and show linear convergence in expectation to the neighborhood of the optimal solution.

4. While distributed gradient methods are known to work with the Euclidean norm, we study distributed mirror descent in Chapter 5, which allows for more general norms. Mirror descent has been shown not only to outperform gradient descent when other norms are considered but also to be applicable to optimization problems formulated in Banach spaces. Our main result is to establish the convergence of the iterates to an optimal solution, which to the best of our knowledge, is not available in the literature. In addition, such convergence is essential in some applications, for example, in our proposed distributed Lagrangian methods in Chapter 7.

5. The focus of Chapter 6 is to study distributed random projection approaches for master-worker architectures (star networks) for solving constrained optimization problems. We show that distributed random projection shares the same convergence rate as distributed stochastic gradient descent, except for a constant factor capturing the regularity condition of the constraint sets.

6. In Chapter 7, we consider network resource allocation problems where we propose distributed Lagrangian methods for solving such problems when the number of resources is uncertain. The key idea of our approach is to utilize distributed gradient methods studied in previous chapters for solving the dual problem.

7. Finally, in Chapter 8 we consider the relaxed resource allocation problem studied in Chapter 7. Our main contribution is to provide a novel distributed algorithm, namely, the distributed gradient balancing protocol, for solving this relaxed problem. In addition, our algorithm achieves a quadratic convergence time, which is an improvement over the existing results by a factor of $n$, the number of nodes in the network.

# Chapter 2

# Distributed Optimization

The focus of this chapter is to provide a foundation for our studies in the subsequent chapters. In particular, we are interested in studying optimization problems defined over a network of nodes. In solving such problems, we are interested in distributed consensus-based methods, a class of truly distributed algorithms. We will present some preliminary results of such methods, which are useful for our studies later.

## 2.1 Problem Statement, Notation, and Assumptions

In network optimization problems, we consider a network of $n$ nodes, where each node has computational capabilities and can send/exchange messages with other nodes. Associated with each node $i$ is a function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, whose sum is the objective problem; see Fig. 2.1 for an example. The goal of the nodes is to solve the following minimization problem:

$$\text{minimize} \sum_{i=1}^{n} f_i(\mathbf{x}) \text{ over } \mathbf{x} \in \mathcal{X}, \tag{2.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a constraint set. Since each node $i$ knows only one function $f_i$, the nodes are required to communicate and cooperatively solve the problem. We assume that there is no central coordination between the nodes and each node is only allowed to interact with a small subset of the nodes, referred to as the node's neighbors. To model this communication structure, we consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the vertex set $\mathcal{V} = \{1, \ldots .n\}$ with the edge set $\mathcal{E} = (\mathcal{V} \times \mathcal{V})$. Here, nodes $i$ and $j$ can communicate with each other if and only if $(i, j) \in \mathcal{E}$. Under this communication constraint, we are interested in distributed algorithms for solving the problem of Eq. (2.1), which are defined as below.

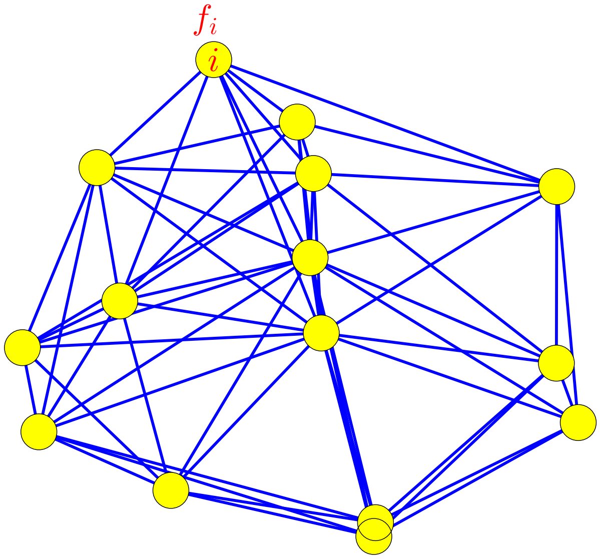

Figure 2.1: A network with 15 nodes, where each node $i$ knows only $f_i$.

**Definition 1.** *A distributed algorithm is an algorithm that is implemented in parallel between the nodes of a graph or a communication network, and is based only on local computation and communication.*

By this definition, the nodes are only allowed to exchange messages with their neighboring nodes that are connected to it based on $\mathcal{G}$. In this thesis, we will focus on studying distributed gradient-based methods, often referred to as distributed first-order methods, for solving problem (2.1). We give a brief introduction and motivation of such methods in Section 2.2. In the next chapters, we study different distributed gradient-based algorithms, which are designed for solving variants of the network optimization problem (2.1). We conclude this section with notation and assumptions frequently used in the remainder of this thesis.

### 2.1.1 Notation

We consider both continuous-time and discrete-time distributed algorithms for solving problem (2.1), where we use $t$ and $k$ to denote continuous and discrete time variables, respectively.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, where $\mathcal{V}$ is the vertex set and $\mathcal{E} = (\mathcal{V} \times \mathcal{V})$ is the edge set. We only consider undirected graphs, meaning that, if $(i, j) \in \mathcal{E}$ then

$(j, i) \in \mathcal{E}$. We denote by $\mathcal{N}_i$ the set of node $i$'s neighbors. In addition, we use $\{\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k))\}$ to denote a time-varying sequence of graphs, where $\mathcal{E}(k)$ is the set of edges at time $k$. Similarly, $\mathcal{N}_i(k)$ denotes the set of node $i$'s neighbors at time $k$.

We use boldface to distinguish between vectors $\mathbf{x}$ in $\mathbb{R}^n$ and scalars $x$ in $\mathbb{R}$. Given any vector $\mathbf{x} \in \mathbb{R}^n$, we write $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and denote by $\bar{x}$ the average of the entries of $\mathbf{x}$, i.e.,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Let $\|\mathbf{x}\|_2$ denote the Euclidean norm of $\mathbf{x} \in \mathbb{R}^n$; we often drop the subscript 2 when it is clear from the context, i.e., $\|\mathbf{x}\|$. Otherwise, when other norms are used we state so explicitly. Given a vector $\mathbf{x}$ and a closed set $\mathcal{X}$ we write the projection of $\mathbf{x}$ on $\mathcal{X}$ as $\mathcal{P}_{\mathcal{X}}[\mathbf{x}]$, i.e.,

$$\mathcal{P}_{\mathcal{X}}[\mathbf{x}] = \arg\min_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|^2. \tag{2.2}$$

We denote by $\mathbf{1}$ and $\mathbf{I}$ a vector whose entries are all 1 and the identity matrix, respectively. Additionally, we use letters in uppercase and boldface to denote matrices, e.g., $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Let $\mathcal{X}^* \subseteq \mathcal{X}$ be the set of optimal solutions to problem (2.1). Finally, we denote $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$ and given a solution $\mathbf{x}^* \in \mathcal{X}^*$ denote the optimal value of (2.1) by

$$f^* = \sum_{i=1}^{n} f_i(\mathbf{x}^*).$$

Our notation is summarized in Table 2.1.

### 2.1.2 Assumptions

We start with some basics of graph theory. For a complete treatment in this area, we refer the readers to the reference [37]. For any fixed graph $\mathcal{G}$, we assume that it is connected.

**Assumption 1.** *$\mathcal{G}$ is connected, i.e., there exists a path between any pair of nodes in $\mathcal{G}$.*

Table 2.1: Notation Table.

| Notation | Meaning |
|---|---|
| $\mathbb{R}$ | set of real numbers |
| $t$ | continuous time |
| $k$ | discrete time |
| $\mathcal{G}$ | undirected graph |
| $\mathcal{V}$ | vertex set |
| $\mathcal{E}$ | set of edges |
| $x$ | real scalar |
| $\mathbf{x}$ | real vector |
| $\mathbf{X}$ | real matrices |
| $\mathbf{1}$ | vector whose entries are 1 |
| $\mathbf{I}$ | identity matrix |
| $\mathcal{X}^*$ | optimal set of (2.1) |
| $f^*$ | optimal value of (2.1) |

On the other hand, we consider the following assumption on the connectivity of time-varying graphs $\mathcal{G}(k)$, which basically states that the sequence of graphs over $k$ can be disconnected at isolated time instants, but are required to satisfy a long-term connectivity.

**Assumption 2.** *There exists an integer $B \geq 1$ such that the graph*

$$(\mathcal{V}, \mathcal{E}(kB) \cup \mathcal{E}(kB+1) \cup \ldots \cup \mathcal{E}((k+1)B-1)) \tag{2.3}$$

*is connected for all non-negative integers $k$.*

We denote by $\mathbf{A}$ the $n \times n$ weighted adjacency matrix corresponding to $\mathcal{G}$, whose $(i, j)$-th entries are $a_{ij}$. We often write $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \ldots \\ \mathbf{a}_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}. \tag{2.4}$$

For fixed graphs, we consider the following assumption on $\mathbf{A}$.

**Assumption 3.** *The weight matrix $\mathbf{A}$ is doubly stochastic, i.e., $\sum_{i=1}^n a_{ij} = \sum_{j=1}^n a_{ij} = 1$. Moreover, $\mathbf{A}$ is assumed to be irreducible and aperiodic. Finally, the weights $a_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$ otherwise $a_{ij} = 0$.*

We note that the assumption on the irreducibility of $\mathbf{A}$ can be satisfied when $\mathcal{G}$ is connected. In addition, the aperiodicity of $\mathbf{A}$ is guaranteed when at least one of its diagonal elements $a_{ii}$ is strictly positive. Similarly, we denote by $\{\mathbf{A}(k)\}$ the sequence of weighted adjacency matrices corresponding to time-varying graphs $\mathcal{G}(k)$, whose $(i,j)$-th entries are $a_{ij}(k)$. In addition, we consider the following assumption when the graph is time-varying.

**Assumption 4.** *There exists a positive constant $\eta$ such that the sequence of matrices $\{\mathbf{A}(k)\}$ satisfies the following conditions:*

1. $a_{ii}(k) \geq \eta$, *for all $i,k$.*

2. $a_{ij}(k) \in [\eta, 1]$ *if $(i,j) \in \mathcal{N}_i(k)$ otherwise $a_{ij}(k) = 0$, for all $i,j,k$.*

3. $\sum_{i=1}^{n} a_{ij}(k) = \sum_{j=1}^{n} a_{ij}(k) = 1$, *for all $i,j,k$.*

Given the weighted adjacency matrix $\mathbf{A}$, we denote by $\sigma_2(\mathbf{A})$ the second largest singular value of $\mathbf{A}$, i.e., $\sigma_2(\mathbf{A})$ is the square root of the second-largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. Since $\mathbf{A}$ satisfies Assumption 3, the Perron-Frobenius theorem [38, Theorem 8.4.4] and the Courant-Fisher theorem [38, Theorem 4.2.11] give

$$\sigma_2(\mathbf{A}) = \max_{\mathbf{x}\neq 0, \mathbf{x}\in\mathbf{1}^{\perp}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \in (0,1). \tag{2.5}$$

In addition, let $\mathbf{L} = \mathbf{I} - \mathbf{A}$ be the corresponding weighted Laplacian matrix of $\mathcal{G}$. We denote by $\lambda_2(\mathbf{L})$ the Fiedler eigenvalue of $\mathbf{L}$ [39], i.e. the second-smallest eigenvalue of $(\mathbf{L}^T+\mathbf{L})/2$, which determines the algebraic connectivity of the graph, similarly defined as,

$$\lambda_2(\mathbf{L}) = \min_{\mathbf{x}\neq 0, \mathbf{1}^T\mathbf{x}=0} \frac{\mathbf{x}(\mathbf{L}+\mathbf{L}^T)\mathbf{x}}{\|\mathbf{x}\|}. \tag{2.6}$$

For convenience, we will often write $\sigma_2$ and $\lambda_2$ when their arguments are clear from the context.

In this thesis, we only consider convex optimization problems, namely, the functions $f_i$ and the set $\mathcal{X}$ are convex, which can be stated as follows:

**Definition 2** (Convexity [40, 41])**.** *The set $\mathcal{X}$ is convex if and only if*

$$\theta\mathbf{x} + (1-\theta)\mathbf{y} \in \mathcal{X}, \qquad \forall\, \mathbf{x},\mathbf{y} \in \mathcal{X} \quad \text{and} \quad \forall\, \theta \in [0,1].$$

*In addition, the function $f : \mathcal{X} \to \mathbb{R}$ is convex if and only if $\mathcal{X}$ is convex and*

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}).$$

We also consider the following two assumptions on $f_i$ and their gradients, respectively.

**Assumption 5.** *For each $i = 1, \ldots, n$, the function $f_i$ is $\mu_i$-strongly convex if there exists a positive constant $\mu_i$ such that*

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) - \partial f_i(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \geq \frac{\mu_i}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2.7}$$

**Assumption 6.** *For each $i = 1, \ldots, n$, the function $f_i$ is $L_i$-smooth if there exists a constant $L_i$ such that*

$$\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\| \leq L_i \|\mathbf{y} - \mathbf{x}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{2.8}$$

Here $\nabla f(\cdot)$ denotes the gradient of the differential function $f$. We denote by $\partial f(\cdot)$ the subgradient of the non-smooth convex function $f$ [42], i.e., the following holds

$$\partial f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \leq f(\mathbf{x}) - f(\mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{2.9}$$

where $\partial f(\cdot) = \nabla f(\cdot)$ when $f$ is differentiable. Finally, we consider the following assumption about the Lipschitz property of functions $f_i$.

**Assumption 7.** *For each $i = 1, \ldots, n$, the function $f_i$ is $C_i$-Lipschitz continuous if there exists a positive constant $C_i$ such that*

$$|f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq C_i \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,\ \forall i \in \mathcal{V}. \tag{2.10}$$

*The condition in Eq.* (2.10) *is also equivalent to the condition that the norm of the subgradient $\partial f_i$ is bounded by $C_i$ [43, Lemma* 2.6*], that is,*

$$\|\partial f_i(\mathbf{x})\| \leq C_i, \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.11}$$

## 2.2 Distributed Consensus-Based Methods

Our main interest in this thesis is to study distributed consensus-based gradient methods for solving problem (2.1). As will be seen, such methods provide a truly distributed approach, that is, they meet the conditions of distributed algorithms given in Definition 1. For brevity, we often call such methods distributed gradient methods in the remainder of this thesis. We start our discussion with distributed consensus methods for solving network consensus problems, a special case of problem (2.1). We then present distributed gradient methods for solving problem (2.1).

### 2.2.1 Network Consensus Problems

We consider here consensus problems defined over a network of $n$ nodes, where their communication structure is imposed by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In these problems, each node has a real-valued initial estimate, where the goal is to compute the average of these values in a distributed framework. In particular, associated with each node $i \in \mathcal{V}$ is a constant $c_i \in \mathbb{R}$. The goal of the nodes is to compute the average $\bar{c}$ given as

$$\bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i.$$

To solve this problem, the nodes consider distributed consensus methods presented below. For convenience, we consider both continuous-time and discrete-time algorithms, which are useful for our later studies.

### Continuous-time distributed consensus methods

In this algorithm, each node $i$, for all $i \in \mathcal{V}$, maintains a local estimate $x_i \in \mathbb{R}$ of $\bar{c}$, which is initialized to $c_i$. The nodes exchange their values with their local neighbors and then iteratively update as

$$\dot{x}_i(t) = \underbrace{\sum_{j\in\mathcal{N}_i} a_{ij}\,(x_j(t) - x_i(t))}_{\text{“consensus step”}}, \qquad \forall\, t \geq 0, \tag{2.12}$$

where $a_{ij}$ is the weight which node $i$ assigns for the estimate $x_j(t)$, received from its neighbor $j$ at time $k$. The goal is to asymptotically drive every $x_i$ to $\bar{c}$, i.e., $\lim_{k\to\infty} x_i(t) = \bar{c}$, elucidating the term consensus problems.[1] The update in Eq. (2.12) is often referred to as a continuous-time consensus step. The following theorem shows that under an appropriate choice of the weights $a_{ij}$ and the connectivity of $\mathcal{G}$, Eq. (2.12) solves the consensus problems.

**Theorem 1** ( [44] ). *Suppose that Assumptions 1 and 3 hold. Then we have*

$$\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\| \leq \|\mathbf{x}(0) - \bar{x}(0)\mathbf{1}\|\, e^{-\lambda_2 t}, \tag{2.13}$$

*where $\lambda_2$ is the Fiedler eigenvalue of the Laplacian $\mathbf{L}$, defined in Eq.* (2.6).

*Proof.* Recall that the Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{A}$. Since $\mathbf{A}$ satisfies Assumption 3 and the graph $\mathcal{G}$ is connected, $\mathbf{L}$ has 0 as the smallest eigenvalue with the corresponding eigenvector $\mathbf{1}$, and all other eigenvalues are strictly greater than 0. Using $\mathbf{L}$ and Eq. (2.12) gives

$$\dot{\mathbf{x}}(t) = -\mathbf{L}\mathbf{x}(t),$$

which implies that $\dot{\bar{x}}(t) = 0$ and

$$-\mathbf{L}\mathbf{x}(t) = -\mathbf{L}(\mathbf{x}(t) - \bar{x}(t)\mathbf{1}).$$

Using the two equations above we then obtain

$$\dot{\mathbf{x}}(t) - \dot{\bar{x}}(t)\mathbf{1} = -\mathbf{L}(\mathbf{x}(t) - \bar{x}(t)\mathbf{1}),$$

which implies

$$\mathbf{x}(t) - \bar{x}(t)\mathbf{1} = e^{-\mathbf{L}t}(\mathbf{x}(0) - \bar{x}(0)\mathbf{1}).$$

Applying the 2-norm to the above and using Eq. (2.6) gives Eq. (2.13). □

[1] In the literature, such problems are referred to as averaging problems. We often call these consensus problems.

### Discrete-time distributed consensus methods

Similarly, the discrete-time counterpart of Eq. (2.12) is given as follows:

$$x_i(k+1) = \underbrace{\sum_{j\in\mathcal{N}_i} a_{ij}x_j(k)}_{\text{“consensus step”}}, \qquad \forall\, k \geq 0. \tag{2.14}$$

The following theorem is a variant of Theorem 1.

**Theorem 2** ( [45] ). *Suppose that Assumptions 1 and 3 hold. Then we have*

$$\|\mathbf{x}(k) - \bar{x}(k)\mathbf{1}\| \leq \|\mathbf{x}(0) - \bar{x}(0)\mathbf{1}\|\, \delta^k, \tag{2.15}$$

*where* $\delta \leq \min\{\left(1 - \frac{1}{2n^3}\right)\ ,\ \sigma_2\}$. *Here* $\sigma_2$ *is defined in Eq.* (2.5).

*Proof.* We provide here a proof for the case of $\delta = \sigma_2$, where the first bound is studied in [24, 45]. First, the double stochasticity of $\mathbf{A}$ gives

$$\bar{x}(k+1)\mathbf{1} = \frac{1}{n}\mathbf{1}^T\mathbf{x}(k+1)\mathbf{1} = \frac{1}{n}\mathbf{1}^T\mathbf{A}(k)\mathbf{x}(k)\mathbf{1} = \mathbf{A}(k)\bar{x}(k)\mathbf{1}.$$

Using the preceding relation and $\sigma_2 \in (0,1)$ in Eq. (2.5), we obtain

$$\|\mathbf{x}(k+1) - \bar{x}(k+1)\mathbf{1}\| = \|\mathbf{A}(\mathbf{x}(k) - \bar{x}(k)\mathbf{1})\| \leq \sigma_2\|\mathbf{x}(k) - \bar{x}(k)\mathbf{1}\|,$$

where we use $\mathbf{1}^T(\mathbf{x}(k) - \bar{x}(k)\mathbf{1}) = 0$. Iterating the above over $k$ gives Eq. (2.15). □

**Remark.** *In continuous-time distributed consensus methods,* $\dot{\bar{x}}(t) = 0$ *gives*

$$\bar{x}(t) = \frac{1}{n}\mathbf{1}^T\mathbf{x}(t) = \frac{1}{n}\mathbf{1}^T\mathbf{x}(t-1) = \cdots = \frac{1}{n}\mathbf{1}^T\mathbf{x}(0) = \bar{x}(0) = \bar{c},$$

*which by Eq.* (2.13) *implies that* $\lim_{t\to\infty} x_i(t) = \bar{c}$, *for all* $i \in \mathcal{V}$. *In addition, the rate of convergence is linear and depends on the algebraic connectivity of the network represented by* $\lambda_2$. *A similar conclusion holds for Eq.* (2.15).

### 2.2.2 Distributed Gradient Methods

We now consider problem (2.1) where we study distributed gradient methods, which are developed based on the consensus methods discussed in Section

2.2.1. To explain the ideas of distributed gradient methods, we start with gradient descent methods [46] for solving problem (2.1) when $\mathcal{X} = \mathbb{R}$,

$$x(k+1) = x(k) - \alpha \sum_{i=1}^{n} f_i'(x(k)), \qquad \forall\, k \geq 0,$$

where $\alpha$ is some positive constant stepsize. It is obvious that the gradient descent method requires the derivatives $f_i'(\cdot)$ of all functions $f_i$ at every iteration. Such a requirement, in general, may not be achievable or may be expensive to compute in distributed frameworks due to the absence of a central coordinator. To circumvent this requirement, we consider distributed gradient methods, which are a combination of distributed consensus steps and local gradient descent steps. In particular, each node $i$ in $\mathcal{G}$ maintains a local estimate $x_i$ of the solution $x^*$ of problem (2.1). The nodes then initialize their estimates arbitrarily and iteratively update them in parallel as

$$x_i(k+1) = \underbrace{\sum_{j\in\mathcal{N}_i} a_{ij}x_j(k)}_{\text{“consensus step”}} - \underbrace{\alpha(k)f_i'(x_i(k))}_{\text{“local gradient step”}}, \qquad \forall\, i \in \mathcal{V},\ \ \forall\, k \geq 0, \tag{2.16}$$

where $\{\alpha(k)\}$ is a sequence of stepsizes. The update in Eq. (2.16) is composed of two parts, namely, a consensus step and a local gradient step, hence the name distributed consensus-based gradient methods. Moreover, the nodes only require local computation and communication with neighboring nodes.

The update in Eq. (2.16) has a simple interpretation: at any time $k \geq 0$, each node $i$ first combines its value $x_i(k)$ with the weighted values received from its neighbors, with the goal of seeking consensus on their estimates. Each node then moves along the gradient of its respective objective function to update its value, pushing the consensus point toward the optimal set $\mathcal{X}^*$.

We now show that under reasonable conditions on the graph connectivity, the weighted adjacency matrix $\mathbf{A}$, and the sequence of stepsizes $\{\alpha(k)\}$, that $\lim_{k\to\infty} x_i(k) = x^*$, for all $i \in \mathcal{V}$; this implies that the nodes achieve a consensus, which is an optimal solution of problem (2.1).

**Main ideas:** The analysis of distributed gradient methods is composed of two key steps. In particular, due to the consensus step the nodes asymptotically agree on some common quantity, which is the average of the nodes' values. Second, this average asymptotically converges to the solution of problem (2.1), which is a consequence of the local gradient steps.

We review here the existing results on the convergence of Eq. (2.16) as well as its continuous-time counterpart for solving problem (2.1). We start with continuous-time distributed gradient methods. For completeness, we present the analysis in Appendix A, which allows us to study the impact of communication delays in Chapter 3. We use the following notation in the remainder of this section,

$$\nabla F(\mathbf{x}) \triangleq [f_1'(x_1), \ldots, f_n'(x_n)]^T.$$

## Continuous-time distributed gradient methods

We consider the continuous-time variant of Eq. (2.16) given as

$$\dot{x}_i(t) = \underbrace{\sum_{j\in\mathcal{N}_i} a_{ij}(x_j(t) - x_i(t))}_{\text{“consensus step”}} - \underbrace{\alpha(t) f_i'(x_i(t))}_{\text{“local gradient step”}}\,, \qquad \forall\, i \in \mathcal{V}. \tag{2.17}$$

We first have the following result, an extension of Theorem 1.

**Lemma 1.** *Suppose that Assumptions 1, 3, and 7 hold. Let the trajectory $\{x_i(t)\}$, for all $i \in \mathcal{V}$, be generated by Eq. (2.17). Let $\{\alpha(t)\}$ be a non-increasing positive scalar sequence with $\alpha(t) = 1$ for $0 \leq t \leq 1$. Then the following statements hold.*

1. *For all $i \in \mathcal{V}$ and $t \geq 0$*

$$\big|x_i(t) - \bar{x}(t)\big| \leq e^{-\lambda_2 t}\|\mathbf{x}(0)\| + \int_0^t e^{-\lambda_2(t-u)}\alpha(u)\big\|\nabla F(\mathbf{x}(u))\big\| du. \tag{2.18}$$

2. *If $\lim_{t\to\infty}\alpha(t) = 0$ then*

$$\lim_{t\to\infty}\big|x_i(t) - \bar{x}(t)\big| = 0 \quad \forall\, i \in \mathcal{V}. \tag{2.19}$$

3. *If $\int_{t=0}^{\infty}\alpha^2(t)dt < \infty$ then we obtain*

$$\int_0^{\infty}\alpha(t)\big|x_i(t) - \bar{x}(t)\big|dt < \infty \quad \forall\, i \in \mathcal{V}. \tag{2.20}$$

Based on Lemma 1, we now state the main result of Eq. (2.17), which is the asymptotic convergence of the nodes' estimates to $x^*$.

**Theorem 3.** *Suppose that Assumptions 1, 3, and 7 hold. Let the trajectory* $\{x_i(t)\}$*, for all* $i \in \mathcal{V}$*, be generated by Eq.* (2.17)*. Let* $\{\alpha(t)\}$ *be a non-increasing positive scalar sequence with* $\alpha(t) = 1$ *for* $0 \leq t \leq 1$ *, and satisfy*

$$\int_{t=0}^{\infty} \alpha(t)dt = \infty \qquad \text{and} \qquad \int_{t=0}^{\infty} \alpha^2(t)dt < \infty. \tag{2.21}$$

*Then we have*

$$\lim_{t\to\infty} \bar{x}(t) = x^*. \tag{2.22}$$

Finally, to study the convergence rate of Eq. (2.17) we consider the stepsizes $\alpha(t) = 1/\sqrt{t}$, as motivated by centralized subgradient methods [46]. Using this stepsize, we now show the rate of convergence of Eq. (2.17).

**Theorem 4.** *Suppose that Assumptions 1, 3, and 7 hold. Let the trajectory* $\{x_i(t)\}$*, for all* $i \in \mathcal{V}$*, be generated by Eq.* (2.17)*. Let* $\alpha(t) = 1/\sqrt{t}$ *for* $t \geq 1$ *and* $\alpha(t) = 1$ *for* $0 \leq t \leq 1$*. Moreover, suppose that each node* $i$*, for all* $i \in \mathcal{V}$*, stores a variable* $z_i \in \mathbb{R}$*, which is initialized arbitrarily and updated as*

$$\dot{z}_i(t) = \frac{\alpha(t)x_i(t) - \alpha(t)z_i(t)}{S(t)}, \quad \forall\, t > 0, \tag{2.23}$$

*where* $S(0) = 0$ *and* $\dot{S}(t) = \alpha(t)$ *for* $t > 0$*. Then for all* $i \in \mathcal{V}$

$$f(z_i(t)) - f^* \leq \mathcal{O}\left(\frac{\ln(t)}{\lambda_2^2\sqrt{t}}\right). \tag{2.24}$$

## Discrete-time distributed gradient methods

Similar to the continuous-time counterpart above, we provide here the convergence results of Eq. (2.16) for solving problem (2.1).

**Lemma 2** ( [17, 24]). *Suppose that Assumptions 1, 3, and 7 hold. Let the sequence* $\{x_i(k)\}$*, for all* $i \in \mathcal{V}$*, be generated by Eq.* (2.16)*. Let* $\{\alpha(k)\}$ *be a non-increasing positive scalar sequence with* $\alpha(0) = 1$*. Then the following statements hold.*

1. *For all $i \in \mathcal{V}$ and $k \geq 0$*

$$\big|x_i(k) - \bar{x}(k)\big| \leq \delta^k \big\|\mathbf{x}(0)\big\| + \sum_{t=0}^{k} \delta^{k-t}\alpha(t)\big\|\nabla F(\mathbf{x}(t))\big\|, \tag{2.25}$$

   *where $\delta \leq \min\{1 - \frac{1}{2n^3}\ ,\ \sigma_2(\mathbf{A})\}$.*

2. *If $\lim_{k\to\infty}\alpha(k) = 0$ then we have*

$$\lim_{k\to\infty}\big|x_i(k) - \bar{x}(k)\big| = 0, \qquad \forall\, i \in \mathcal{V}. \tag{2.26}$$

3. *If $\sum_{k=0}^{\infty}\alpha^2(k) < \infty$ then we obtain*

$$\sum_{k=0}^{\infty}\alpha(k)\big|x_i(k) - \bar{x}(k)\big| < \infty, \qquad \forall\, i \in \mathcal{V}. \tag{2.27}$$

**Theorem 5** ( [18])**.** *Suppose that Assumptions 1, 3, and 7 hold. Let the sequence $\{x_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Eq. (2.16). Let $\{\alpha(k)\}$ be a non-increasing positive scalar sequence with $\alpha(0) = 1$, and satisfy*

$$\sum_{k=0}^{\infty}\alpha(k) = \infty \qquad \text{and} \qquad \sum_{k=0}^{\infty}\alpha^2(k) < \infty. \tag{2.28}$$

*Then we have*

$$\lim_{k\to\infty}\bar{x}(k) = x^*. \tag{2.29}$$

**Theorem 6.** *Suppose that Assumptions 1, 3, and 7 hold. Let the trajectory $\{x_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Eq. (2.16). Let $\alpha(k) = 1/\sqrt{k}$ for $k \geq 1$ and $\alpha(0) = 1$. Moreover, suppose that each node $i$, for all $i \in \mathcal{V}$, stores a variable $z_i(k) \in \mathbb{R}$, which is initialized arbitrarily and updated as*

$$z_i(k+1) = \frac{\alpha(k+1)x_i(k+1) + \alpha(k)z_i(k)}{S(k+1)}, \tag{2.30}$$

*where $S(0) = 0$ and $S(k) = \sum_{t=0}^{k}\alpha(t)$ for $k > 0$. Then for all $i \in \mathcal{V}$*

$$f(z_i(k)) - f^* \leq \mathcal{O}\left(\frac{\ln(k)}{(1-\sigma_2)^2\sqrt{k}}\right)\cdot \tag{2.31}$$

**Remark.** *First, we note that the stepsizes in Eq.* (2.28) *also satisfy the conditions of stepsizes in Eqs.* (2.26) *and* (2.27) *in Lemma 2. Thus, it is immediate to see that Eq.* (2.29) *implies* $\lim_{k\to\infty} x_i(k) = x^*$*, for all* $i \in \mathcal{V}$*.*

*Second, Theorem 6 reveals that* $f$ *evaluated at a time-weighted average of each node's values converges to the optimal value* $f^*$*. Moreover, it shows that this convergence occurs at a rate* $\mathcal{O}\left(\ln(k)/\sqrt{k}\right)$*. The term* $\mathcal{O}\left(1/\sqrt{k}\right)$ *mirrors the convergence results for centralized subgradient algorithms, for example, see [46, Chapter 3]. However, the distributed nature of the algorithms slows the convergence by a factor of* $\ln(k)$*. When the objective functions are strongly convex, i.e., Assumption 5 holds, we can further show that Eq.* (2.16) *achieves a rate* $\mathcal{O}(\ln(k)/k)$ *when* $\alpha(k) = 1/(k+1)$ *[26]. In addition, the convergence rate depends inversely on* $1-\sigma_2$*, the spectral gap of* $\mathbf{A}$*. Here,* $\sigma_2$ *presents the speed of information among the nodes is diffused over networks.*

*Third, the results presented above are straightforward to extend to the multidimentional case* $d > 1$ *and constrained problems,* $\mathcal{X} \subset \mathbb{R}$*, which can be found in Appendix B. In particular, for constrained problems we consider the following distributed projected gradient methods [18]*

$$x_i(k+1) = \mathcal{P}_{\mathcal{X}}\left[\sum_{j\in\mathcal{N}_i} a_{ij}x_j(k) - \alpha(k)f_i'(x_i(k))\right], \quad \forall\, i \in \mathcal{V}, \ \ \forall\, k \geq 0.$$

*We can also extend these results to the case of time-varying graphs* $\mathcal{G}(k)$ *under Assumptions 2 and 4 as studied in Chapter 7. Finally, all the statements above hold for the continuous-time counterpart.*

# Chapter 3

# Convergence Rate of Distributed Gradient Methods with Communication Delays

## 3.1 Motivation and Contribution

In this chapter, we consider the optimization problem (2.1), i.e.,

$$\text{minimize } \sum_{i=1}^{n} f_i(\mathbf{x}) \text{ over } \mathbf{x} \in \mathcal{X},$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact convex set known by the nodes. Our focus is to study distributed gradient methods for solving the problem of Eq. (2.1), while explicitly accounting for network delays, one of the most critical issues in distributed systems. In particular, we focus on the convergence rate of these methods in the presence of inter-node communication delays, which has been identified as an important problem in [47, see Chapter 10]. Communication delay has been studied in other contexts, such as distributed dual averaging [48]. The analysis in [48] is based on adding fictitious nodes corresponding to the number of time delay steps, thus requiring a modification of the true network topology. As a result, the influence of the delays on the convergence rate for the original network topology is not clear. Convergence under delays is also considered in distributed consensus algorithms [49–53]. However, these results do not apply to the distributed gradient algorithms. Our goal, therefore, is to address this important problem of proving convergence and obtaining convergence rates for distibuted gradient algorithms with inter-node communication delays.

**Main Contributions**. The main contribution of this chapter is to derive the convergence rate of distributed gradient algorithms under uniform communication delays between nodes. Due to the delays, we first redesign the algorithm by introducing a free parameter, which allows us to establish the rate of convergence of the algorithm. We show that the algorithm achieves the same rate as in the delay-free case in Eq. (2.24), except for a factor which captures the impact of delays. In particular, the convergence occurs at rate $\mathcal{O}\Big(n\tau^3\ln(t)/(1-\gamma)^2\sqrt{t}\Big)$, where $n$ is the number of nodes, $t$ is the time variable, and $\tau$ is the delay constant. In addition, $\gamma$ is a constant in $(0,1)$ that depends on $\tau$ and $\sigma_2$, which reflects the spectral properties of the network connectivity of the nodes. We note that such an explicit formula for the convergence rate is not available for dual averaging methods. As remarked, the existing analysis in distributed optimization literature cannot be extended to show this result. We, therefore, introduce a new approach by considering a new candidate Krasovskii Lyapunov function, that directly takes into account the impact of delays. Finally, while we do not analyze dual averaging methods in the presence of delays, we provide simulation results comparing it to distributed gradient methods, which indicate that distributed gradient methods perform significantly better.

For ease of exposition, we study the convergence rate for the continuous-time version of distributed gradient methods, Eq. (2.17), with communication delays in this chapter. In addition, we consider problem (2.1) when the variable $\mathbf{x}$ is a scalar, i.e., $d = 1$. Extensions for the case $d > 1$ and the results of the discrete-time version are presented in Appendix B. The results in this chapter have been presented in [54, 55]

## 3.2 Continuous-Time Distributed Gradient Methods with Delays

We consider the continuous-time distributed gradient method in Eq. (2.17) with uniform communication delays between nodes. In particular, we assume that at any time $t \geq 0$ each node $i$, for all $i \in \mathcal{V}$, only receives a delayed value $x_j(t-\tau)$ of $x_j(t)$ from node $j \in \mathcal{N}_i$, where $\tau$ is a constant representing the time delay of communication between nodes. Each node $i$ then uses these values to update its estimate as stated in Eq. (3.1), where $\mathcal{T}_{\mathcal{X}(x_i(t))}$ is the tangent cone of $\mathcal{X}$ at $x_i(t)$, $\beta$ is some positive constant, and $\alpha(t)$ is a sequence of positive stepsizes. The conditions of $\beta$ and $\alpha(t)$ to guarantee the convergence of the algorithm will be given explicitly later. Here, the initial conditions, $\phi_i(t)$, are assumed to be continuous functions of time. Thus, the estimates $x_i(t)$ are functional since they are functions of $\phi_i(t)$. The continuous-time distributed gradient algorithm with communication delays is formulated in Algorithm 1.

```
Algorithm 1 Continuous-Time Distributed Gradient Algorithm with Delays
1. Initialize: Each node $i$ is initiated with $x_i(t) = \phi_i(t) \in \mathcal{X},\ t \in [-\tau, 0]$.
2. Iteration: For $t \geq 0$ each node $i \in \mathcal{V}$ executes
    $$\dot{x}_i(t) = \mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}}\left[-\beta x_i(t) + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - \alpha(t)f_i'(x_i(t))\right]. \qquad (3.1)$$
```

## 3.3 Main Results

The focus of this section is to analyze the performance of distributed gradient methods under communication delays, given in Algorithm 1. In particular, we provide a rigorous analysis which establishes the convergence rate of Algorithm 1. The main steps of the analysis are as follows.

We first show that the distances between the estimates $x_i(t)$ to their average $\bar{x}(t)$ asymptotically converge to zero. We then study the convergence rate of Algorithm 1, where we utilize the standard techniques used in the centralized version of subgradient methods. The key idea of this step is to introduce a candidate Krasovskii Lyapunov functional, which takes into account the impact of delays on the system. By using this function, we can show that the impact of delays is asymptotically negligible. In particular, we show that if each node maintains a variable $z_i(t)$ to compute the time-weighted average of the estimate $x_i(t)$ and if the stepsizes decay with the rate $\alpha(t) = 1/\sqrt{t}$, the algorithm achieves an asymptotic convergence to the optimal value estimated on the variable $z_i(t)$ at a rate

$$\mathcal{O}\left(\frac{n\tau^3 \ln(t)}{(1-\gamma)^2\sqrt{t}}\right),$$

where $\beta \in (0\,,\, \ln(1/\sigma_2)/\tau)$, $\gamma = \sigma_2 e^{\beta\tau} \in (0\,,\,1)$, and $\sigma_2$ is given in Eq. (2.5).

We start our analysis by introducing a bit more notation. We denote by $\mathcal{D}_{\mathcal{X}}(x)$ the set of feasible directions of $x$ in $\mathcal{X}$, i.e.,

$$\mathcal{D}_{\mathcal{X}}(x) = \{y \in \mathbb{R} \mid \exists\, \theta > 0 \text{ s.t. } x + \theta y \in \mathcal{X}\}. \qquad (3.2)$$

In the sequel we use the following result from [40].

**Proposition 1** (Proposition 4.6.2 [40])**.** *Let $\mathcal{X}$ be a closed convex set. Then the tangent cone $\mathcal{T}_{\mathcal{X}}(x)$ at $x \in \mathcal{X}$ is closed, convex, and $\mathcal{T}_{\mathcal{X}}(x) = cl(\mathcal{D}_{\mathcal{X}}(x))$, where $cl(\mathcal{D}_{\mathcal{X}}(x))$ is the closure of $\mathcal{D}_{\mathcal{X}}(x)$.*

In addition, we use the following notation

$$F(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(x_i), \quad \nabla F(\mathbf{x}(t)) \triangleq [f_1'(x_1), \ldots, f_n'(x_n)]^T, \quad C \triangleq \sum_{i=1}^{n} C_i,$$

where recall that $C_i$ is the Lipschitz constant of $f_i$ given in Eq. (2.10). In this case, since the set $\mathcal{X}$ is compact, Assumption 7 is always satisfied.

Without loss of generality we consider $\mathcal{X} = [a, b]$ for some real numbers $a \leq b \in \mathbb{R}$. This simplification will allow us to write explicitly the projection on the tangent cone in Eq. (3.1). In particular, given a real number $v$ we denote $v^+ = \max(0, v)$, the positive part of $v$. Similarly, we denote $v^- = \max(0, -v)$, the negative part of $v$. The update in Eq. (3.1) can now be rewritten as

$$v_i(t) = -\beta x_i(t) + \beta \sum_{j \in \mathcal{N}_i} a_{ij} x_j(t - \tau) - \alpha(t) f_i'(x_i(t)) \tag{3.3}$$

$$\dot{x}_i(t) = \mathcal{P}\left(v_i(t)\right) = \begin{cases} v_i(t) & \text{if } \; x_i(t) \in (a, b) \\ v_i^+(t) & \text{if } \; x_i(t) = a \\ -v_i^-(t) & \text{if } \; x_i(t) = b. \end{cases} \tag{3.4}$$

Given $v_i \in \mathcal{X}$ we denote by $\zeta_i$ the error due to projection of $v_i$ to $\mathcal{T}_{\mathcal{X}(x_i)}$, i.e., $\zeta_i(v_i) = v_i - \mathcal{P}\left(v_i\right)$. Using this notation and the weighted adjacency matrix $\mathbf{A}$ defined in Eq. (2.4), Eqs. (3.3) and (3.4) can be rewritten compactly as

$$\mathbf{v}(t) = -\beta \mathbf{x}(t) + \beta \mathbf{A} \mathbf{x}(t - \tau) - \alpha(t) \nabla F(\mathbf{x}(t)) \tag{3.5}$$

$$\dot{\mathbf{x}}(t) = \mathcal{P}(\mathbf{v}(t)) = \mathbf{v}(t) - \zeta(\mathbf{v}(t)), \tag{3.6}$$

where $\mathcal{P}(\mathbf{v}(t))$ denotes the component-wise projection. Moreover, we have

$$\bar{v}(t) = -\beta \bar{x}(t) + \beta \bar{x}(t - \tau) - \frac{\alpha(t)}{n} \sum_{i=1}^{n} f_i'(x_i(t)) \tag{3.7}$$

$$\dot{\bar{x}}(t) = \bar{z}(t) - \zeta(\mathbf{v}(t)). \tag{3.8}$$

As remarked, the first step in our analysis is to show the asymptotic convergence of $\big\|\mathbf{x}(t)-\bar{x}(t)\mathbf{1}\big\|$ to zero under some appropriate choice of stepsizes. The following lemma, which will be essential for our analysis later, is an important facet of this result.

**Lemma 3.** *Suppose that Assumptions 1 and 3 hold. Let the trajectories of $x_i(t)$, for all $i\in\mathcal{V}$, be updated by Algorithm 1. Let $\{\alpha(t)\}$ be a given non-increasing positive scalar sequence with $\alpha(0)=1$. Moreover, let*

$$\beta\in\left(0\,,\,\frac{\ln(1/\sigma_2)}{\tau}\right)\qquad \textit{and}\qquad \gamma=\sigma_2e^{\beta\tau}\in(0\,,\,1).$$

*Then the following statements hold.*

1. *For all $t\geq 0$ we have*

$$\big\|\mathbf{x}(t)-\bar{x}(t)\mathbf{1}\big\|\leq\mu(t)+\beta\sigma_2\int_0^t e^{-\beta(1-\gamma)(t-u)}\mu(u-\tau)du,\tag{3.9}$$

   *where*

$$\mu(t)=\frac{\|\mathbf{x}(0)\|+2C}{\beta}e^{-\beta t/2}+\frac{2C\alpha(t/2)}{\beta}.\tag{3.10}$$

2. *If $\lim\limits_{t\rightarrow\infty}\alpha(t)=0$ then we have*

$$\lim_{t\rightarrow\infty}\big|x_i(t)-\bar{x}(t)\big|=0,\qquad\forall\, i\in\mathcal{V}.\tag{3.11}$$

3. *Further we have*

$$\begin{aligned}&\int_0^t\alpha(u)\big\|\mathbf{x}(u)-\bar{x}(u)\mathbf{1}\big\|du\\&\qquad\leq\frac{8\left(\|\mathbf{x}(0)\|+2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2}+\frac{4C}{\beta^2(1-\gamma)}\int_0^t\alpha^2(\gamma u/4-\tau)du.\end{aligned}\tag{3.12}$$

*Proof sketch.* The main idea in the proof of Lemma 3 is to show Eq. (3.9). The analysis of Eqs. (3.11) and (3.12) are the consequences of Eq. (3.9) with the given assumptions on stepsizes and proper algebraic manipulations. We, therefore, provide here the key steps for the proof of Eq. (3.9), where the details are delayed to Section 3.5.

(a) Denote $\mathbf{y}(t) \triangleq \mathbf{x}(t) - \bar{x}(t)\mathbf{1}$. By Eqs. (3.6) and (3.8), $\dot{\mathbf{y}}(t)$ is given as

$$\dot{\mathbf{y}}(t) = -\beta\mathbf{y}(t) + \beta\mathbf{A}\mathbf{y}(t-\tau) - \alpha(t)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t)) \\ - \alpha(t)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{v}(t)). \qquad (3.13)$$

Due to the delay term $\mathbf{A}\mathbf{y}(t-\tau)$, we would expect an accumulation of this term for the solution $\mathbf{y}(t)$ of Eq. (3.13). Indeed, we have

$$\mathbf{y}(t) = e^{-\beta t}\mathbf{y}(0) + \beta\int_0^t e^{-\beta(t-u)}\mathbf{A}\mathbf{y}(u-\tau)du \\ - \int_0^t e^{-\beta(t-u)}\alpha(u)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\Big(\nabla F(\mathbf{x}(u)) + \zeta(\mathbf{v}(u))\Big)du.$$

(b) Next, taking the 2-norm of above and using the triangle inequality give

$$\big\|\mathbf{y}(t)\big\| \leq e^{-\beta t}\big\|\mathbf{y}(0)\big\| + \beta\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du \\ + \int_0^t e^{-\beta(t-u)}\left\|\alpha(u)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\Big(\nabla F(\mathbf{x}(u)) + \zeta(\mathbf{v}(u))\Big)\right\|du.$$

By the Cauchy-Schwarz inequality we can show that

$$\left\|\alpha(u)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t))\right\| \leq \alpha(u)C.$$

Furthermore, by Eq. (3.4) we can obtain

$$\left\|\alpha(u)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{v}(u))\right\| \leq \alpha(u)C.$$

(c) Finally, the key step of our analysis is to provide an upper bound for

$$\beta\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du,$$

which is done by applying the *Grönwall-Bellman* inequality [56].

□

We are now ready to state the main result of this section, which is the convergence rate of Algorithm 1 to the optimal value using standard techniques from the analysis of centralized subgradient methods. One can view the update $\bar{x}(t)$ in Eq. (3.8) as a centralized projected subgradient used to solve problem (2.1). Specifically, at any time $t \geq 0$ if each node $i \in \mathcal{V}$ maintains a variable $z_i(t)$ to compute the time-weighted average of $x_i(t)$ and if the stepsizes $\alpha(t)$ decay as $\alpha(t) = 1/\sqrt{t}$, then the objective function value $f$ in problem (2.1), estimated at any $z_i(t)$, converges to the optimal value with a rate

$$\mathcal{O}\left(\frac{n\tau^3 \ln(t)}{(1-\gamma)^2\sqrt{t}}\right),$$

where $\gamma = \sigma_2 e^{\beta\tau} \in (0\,,\,1)$ and $\beta \in (0\,,\,\ln(1/\sigma_2)/\tau)$. We also note that this condition on the stepsizes is also used to study the convergence rate of centralized subgradient methods [46]. The following theorem is used to show the convergence rate of Algorithm 1, and its proof is given in Section 3.5.

**Theorem 7.** *Suppose that Assumptions 1 and 3 hold. Let the trajectories of $x_i(t)$, for all $i \in \mathcal{V}$, be updated by Algorithm 1. Let $\beta \in (0\,,\,\ln(1/\sigma_2)/\tau)$ and $\gamma = \sigma_2 e^{\beta\tau} \in (0\,,\,1)$. Let $\alpha(t) = 1/\sqrt{t}$ for $t \geq 1$ and $\alpha(t) = 1$ for $t \leq 1$. Moreover, suppose that each node $i$, for all $i \in \mathcal{V}$, stores a variable $z_i \in \mathbb{R}$, which is initialized arbitrarily and updated as*

$$\dot{z}_i(t) = \frac{\alpha(t)x_i(t) - \alpha(t)z_i(t)}{S(t)}, \quad \forall\, t \geq 0, \tag{3.14}$$

*where $S(0) = 0$ and $\dot{S}(t) = \alpha(t)$ for $t > 0$. Then*

$$f\left(z_i(t)\right) - f^* \leq \frac{2\Gamma_0(t) + nV(\bar{x}(0))}{2\sqrt{t}-1}, \qquad \forall\, i \in \mathcal{V}, \tag{3.15}$$

*where*

$$\Gamma_0(t) \triangleq \frac{24C\left(\|\mathbf{x}(0)\| + 2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{48C^2(1+\tau)}{\beta^2\gamma(1-\gamma)} + C^2\ln(t) + \frac{48C^2\ln(\gamma t - 4\tau)}{\beta^2\gamma(1-\gamma)}.$$

*Sketch of Proof.* As mentioned previously, the main idea of this proof is to introduce a candidate Lyapunov functional, which takes into account the impact of delays. In particular, a quadractic candidate Lyapunov function, i.e., $(\bar{x}(t)-x^*)^2$, is often used in the case of no communication delay. However, since the estimates $x_i(t)$ depend on the interval $[t-\tau, t]$ we consider an extra term to study this impact. Specifically, we consider the following candidate

Krasovskii Lyapunov functional $V$ [57]

$$V(\bar{x}(t)) = \frac{1}{2}(\bar{x}(t) - x^*)^2 + \frac{\beta}{2}\int_{t-\tau}^{t}(\bar{x}(s) - x^*)^2 ds.$$

We then show that $V$ is sufficiently decreasing by considering the following two main steps.

1. One can show that the derivative of $V$ satisfies

$$\dot{V}(\bar{x}(t)) \leq \frac{2C\alpha(t)}{n}\left\|\mathbf{x}(t) - \bar{x}(t)\right\| + \frac{C^2\alpha^2(t)}{n} - \frac{\alpha(t)}{n}(f(\bar{x}(t)) - f^*).$$

2. Integrating the above and using Eq. (3.12) gives the rate in Eq. (3.15).

□

## 3.4 Simulations

In this section, we apply the distributed gradient algorithm to study the well-known linear regression problem in statistical machine learning, which is the most popular technique for data fitting [1,2]. The goal of this problem is to find a linear relationship between a set of variables and some real value outcome. Here, we focus on quadratic loss functions, that is, given a training set $S = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}$ for $i = 1, \ldots, n$, we want to learn a parameter $\mathbf{w}$ that minimizes the following least squares problem,

$$\min_{\mathbf{w}\in\mathcal{X}} \sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w} - y_i)^2. \tag{3.16}$$

We assume that the datasets are distributedly stored in a network of $n$ processors, i.e., each processor $i$ knows only the pair $(\mathbf{x}_i, y_i)$.

For the purpose of simulations, we consider the discrete-time version of Algorithm 1, i.e., Eq. (2.16) with communication delays $\tau$. We simulate the case when $\mathcal{X} = [-5,\ 5]^d$ where $d = 10$, i.e., $w,\ \mathbf{x}_i \in \mathbb{R}^{10}$. We consider simulated training datasets, i.e., $(\mathbf{x}_i, y_i)$ are generated randomly with uniform distribution between $[0, 1]$. We consider the performance of the distributed gradient algorithm on different sizes of network $\mathcal{G}$, where each network is generated as follows:

1. We first randomly generate the nodes' coordinates in the plane with uniform distribution.

2. Then any two nodes are connected if their distance is less than a reference number $r$, e.g., $r = 0.6$ for our simulations.

3. Finally we check whether the network is connected. If not we return to step 1 and run the program again.

To implement our algorithm, the communication matrix $\mathbf{A}$ is chosen as a lazy Metropolis matrix corresponding to $\mathcal{G}$, i.e.,

$$\mathbf{A} = [a_{ij}] = \begin{cases} \frac{1}{2(\max\{|\mathcal{N}_i|,|\mathcal{N}_j|\})}, & \text{if } (i,j) \in \mathcal{E} \\ 0, & \text{if } (i,j) \notin \mathcal{E} \text{ and } i \neq j \\ 1 - \sum_{j\in\mathcal{N}_i} a_{ij}, & \text{if } i = j. \end{cases}$$

It is straightforward to verify that the lazy Metropolis matrix $\mathbf{A}$ satisfies Assumption 3. In all simulations considered herein, we set the stepsizes $\alpha(k) = 1/\sqrt{k}$ for $k = 1, 2, \ldots$ and $\alpha(0) = 1$.

In the sequel, we will compare the performance of the discretized version of distributed gradient (DG) with distributed dual averaging (DA) [48, 58] for solving problem (3.16) in the delay-free case as well as in the case of constant delays. For DA, we chose the same stepsize $\alpha(k) = 1/\sqrt{k}$ as used in our algorithm. Simulations show that the distributed gradient algorithm outperforms distributed dual averaging in both cases.

### 3.4.1 Delay-free case

In the delay-free case, i.e., $\tau = 0$, we simulate DG and DA for two networks, namely, with $n = 40$ and $n = 50$. In each simulation, we fix the number of iterations $k = 1000$ and output the worst-case distance of the function value to the optimal value, i.e., $\max_i |f(z_i(k)) - f^*|$, where $z_i(k) = \frac{1}{k}\sum_{t=1}^{k} x_i(t)$. The simulations are shown in Fig. 3.1, which show that the performance of DG is slightly better than that of DA. However, overall they seem to share the same rate $\mathcal{O}(\ln(k)/\sqrt{k})$, which agrees with the result in Theorem 6.

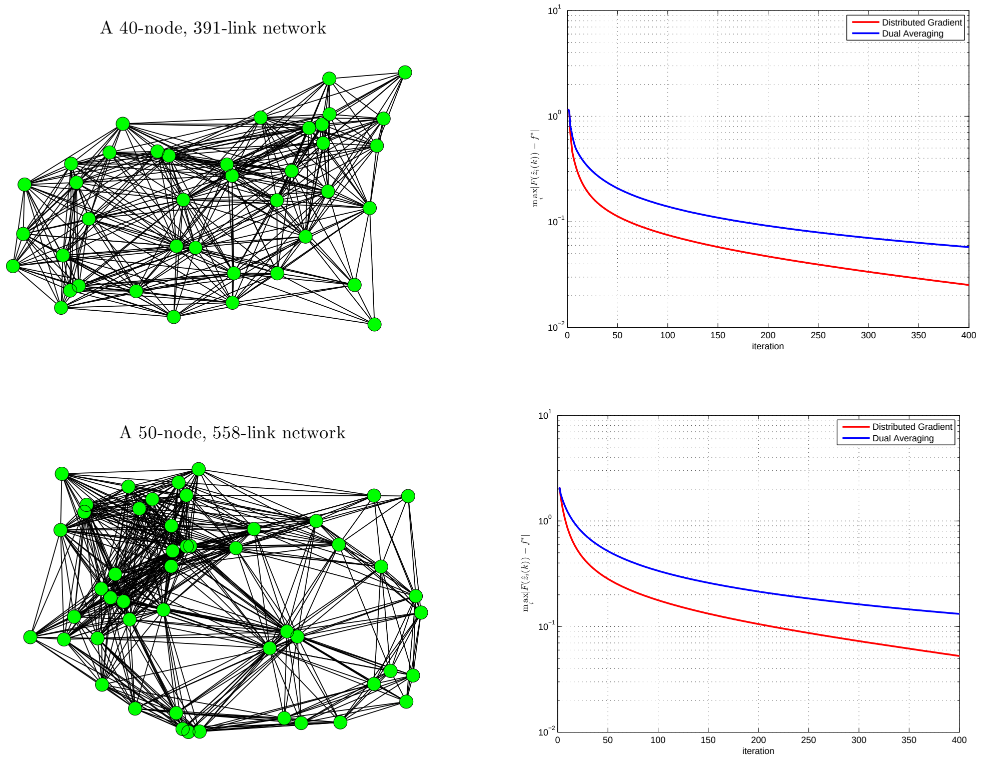

Figure 3.1: Performance of DG and DA in delay-free case over two networks with 40 and 50 nodes on the top and the bottom plots, respectively.

### 3.4.2 Uniform delays

To study the impact of uniform communication delays on the performance of DG and DA, we simulate the two algorithms for two networks above. We implement DG and DA for each network, and terminate them when $\max_i |f(z_i(k)) - f^*| \leq 0.2$. We let the delay constant $\tau$ run from 0 to 10 and output the number of iterations as a function on $\tau$. We plot the number of iterations as a function on the number of delay steps. The simulations are shown in Fig. 3.2.

We first note that the delays do influence the convergence rate of the two algorithms, that is, the greater the delay the more time the algorithms need to terminate. Second, as shown by the curve for DG the number of iterations seems to increase as a cubic function of the number of delay steps, which agrees with our analysis in Theorem 7. Finally, in this example, uniform delays have a bigger impact on the performance of DA, that is, DA requires more iterations to converge than DG under the same number of delay steps.

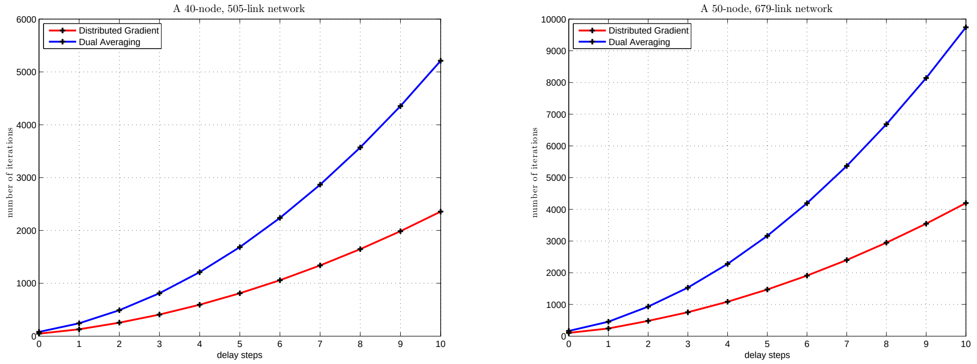

Figure 3.2: Performance of DG and DA with delays over two networks with 40 and 50 nodes on the left and the right plots, respectively.

## 3.5 Proofs for Main Results

We provide here the complete proof of the main results presented in Section 3.3. In the following lemma, we first study some important properties for the projection error $\zeta_i$ , which can be viewed as the one-dimension version of Lemma 16 for the general convex set $\mathcal{X}$, stated in Appendix B.

**Lemma 4.** *Suppose Assumptions 1 and 3 hold. Let $v_i(t)$ and $x_i(t)$ be updated by Eqs.* (3.3) *and* (3.4)*, respectively. Then*

1. *For all $t \geq 0$*

$$\left|\zeta_i(v_i(t))\right| \leq \left|\alpha(t)f_i'(x_i(t))\right| \leq C_i\alpha(t). \tag{3.17}$$

2. *Given any feasible direction $r_i$, i.e.,*

$$\begin{cases} r_i \leq 0 & \textit{if} \;\; x_i(t) = b \\ r_i \geq 0 & \textit{if} \;\; x_i(t) = a. \end{cases} \tag{3.18}$$

*We have*

$$(v_i(t) - r_i)\,\zeta_i(v_i(t)) \geq \Big[\zeta_i(v_i(t))\Big]^2. \tag{3.19}$$

*Proof.* 1. Recall that $\zeta_i(v_i(t)) = v_i(t) - \mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}}$. By Eq. (3.4) consider the following three cases.

(a) If $x_i(t) \in \mathcal{X} = (a, b)$ then $\zeta_i(v_i(t)) = v_i(t) - v_i(t) = 0$.

(b) If $x_i(t) = a$ then $0 \leq \mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}} = \max(0, v_i(t))$. If $v_i(t) \geq 0$ then $\zeta_i(v_i(t)) = 0$. Otherwise if

$$v_i(t) = -\beta a + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - \alpha(t)f_i'(x_i(t)) < 0,$$

then since $x_j(t-\tau) \in (a, b)$

$$0 \leq -\beta a + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau).$$

This gives

$$-\alpha(t)f_i'(x_i(t)) \leq -\beta a + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - \alpha(t)f_i'(x_i(t)) \leq 0.$$

Thus, since $\zeta_i(v_i(t)) = v_i(t) - \mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}}$ we obtain

$$\Big|\,\zeta_i(v_i(t))\,\Big| = \left|\,\beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) + \alpha(t)f_i'(x_i(t))\,\right| \leq \Big|\,\alpha(t)f_i'(x_i(t))\,\Big|.$$

(c) Finally, if $x_i(t) = b$ then

$$\mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}} = -v_i^-(t) = -\max(0, -v) \leq 0.$$

If $v_i(t) < 0$ then $\mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}} = v_i(t)$ implying $\zeta_i(v_i(t)) = 0$. Otherwise, if $v_i(t) \geq 0$ then $\mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}} = 0$, which implies

$$\begin{aligned}
0 &\leq -\beta x_i(t) + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - \alpha(t)f_i'(x_i(t)) \\
&= -\beta b + \beta \sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - \alpha(t)f_i'(x_i(t)) \\
&\leq \beta(b - \sum_{j\in\mathcal{N}_i} a_{ij}b) - \alpha(t)f_i'(x_i(t)) = -\alpha(t)f_i'(x_i(t)).
\end{aligned}$$

Thus we have

$$\Big|\zeta_i(v_i(t))\Big|\Bigg|-\beta x_i(t)+\beta\sum_{j\in\mathcal{N}_i}a_{ij}x_j(t-\tau)-\alpha(t)f_i'(x_i(t))\Bigg|\leq\Big|\alpha(t)f_i'(x_i(t))\Big|.$$

From these three cases, we have $\big|\zeta_i(v_i(t))\big|\leq\big|\alpha(t)f_i'(x_i(t))\big|$, which by Eq. (2.11) implies $\Big|\zeta_i(v_i(t))\Big|\leq C_i\alpha(t)$.

2. Let $r_i$ be a feasible direction, i.e., $r_i$ satisfies Eq. (3.18). Consider

$$\begin{aligned}(v_i(t)-r_i)\zeta_i(v_i(t)) &= (v_i(t)-\mathcal{P}(v_i(t))+\mathcal{P}(v_i(t))-r_i)\zeta_i(v_i(t))\\ &= \zeta_i^2(v_i(t))+(\mathcal{P}(v_i(t))-r_i(t))\zeta_i(v_i(t))\\ &= \zeta_i^2(v_i(t))+(\mathcal{P}(v_i(t))-r_i(t))(v_i(t)-\mathcal{P}(v_i(t))),\end{aligned} \tag{3.20}$$

where for convenience we define $q_i$ as

$$q_i(\mathcal{P}(v_i(t))-r_i(t))(v_i(t)-\mathcal{P}(v_i(t))).$$

We investigate the second term of the previous relation for three cases.

(a) If $x_i(t)\in\mathcal{X}=(a,b)$ then $\mathcal{P}(v_i(t))=v_i(t)$ implying $q_1=0$.

(b) If $x_i(t)=a$ then

$$0\leq\mathcal{P}_{\mathcal{T}_{\mathcal{X}(x_i(t))}}=v_i^+(t)=\max(0,v_i(t)).$$

If $v_i(t)\geq 0$ then $\mathcal{P}(v_i(t))=v_i(t)$ implying $q_i=0$. Otherwise if $v_i(t)<0$ then $\mathcal{P}(v_i(t))=0$. Since $x_i(t)=a$ we have $r_i\geq 0$, which implies $q_i\geq 0$ since $v_i(t)\leq 0$.

(c) Finally, if $x_i(t)=b$ then

$$\mathcal{P}(v_i(t))=-\max(0,-v)\leq 0.$$

If $v_i(t)<0$ then $\mathcal{P}(v_i(t))=v_i(t)$, implying $q_i=0$. Otherwise, if $v_i(t)\geq 0$ then $\mathcal{P}(v_i(t))=0$. Since $x_i(t)=b$, $r_i\leq 0$, this gives $q_1\geq 0$ due to $v_i(t)\geq 0$.

Combining these three cases and by Eq. (3.20) we have Eq. (3.19).

□

### 3.5.1 Proof of Lemma 3

*Proof.* Let $\mathbf{y}(t) = \mathbf{x}(t) - \bar{x}(t)\mathbf{1}$ and consider the following notation

$$\mathbf{g}(t) = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t)), \qquad \mathbf{h}(t) = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{v}(t)).$$

1. We first show steps $1-3$ stated in the proof sketch of Lemma 3.

   (a) Using Eqs. (3.6) and (3.8), and since $\mathbf{1}^T\mathbf{A} = \mathbf{A}\mathbf{1} = \mathbf{1}$ we have

$$\begin{aligned}\dot{\mathbf{y}}(t) &= \dot{\mathbf{x}}(t) - \dot{\bar{x}}(t)\mathbf{1}\\ &= -\beta\mathbf{x}(t) + \beta\mathbf{A}\mathbf{x}(t-\tau) + \beta\bar{x}(t)\mathbf{1} - \beta\bar{x}(t-\tau)\mathbf{1} - \alpha(t)\nabla F(\mathbf{x}(t))\\ &\qquad + \frac{\alpha(t)}{n}\mathbf{1}\mathbf{1}^T\nabla F(\mathbf{x}(t)) - \zeta(\mathbf{v}(t)) + \frac{1}{n}\mathbf{1}\mathbf{1}^T\zeta(\mathbf{v}(t))\\ &= -\beta(\mathbf{x}(t) - \bar{x}(t)\mathbf{1}) + \beta\mathbf{A}(\mathbf{x}(t-\tau) - \bar{x}(t-\tau)\mathbf{1})\\ &\qquad - \alpha(t)\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t)) - \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{v}(t))\\ &= -\beta\mathbf{y}(t) + \beta A\mathbf{y}(t-\tau) - \alpha(t)\mathbf{g}(t) - \mathbf{h}(t). \end{aligned} \tag{3.21}$$

   The preceding relation gives

$$\mathbf{y}(t) = e^{-\beta t}\mathbf{y}(0) + \int_0^t e^{-\beta(t-u)}\Big(\beta\mathbf{A}\mathbf{y}(u-\tau) - \alpha(u)\mathbf{g}(u) - \mathbf{h}(u)\Big)du. \tag{3.22}$$

   (b) Taking the 2-norm of Eq. (3.22) and using the triangle inequality gives

$$\begin{aligned}\big\|\mathbf{y}(t)\big\| &\leq e^{-\beta t}\big\|\mathbf{y}(0)\big\| + \int_0^t e^{-\beta(t-u)}\Big(\alpha(u)\big\|\mathbf{g}(u)\big\| + \big\|\mathbf{h}(u)\big\|\Big)du\\ &\qquad + \beta\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du. \end{aligned} \tag{3.23}$$

   First, using the triangle inequality and Eq. (2.11) gives

$$\begin{aligned}\big\|\mathbf{g}(t)\big\| \leq \big\|\nabla F(\mathbf{x}(t))\big\| &= \sqrt{\sum_{i=1}^n \Big[f_i'(x_i(t))\Big]^2}\\ &\overset{(2.11)}{\leq} \sqrt{\sum_{i=1}^n C_i^2} \leq C. \end{aligned} \tag{3.24}$$

Second, by Eq. (3.17) we have

$$\big\|\mathbf{h}(t)\big\| = \left\|\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{v}(t))\right\| \leq C\alpha(t).$$

Substituting the above and Eq. (3.24) into Eq. (3.23) we have

$$\begin{aligned}\big\|\mathbf{y}(t)\big\| \leq e^{-\beta t}\big\|\mathbf{y}(0)\big\| + 2C\int_0^t e^{-\beta(t-u)}\alpha(u)du \\ + \beta\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du. \qquad (3.25)\end{aligned}$$

Note that $\alpha(t)$ is non-increasing with $\alpha(0) = 1$. Consider the second term on the right-hand side of Eq. (3.25)

$$\begin{aligned}\int_0^t e^{-\beta(t-u)}\alpha(u)du &= \int_0^{t/2} e^{-\beta(t-u)}\alpha(u)du + \int_0^{t/2} e^{-\beta(t-u)}\alpha(u)du \\ &\leq \int_0^{t/2} e^{-\beta(t-u)}du + \alpha(t/2)\int_0^{t/2} e^{-\beta(t-u)}du \\ &\leq \frac{1}{\beta}e^{-\beta t/2} + \frac{\alpha(t/2)}{\beta}\cdot\end{aligned}$$

Substituting the above in Eq. (3.25) and using $\big\|\mathbf{y}(0)\big\| \leq \big\|\mathbf{x}(0)\big\|$ gives

$$\begin{aligned}\big\|\mathbf{y}(t)\big\| \leq e^{-\beta t}\big\|\mathbf{x}(0)\big\| + \frac{2C}{\beta}e^{-\beta t/2} + \frac{2C\alpha(t/2)}{\beta} \\ + \beta\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du. \qquad (3.26)\end{aligned}$$

(c) Using Eq. (2.5) in the last term of Eq. (3.26) gives

$$\int_0^t e^{-\beta(t-u)}\big\|\mathbf{A}\mathbf{y}(u-\tau)\big\|du \leq \sigma_2\int_0^t e^{-\beta(t-u)}\big\|\mathbf{y}(u-\tau)\big\|du.$$

Using $\beta \in (0,1)$ and the preceding relation in Eq. (3.26) yields

$$\begin{aligned}\big\|\mathbf{y}(t)\big\| &\leq \frac{\big\|\mathbf{x}(0)\big\|+2C}{\beta}e^{-\beta t/2} + \frac{2C\alpha(t/2)}{\beta} \\ &\quad + \beta\sigma_2\int_0^t e^{-\beta(t-u)}\big\|\mathbf{y}(u-\tau)\big\|du \\ &= \mu(t) + \beta\sigma_2\int_0^t e^{-\beta(t-u)}\big\|\mathbf{y}(u-\tau)\big\|du, \qquad (3.27)\end{aligned}$$

where $\mu(t)$ is defined as

$$\mu(t) = \frac{\|\mathbf{x}(0)\| + 2C}{\beta} e^{-\beta t/2} + \frac{2C\alpha(t/2)}{\beta}\cdot \tag{3.28}$$

We now apply a delayed version of the *Grönwall-Bellman* inequality to obtain a bound for Eq. (3.27). Define $w(t)$ be a function of $t$ as

$$w(t) = \int_0^t e^{\beta u} \|\mathbf{y}(u-\tau)\| du.$$

By Eq. (3.27) we have

$$\|\mathbf{y}(t)\| \leq \mu(t) + \beta\sigma_2 e^{-\beta t} w(t).$$

Moreover, $w(t)$ is non-decreasing and $w(0) = 0$. Consider

$$\begin{aligned}\dot{w}(t) = e^{\beta t}\|\mathbf{y}(t-\tau)\| &\leq e^{\beta t}\Big(\mu(t-\tau) + \beta\sigma_2 e^{-\beta(t-\tau)} w(t-\tau)\Big)\\ &= e^{\beta t}\mu(t-\tau) + \sigma_2\beta e^{\beta\tau} w(t-\tau) \leq e^{\beta t}\mu(t-\tau) + \sigma_2\beta e^{\beta\tau} w(t),\end{aligned}$$

where the last inequality is due to the fact that $w(t)$ is non-decreasing, i.e., $w(t) \geq w(t-\tau)$. The preceding relation implies

$$\dot{w}(t) - \sigma_2\beta e^{\beta\tau} w(t) \leq e^{\beta t}\mu(t-\tau),$$

which by multiplying both sides by $e^{-\sigma_2\beta e^{\beta\tau} t}$ gives

$$\frac{d}{dt}\left(e^{-\sigma_2\beta e^{\beta\tau} t} w(t)\right) \leq e^{-\sigma_2\beta e^{\beta\tau} t} e^{\beta t}\mu(t-\tau).$$

Taking the integral on both sides of above and using $w(0) = 0$ gives

$$w(t) \leq e^{\sigma_2\beta e^{\beta\tau} t} \int_0^t e^{\beta(1-\sigma_2 e^{\beta\tau})u} \mu(u-\tau) du. \tag{3.29}$$

Thus, since $\|\mathbf{y}(t)\| \leq \mu(t) + \beta\sigma_2 e^{-\beta t} w(t)$ and by Eq. (3.29) we have

$$\|\mathbf{y}(t)\| \leq \mu(t) + \beta\sigma_2 \int_0^t e^{-\beta(1-\sigma_2 e^{\beta\tau})(t-u)} \mu(u-\tau) du, \tag{3.30}$$

which is Eq. (3.9) since $\gamma = \sigma_2 e^{\beta\tau}$.

2. We now show Eq. (3.11). Since $\lim\limits_{t\to\infty}\alpha(t) = 0$, $\lim\limits_{t\to\infty}\mu(t) = 0$ by Eq. (3.28). Consider the second term on the right-hand side of Eq. (3.30)

$$\begin{aligned}\int_0^t e^{-\beta(1-\gamma)(t-u)}\mu(u-\tau)du = \frac{\|\mathbf{x}(0)\| + 2C}{\beta}\int_0^t e^{-\beta(1-\gamma)(t-u)}e^{-\beta(u-\tau)/2}du\\ + \frac{2C}{\beta}\int_0^t e^{-\beta(1-\gamma)(t-u)}\alpha((u-\tau)/2)du. \quad (3.31)\end{aligned}$$

First, consider the first term on the right-hand side of Eq. (3.31)

$$\begin{aligned}\lim_{t\to\infty}\int_0^t e^{-\beta(1-\gamma)(t-u)}e^{-\beta(u-\tau)/2}du &= \lim_{t\to\infty} e^{-\beta(1-\gamma)t+\beta\tau/2}\int_0^t e^{\beta(1/2-\gamma)u}du\\ &= e^{\beta\tau/2}\lim_{t\to\infty} e^{-\beta(1-\gamma)t}\frac{e^{\beta(1/2-\gamma)t}-1}{\beta(1/2-\gamma)} = 0. \quad (3.32)\end{aligned}$$

Second, consider the second term on the right-hand side of Eq. (3.31)

$$\begin{aligned}&\lim_{t\to\infty}\int_0^t e^{-\beta(1-\gamma)(t-u)}\alpha((u-\tau)/2)du\\ &= \lim_{t\to\infty}\int_0^{t/2} e^{-\beta(1-\gamma)(t-u)}\alpha((u-\tau)/2)du\\ &\qquad + \lim_{t\to\infty}\int_{t/2}^t e^{-\beta(1-\gamma)(t-u)}\alpha((u-\tau)/2)du\\ &\leq \lim_{t\to\infty}\int_0^{t/2} e^{-\beta(1-\gamma)(t-u)}du + \lim_{t\to\infty}\alpha((u-2\tau)/4)\int_{t/2}^t e^{-\beta(1-\gamma)(t-u)}du\\ &\leq \lim_{t\to\infty}\frac{e^{-\beta(1-\gamma)t/2}}{\beta(1-\gamma)} + \lim_{t\to\infty}\frac{\alpha((u-2\tau)/4)}{\beta(1-\gamma)} = 0, \quad (3.33)\end{aligned}$$

where the last equality is due to $\gamma \in (0,1)$ and $\lim\limits_{t\to\infty}\alpha(t) = 0$. Using the preceding relation and Eq. (3.32) in Eq. (3.31) we have

$$\lim_{t\to\infty}\int_0^t e^{-\beta(1-\gamma)(t-u)}\mu(u-\tau) = 0, \quad (3.34)$$

which together with $\lim\limits_{t\to\infty}\mu(t) = 0$ and Eq. (3.9) gives Eq. (3.11).

3. We now consider Eq. (3.30) where $\mu(t)$ is given in Eq. (3.28). Indeed, we provide a bound for the first term on the right-hand side of Eq. (3.30)

$$\int_0^t \alpha(u)\mu(u)du \leq \frac{\|\mathbf{x}(0)\| + 2C}{\beta}\int_0^t e^{-\beta u/2}du + \frac{2C}{\beta}\int_0^t \alpha^2(u/2)du$$

$$\leq \frac{2\|\mathbf{x}(0)\| + 4C}{\beta^2} + \frac{2C}{\beta}\int_0^t \alpha^2(u/2)du. \tag{3.35}$$

Second, consider the second term of Eq. (3.30). We first have

$$\begin{aligned}
&\int_{u=0}^t \alpha(u)\int_{s=0}^u e^{-\beta(1-\gamma)(u-s)}e^{-\beta(s-\tau)/2}dsdu \\
&\qquad \leq e^{\beta\tau/2}\int_{u=0}^t\int_{s=0}^u e^{-\beta(1-\gamma)u}e^{\beta(1-\gamma)s/2}dsdu \\
&\qquad \leq \frac{2e^{\beta\tau/2}}{\beta(1-\gamma)}\int_0^t e^{-\beta(1-\gamma)u/2}du \leq \frac{4e^{\beta\tau/2}}{\beta^2(1-\gamma)^2}.
\end{aligned} \tag{3.36}$$

Next, consider

$$\begin{aligned}
&\int_{u=0}^t \alpha(u)\int_{s=0}^u e^{-\beta(1-\gamma)(u-s)}\alpha((t-\tau)/2)dsdu \\
&\qquad \leq \int_{u=0}^t\int_{s=0}^u e^{-\beta(1-\gamma)(u-s)}\alpha^2((s-\tau)/2)dsdu \\
&\qquad = \int_{u=0}^t e^{-\beta(1-\gamma)u}\int_{s=0}^{u/2} e^{\beta(1-\gamma)s}\alpha^2((s-\tau)/2)dsdu \\
&\qquad\qquad + \int_{u=0}^t e^{-\beta(1-\gamma)u}\int_{s=u/2}^{u} e^{\beta(1-\gamma)s}\alpha^2((s-\tau)/2)dsdu \\
&\qquad \leq \int_{u=0}^t e^{-\beta(1-\gamma)u}\int_{s=0}^{u/2} e^{\beta(1-\gamma)s}dsdu \\
&\qquad\qquad + \int_{u=0}^t e^{-\beta(1-\gamma)u}\alpha^2((s-2\tau)/4)\int_{s=u/2}^{u} e^{\beta(1-\gamma)s}dsdu \\
&\qquad \leq \frac{1}{\beta(1-\gamma)}\int_u^t e^{-\beta(1-\gamma)u/2}du \\
&\qquad\qquad + \frac{1}{\beta(1-\gamma)}\int_u^t \alpha^2((s-2\tau)/4)du \\
&\qquad \leq \frac{2}{\beta^2(1-\gamma)^2} + \frac{1}{\beta(1-\gamma)}\int_0^t \alpha^2((s-2\tau)/4)du.
\end{aligned} \tag{3.37}$$

Using Eqs. (3.36) and (3.37) in the second term of Eq. (3.30) gives

$$\begin{aligned}
&\int_{u=0}^t \alpha(u)\int_{s=0}^u e^{-\beta(1-\gamma)(u-s)}\mu(s-\tau)dsdu \\
&\qquad \leq \frac{4\left(\|\mathbf{x}(0)\| + 3C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{2C}{\beta^2(1-\gamma)}\int_0^t \alpha^2(\gamma u/4-\tau)du.
\end{aligned} \tag{3.38}$$

By adding Eq. (3.38) to Eq. (3.35) and using Eq. (3.30) give

$$\begin{aligned}
&\int_0^t \alpha(u)\big\|\mathbf{y}(u)\big\|du \\
&\qquad \leq \frac{2\big\|\mathbf{x}(0)\big\| + 4C}{\beta^2} + \frac{2C}{\beta}\int_0^t \alpha^2(u/2)du \\
&\qquad\qquad + \frac{4\left(\big\|\mathbf{x}(0)\big\| + 3C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{2C}{\beta^2(1-\gamma)}\int_0^t \alpha^2(\gamma u/4 - \tau)du \\
&\qquad \leq \frac{8\left(\big\|\mathbf{x}(0)\big\| + 2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{4C}{\beta^2(1-\gamma)}\int_0^t \alpha^2(\gamma u/4 - \tau)du, \qquad (3.39)
\end{aligned}$$

where in the last inequality we use $\gamma \in (0,1)$ and $\alpha(t)$ is non-increasing, i.e., $\alpha^2(u/2) \leq \alpha^2(\gamma u/4 - \tau)$ for $\tau > 0$. This shows Eq. (3.12).

□

### 3.5.2 Proof of Theorem 7

*Proof.* Let $x^*$ be a solution of problem (2.1). Consider a candidate Krasovskii Lyapunov functional $V$ [57] defined as

$$V(\bar{x}(t)) = \frac{1}{2}(\bar{x}(t) - x^*)^2 + \frac{\beta}{2}\int_{t-\tau}^t (\bar{x}(s) - x^*)^2 ds, \quad t \geq 0. \qquad (3.40)$$

Taking the derivative of $V$ above gives

$$\begin{aligned}
\dot{V}(\bar{x}(t)) &= (\bar{x}(t) - x^*)\dot{\bar{x}} + \frac{\beta}{2}\Big[(\bar{x}(t) - x^*)^2 - (\bar{x}(t-\tau) - x^*)^2\Big] \\
&= (\bar{x}(t) - x^*)\left(-\beta\bar{x}(t) + \beta\bar{x}(t-\tau) - \frac{\alpha(t)}{n}\sum_{i=1}^n f_i'(x_i(t)) - \bar{\zeta}(t)\right) \\
&\qquad + \frac{\beta(\bar{x}(t) - x^*)^2 - \beta(\bar{x}(t-\tau) - x^*)^2}{2} \\
&= \beta(\bar{x}(t) - x^*)(\bar{x}(t-\tau) - \bar{x}(t)) - \frac{\alpha(t)}{n}\sum_{i=1}^n (\bar{x}(t) - x^*)f_i'(x_i(t)) \\
&\qquad + \frac{\beta(\bar{x}(t) - x^*)^2 - \beta(\bar{x}(t-\tau) - x^*)^2}{2} - \frac{1}{n}\sum_{i=1}^n (\bar{x}(t) - x^*)z_i(x_i(t)) \\
&= W_1 + W_2 - \frac{\beta}{2}\Big(\bar{x}(t) - \bar{x}(t-\tau)\Big)^2 \\
&\leq W_1 + W_2, \qquad (3.41)
\end{aligned}$$

where $W_1, W_2$ are defined as

$$\begin{aligned} W_1 &= -\frac{\alpha(t)}{n}\sum_{i=1}^{n}(\bar{x}(t) - x^*)f_i'(x_i(t)) \\ W_2 &= -\frac{1}{n}\sum_{i=1}^{n}(\bar{x}(t) - x^*)z_i(x_i(t)). \end{aligned}$$

We now analyze the terms $W_1$ and $W_2$. First, consider $W_1$

$$\begin{aligned} W_1 &= -\frac{\alpha(t)}{n}\sum_{i=1}^{n}(\bar{x}(t) - x_i(t) + x_i(t) - x^*)f_i'(x_i(t)) \\ &= -\frac{\alpha(t)}{n}\sum_{i=1}^{n}\Big(\bar{x}(t) - x_i(t)\Big)f_i'(x_i(t)) - \frac{\alpha(t)}{n}\sum_{i=1}^{n}\Big(x_i(t) - x^*\Big)f_i'(x_i(t)) \\ &\leq \frac{\alpha(t)}{n}\sum_{i=1}^{n}\big|\bar{x}(t) - x_i(t)\big|\,\Big|f_i'(x_i(t))\Big| - \frac{\alpha(t)}{n}\Big(F(\mathbf{x}(t)) - f^*\Big) \\ &\leq \frac{\alpha(t)C}{n}\big\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\big\| - \frac{\alpha(t)}{n}\Big(F(\mathbf{x}(t)) - f^*\Big) \\ &= \frac{\alpha(t)C}{n}\big\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\big\| - \frac{\alpha(t)}{n}\Big(F(\mathbf{x}(t)) - f(\bar{x}(t))\Big) \\ &\qquad\qquad - \frac{\alpha(t)}{n}\Big(f(\bar{x}(t)) - f^*\Big) \\ &\leq \frac{2\alpha(t)C}{n}\big\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\big\| - \frac{\alpha(t)}{n}\Big(f(\bar{x}(t)) - f^*\Big). \end{aligned} \tag{3.42}$$

Second, let $r_i(t)$ be defined as

$$r_i(t) = x^* - x_i(t) - \beta x_i(t) + \beta\sum_{j\in\mathcal{N}_i}a_{ij}x_j(t-\tau),$$

and recall from Eq. (3.18) that $r_i(t)$ is a feasible direction if

$$\begin{cases} r_i \leq 0 & \text{if } \ x_i(t) = b \\ r_i \geq 0 & \text{if } \ x_i(t) = 0. \end{cases} \tag{3.43}$$

Indeed, if $x_i(t) = 0$ then $r_i(t) \geq 0$ since $x^*, x_j(t-\tau) \in (0, b)$, for all $j \in \mathcal{V}$, and $\mathbf{A}$ is doubly stochastic. Otherwise, if $x_i(t) = b$ then $r_i(t) \leq 0$. Thus, $r_i(t)$ is a feasible direction, i.e., $r_i(t)$ satisfies Eq. (3.43). We now provide an upper bound for $W_2$. Indeed, using the definition of $r_i(t)$ above and $v_i(t)$ we consider $W_2$

$$
\begin{aligned}
W_2 &= -\frac{1}{n}\sum_{i=1}^{n}(\bar{x}(t) - x^*)\zeta_i(t)) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}(t) - (1+\beta)x_i(t) + \beta\sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - v_i(t)\right)\zeta_i(t) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(v_i(t) + (1+\beta)x_i(t) - \beta\sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - x^*\right)\zeta_i(t) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}(t) - (1+\beta)x_i(t) + \beta\sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - v_i(t)\right)\zeta_i(t) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(v_i(t) - r_i(t)\right)\zeta_i(t), \qquad (3.44)
\end{aligned}
$$

where by Eq. (3.3) the first sum is equivalent to

$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}(t) - (1+\beta)x_i(t) + \beta\sum_{j\in\mathcal{N}_i} a_{ij}x_j(t-\tau) - v_i(t)\right)\zeta_i(t) \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}(t) - x_i(t) + \alpha(t)f_i'(x_i(t))\right)\zeta_i(t) \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n}\left|\bar{x}(t) - x_i(t)\right|\left|\zeta_i(t)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\alpha(t)f_i'(x_i(t))\right|\left|\zeta_i(t)\right| \\
&\quad \overset{(3.17)}{\leq} \frac{C\alpha(t)}{n}\left\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\right\| + \frac{C^2\alpha^2(t)}{n}\cdot
\end{aligned}
$$

Since $r_i(t)$ is a feasible direction and by Eq. (3.19), we have

$$
-\frac{1}{n}\sum_{i=1}^{n}\left(v_i(t) - r_i(t)\right)\zeta_i(t) \leq -\frac{1}{n}\sum_{i=1}^{n}\zeta_i^2(t) = -\frac{1}{n}\left\|\zeta(t)\right\|^2.
$$

Applying the preceding two relations in Eq. (3.44) we obtain

$$
W_2 \leq \frac{C\alpha(t)}{n}\left\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\right\|_2 + \frac{C^2\alpha^2(t)}{n}\cdot \qquad (3.45)
$$

Thus, substituting Eqs. (3.42) and (3.45) into Eq. (3.41) we obtain

$$
\dot{V}(\bar{x}(t)) \leq \frac{3\alpha(t)C}{n}\left\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\right\| + \frac{C^2\alpha^2(t)}{n} - \frac{\alpha(t)}{n}(f(\bar{x}(t)) - f^*). \qquad (3.46)
$$

By Eq. (3.12) we have

$$\begin{aligned}
&\int_0^t \alpha(u)\big\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\big\|du \\
&\qquad \leq \frac{8\left(\big\|\mathbf{x}(0)\big\| + 2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{4C}{\beta^2(1-\gamma)}\int_0^t \alpha^2(\gamma u/4 - \tau)du\cdot
\end{aligned} \tag{3.47}$$

Using $\alpha(t) = 1$ for $t \leq 1$ and $\alpha(t) = 1/\sqrt{t}$ for $t \geq 1$ gives

$$\begin{aligned}
\int_0^t \alpha^2(\gamma u/4 - \tau)du &= \frac{4}{\gamma}\int_{-\tau}^{\frac{\gamma t}{4}-\tau} \alpha^2(u)du \\
&= \frac{4}{\gamma}\int_{-\tau}^{1} \alpha^2(u)du + \frac{4}{\gamma}\int_{1}^{\frac{\gamma t}{4}-\tau} \alpha^2(u)du = \frac{4(1+\tau)}{\gamma} + \frac{4}{\gamma}\int_{1}^{\frac{\gamma t}{4}-\tau} \frac{1}{t}du \\
&\leq \frac{4(1+\tau)}{\gamma} + \frac{4\ln(\gamma t - 4\tau)}{\gamma}\cdot
\end{aligned}$$

Substituting above into Eq. (3.47) we obtain

$$3C\int_0^t \alpha(u)\big\|\mathbf{x}(u) - \bar{x}(t)\mathbf{1}\big\|du + \frac{C^2}{n}\int_0^t \alpha^2(u)du \leq \Gamma_0(t), \tag{3.48}$$

where $\Gamma_0(t)$ is defined as

$$\begin{aligned}
\Gamma_0(t) \triangleq \frac{24C\left(\big\|\mathbf{x}(0)\big\| + 2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{48C^2(1+\tau)}{\beta^2\gamma(1-\gamma)} \\
+ C^2\ln(t) + \frac{48C^2\ln(\gamma t - 4\tau)}{\beta^2\gamma(1-\gamma)}\cdot
\end{aligned}$$

Taking the integral of both sides in Eq. (3.41) and using Eq. (3.48) gives

$$\begin{aligned}
V(\bar{x}(t)) - V(\bar{x}(0)) &\leq \frac{3C}{n}\int_0^t \alpha(u)\big\|\mathbf{x}(u) - \bar{x}(t)\mathbf{1}\big\|du + \frac{C^2}{n}\int_0^t \alpha^2(u)du \\
&\qquad - \frac{1}{n}\int_0^t \alpha(u)(F(\bar{x}(u)\mathbf{1}) - f^*)du \\
&\leq \frac{\Gamma_0(t)}{n} - \frac{1}{n}\int_0^t \alpha(u)(f(\bar{x}(u) - f^*).
\end{aligned} \tag{3.49}$$

Rearranging Eq. (3.49) and dropping $V(\bar{x}(t))$ gives

$$\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du \leq 2\Gamma_0(t) + nV(\bar{x}(0)).$$

Dividing both sides of the preceding equation by

$$\int_0^t \alpha(u)du = 1 + \int_1^t \frac{1}{\sqrt{u}}du = 2\sqrt{t} - 1$$

gives

$$\frac{\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du}{\int_0^t \alpha(u)du} \leq \frac{\Gamma_0(t) + nV(\bar{x}(0))}{2\sqrt{t} - 1},$$

which by the Jensen inequality implies

$$f\left(\frac{\int_0^t \alpha(u)\bar{x}(u)du}{\int_0^t \alpha(u)du}\right) - f^* \leq \frac{\Gamma_0(t) + nV(\bar{x}(0))}{2\sqrt{t} - 1}. \tag{3.50}$$

Moreover, we have

$$\begin{aligned} f\left(\frac{\int_0^t \alpha(u)x_i(u)du}{\int_0^t \alpha(u)du}\right) &- f\left(\frac{\int_0^t \alpha(u)\bar{x}(u)du}{\int_0^t \alpha(u)du}\right) \\ &\leq C\left|\frac{\int_0^t \alpha(u)(x_i(u) - \bar{x}(u))du}{\int_0^t \alpha(u)du}\right| \overset{(3.46)}{\leq} \frac{\Gamma_0(t)}{2\sqrt{t} - 1}. \end{aligned} \tag{3.51}$$

By Eq. (3.14) we further obtain

$$\begin{aligned} &\frac{d}{dt}(S(t)z_i(t)) = \dot{S}(t)z_i(t) + S(t)\dot{z}_i(t) = \alpha(t)x_i(t) \\ &\quad\Rightarrow z_i(t) = \frac{\int_0^t \alpha(u)x_i(u)du}{\int_0^t \alpha(u)du}, \qquad \forall i \in \mathcal{V}. \end{aligned}$$

Thus, by adding Eq. (3.50) into Eq. (3.51) and using the preceding relation of $z_i(t)$ we obtain Eq. (3.15), which concludes our proof. □

# Chapter 4

# Distributed Aggregated Stochastic Gradient Methods

## 4.1 Motivation and Contribution

In general, distributed gradient methods for solving problem (2.1) achieve sublinear convergence rates $\mathcal{O}(\ln(k)/\sqrt{k})$ and $\mathcal{O}(\ln(k)/k)$ to the optimal value for non-smooth convex and strongly convex functions, respectively. One critical assumption required by these methods is the Lipschitz continuity of the objective functions, which does not often hold in general. For example, the common quadratic function, $x^2$, is not Lipschitz continuous unless the feasible set is bounded. The second condition to guarantee for the asymptotic convergence of distributed gradient methods is a diminishing sequence of stepsizes, which, however, decreases the performance of these methods. Unlike the standard gradient method, distributed gradient methods of Eq. (2.16) do not achieve linear convergence rate for strongly convex and $L$-smooth objective functions (cf. Assumptions 5 and 6, respectively) [20].

To improve the sublinear rate of distributed gradient methods, distributed gradient tracking methods have been simultaneously studied in [20, 21]. In particular, distributed gradient tracking methods achieve the same rate as the standard gradient descent, that is, a linear rate for $L$-smooth and strongly convex functions, while relaxing the previous two critical assumptions in distributed gradient methods.

Our interest in this chapter is to consider distributed gradient tracking methods for solving problem (2.1), when there is noise in gradient estimates, named as a distributed aggregated stochastic gradient (DASG) method. For ease of exposition, in this chapter we consider problem (2.1) when $\mathbf{x}$ is a scalar, i.e., $d = 1$. In DASG, we assume that each node $i$, for all $i \in \mathcal{V}$, can only estimate noisy samples of the gradient of its function $f_i$, i.e., given a point $y \in \mathbb{R}$ each node $i$ can generate

$$g_i(y) = f_i'(y) + \xi_i(y), \tag{4.1}$$

where $\xi_i$ are independent random variables with zero mean, i.e., $\mathbb{E}[\xi_i] = 0$, implying $g_i(\cdot)$ are unbiased estimates of the derivatives $f_i'(\cdot)$. In addition, we assume that the noise-norm $\big|\xi_i(y)\big|$ is almost surely bounded, i.e., there exists a positive constant $C_i$, for all $i \in \mathcal{V}$, such that the following holds with probability 1

$$\big|\, \xi_i(y) \,\big| \leq C_i, \qquad \forall\, y \in \mathbb{R}. \tag{4.2}$$

**Main Contributions**. We will show that our method achieves a linear convergence rate in expectation to the neighborhood of the optimal solution, which depends on the noise variance of gradient estimates. The key idea of our approach is to reduce the difference between the nodes' gradient estimates at a linear rate, which is done through consensus steps following correction steps. This will help to speed up the convergence of the algorithm. Finally, we provide simulation results comparing the performance of our method with distributed stochastic gradient (DSG) methods [26] for solving linear regression problems over networks; these simulations indicate that the distributed aggregated stochastic gradient method outperforms distributed stochastic gradient methods. The results in this chapter have been appeared in [59].

## 4.2 Distributed Aggregated Stochastic Gradient Methods

DASG is formally stated in Algorithm 2 for solving problem (2.1) over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Our main motivation is based on the following key observation. Note that the local stochastic gradient step in Eq. (4.3) is not the true (global) stochastic gradient step of problem 2.1, i.e., $\sum\limits_{i\in\mathcal{V}} g_i(x)$. Therefore, the variable $y_i$ in Eq. (4.4) is used to estimate for this quantity, i.e., the goal is to reduce the difference (variance) between these two quantities. This is done by making an appropriate combination with the gradient estimate

**Algorithm 2** Distributed Aggregated Stochastic Gradient Method

```
1. Initialize: Each node $i$ initializes $x_i$ arbitrarily and $y_i(0) = g_i(x_i(0))$.
2. Iteration: For $k \geq 0$ each node $i \in \mathcal{V}$ updates
       $x_i(k+1) = \sum_{j \in \mathcal{N}_i} a_{ij} x_j(k) - \alpha y_i(k)$                       (4.3)
       $y_i(k+1) = \sum_{j \in \mathcal{N}_i} a_{ij} y_j(k) + g_i(x_i(k+1)) - g_i(x_i(k))$.     (4.4)
```

values received from the node $i$' neighbors, following a correction step, i.e., $g_i(x_i(k+1)) - g_i(x_i(k))$. We will show that DASG achieves a linear convergence rate in expectation to the neighborhood of the optimal solution.

## 4.3 Main Results

The focus of this section is to analyze the convergence rate of DASG. In particular, we show that this method achieves a linear convergence rate in expectation to the neighborhood of the solution of problem (2.1).

Let $\bar{g}(k) \triangleq (1/n)\sum_{i\in\mathcal{V}} g_i(x_i(k))$. By Eqs. (4.3) and (4.4) we have

$$\bar{x}(k+1) = \bar{x}(k) - \alpha \bar{y}(k) \tag{4.5}$$

$$\bar{y}(k+1) = \bar{y}(k) + \bar{g}(k+1) - \bar{g}(k). \tag{4.6}$$

We now explain the motivation of our analysis. The convergence analysis of DASG is composed of two steps. First, we show that the distance of $\left|x_i(k) - \bar{x}(k)\right|$ is linearly decreasing, resulting a consensus between the $x_i$. Second, we show that $\bar{x}(k)$ linearly converges to the minimizer $x^*$, implying linear convergence of $x_i(k)$ to $x^*$. These steps are built on the following fundamental inequality for Eq. (4.3), a randomized version of Lemma 2, which has been studied in [22, 26].

**Lemma 5.** *Suppose that Assumption 3 holds. Let the sequence $\{x_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Eq. (4.3). Then we have*

$$\mathbb{E}\Big[\left\|\mathbf{x}^\dagger(k)\right\|\Big] \leq \sigma_2^k \mathbb{E}\Big[\left\|\mathbf{x}^\dagger(0)\right\|\Big] + \alpha\sum_{t=0}^{k-1}\sigma_2^{k-1-t}\mathbb{E}\Big[\left\|\mathbf{y}^\dagger(t)\right\|\Big]. \tag{4.7}$$

### 4.3.1 Linear Convergence Rate of DASG

The linear convergence rate of DASG algorithm is established under the following condition on the convergence of $\left\|\mathbf{y}^{\dagger}(k)\right\|$, which we call the gradient reduction at a linear rate. To help presenting the analysis of our main result more rigorously, we discuss this condition in the next section.

**Gradient Reduction at a Linear Rate**: Let $B, D$ be positive constants, and $\gamma \in (\sigma_2, 1)$. Suppose that the sequence $\{\mathbf{y}(k)\}$ generated by Eq. (4.4) satisfies

$$\mathbb{E}\Big[\left\|\mathbf{y}^{\dagger}(k)\right\|\Big] \leq D\gamma^k + B, \quad \forall k \geq 0. \tag{4.8}$$

By Eq. (4.7), one can see that the rate of $\mathbb{E}\Big[\left\|\mathbf{x}^{\dagger}(k)\right\|\Big]$ depends on the rate of $\mathbb{E}\Big[\left\|\mathbf{y}^{\dagger}(k)\right\|\Big]$, as given in the following lemma.

**Lemma 6.** *Suppose that Assumptions 1, 3, 5, and 6 hold. Let the sequences $\{x_i(k)\}$ and $\{y_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 2. Then*

$$\mathbb{E}\Big[\left\|\mathbf{x}^{\dagger}(k)\right\|\Big] \leq \left(\frac{\mathbb{E}\Big[\left\|\mathbf{x}^{\dagger}(0)\right\|\Big] + D\alpha}{\gamma - \sigma_2}\right)\gamma^k + \frac{B\alpha}{1-\sigma_2}\cdot \tag{4.9}$$

*Proof.* Using Eq. (4.8) and $\gamma \in (\sigma_2, 1)$ into the second term of Eq. (4.7) gives

$$\sum_{t=0}^{k-1} \sigma_2^{k-1-t}\mathbb{E}\Big[\left\|\mathbf{y}^{\dagger}(t)\right\|\Big] \leq \sum_{t=0}^{k-1} \sigma_2^{k-1-t}\left(D\gamma^t + B\right) \leq \frac{D\gamma^k}{\gamma - \sigma_2} + \frac{B}{1-\sigma_2},$$

which implies Eq. (4.9) by Eq. (4.7) and $\gamma \in (\sigma_2, 1)$. □

As mentioned, $y_i(k)$ is used to estimate for $\bar{g}(k)$, that is, $\bar{y}(0) = \bar{g}(0)$ implies $\bar{y}(k+1) = \bar{y}(k) + \bar{g}(k+1) - \bar{g}(k)$. Thus we obtain

$$\bar{y}(k+1) - \bar{g}(k+1) = \bar{y}(k) - \bar{g}(k) = \ldots = \bar{y}(0) - \bar{g}(0) = 0.$$

Hence, the update in Eq. (4.4) is used to steer $y_i(k)$ to $\bar{y}(k)$, which is $\bar{g}(k)$. On the other hand, if $x_i(k)$ converges to $\bar{x}(k)$, $\bar{g}(k)$ converges to the noisy gradient estimates of problem (2.1). This is the main motivation of our analysis for the linear convergence rate of DASG. Indeed, under the condition in Eq. (4.8), we show that $\bar{x}(k)$ converges linearly to the neighborhood of $x^*$

in expectation, implying $x_i(k)$ converges linearly to the neighborhood of $x^*$ in expectation by Lemma 6 . We introduce a bit more notation as follows.

$$C \triangleq \sum_{i\in\mathcal{V}} C_i, \quad L \triangleq \sum_{i\in\mathcal{V}} L_i, \quad \mu \triangleq \sum_{i\in\mathcal{V}} \mu_i, \quad \mathbf{g}(\mathbf{x}) \triangleq \left(g_1(x_1), \ldots, g_n(x_n)\right)^T.$$

**Theorem 8.** *Suppose that Assumptions 1, 3, 5, and 6 hold. Let the sequences $\{x_i(k)\}$ and $\{y_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 2. Given $\gamma \in (\sigma_2, 1)$, let the stepsize $\alpha$ satisfy*

$$\frac{(1-\gamma^2)n(\mu+L)}{2L\mu} \leq \alpha \leq \frac{(1-\sigma_2^2)n(\mu+L)}{2(1+\tau)L\mu}, \tag{4.10}$$

*where $\tau > 0$ is the tuning parameter. Then we have*

$$\mathbb{E}\Big[|\bar{x}(k+1)-x^*|\Big] \leq \left(\mathbb{E}\Big[|\bar{x}(0)-x^*|\Big] + \frac{L\alpha\mathbb{E}\Big[\left\|\mathbf{x}^\dagger(0)\right\|\Big]}{n(\gamma-\sigma_2)}\right)\gamma^{k+1} + \frac{C\alpha}{n(1-\gamma)} + \frac{L(B+D)\alpha^2}{n(1-\sigma_2)(1-\gamma)}. \tag{4.11}$$

*Proof.* By Eq. (4.5) and since $\bar{y}(k) = (1/n)\sum_{i\in\mathcal{V}} g_i(x_i(k))$ we have

$$\begin{aligned}\bar{x}(k+1)-x^* &= \bar{x}(k)-x^* - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} g_i(x_i(k))\\ &= \bar{x}(k)-x^* - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} f_i'(\bar{x}(k)) - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} g_i(x_i(k)) - f_i'(\bar{x}(k)),\end{aligned}$$

which by Eq. (4.1) and the triangle inequality implies that

$$\begin{aligned}&\left|\bar{x}(k+1)-x^*\right|\\ &\qquad \leq \left|\bar{x}(k)-x^* - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} f_i'(\bar{x}(k))\right| + \frac{\alpha}{n}\sum_{i\in\mathcal{V}}\left|g_i(x_i(k)) - f_i'(\bar{x}(k))\right|.\end{aligned}$$

Using Eq. (4.1) into the preceding equation gives

$$\left|\bar{x}(k+1)-x^*\right| \leq \left|\bar{x}(k)-x^* - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} f_i'(\bar{x}(k))\right| + \frac{C\alpha}{n} + \frac{L\alpha}{n}\left\|\mathbf{x}^\dagger(k)\right\|. \tag{4.12}$$

Observe that the first term of Eq. (4.12) is often used in the analysis of the standard gradient method for solving problem (2.1). Thus, by Theorem 2.1.15 in [46] ( [60, Theorem 3.12]) and since $\alpha$ satisfies Eq. (4.10) we have

$$\Big|\bar{x}(k) - x^* - \frac{\alpha}{n}\sum_{i\in\mathcal{V}} f_i'(\bar{x}(k))\Big| \leq \sqrt{1-\alpha\frac{2L\mu}{n(\mu+L)}}\Big|\bar{x}(k)-x^*\Big|.$$

Note that

$$\sigma_2 < \sqrt{1-\alpha\frac{2L\mu}{n(\mu+L)}} \leq \gamma < 1.$$

Taking the expectation on both sides of Eq. (4.12) and using the previous relation we obtain

$$\mathbb{E}\Big[\big|\bar{x}(k+1)-x^*\big|\Big] \leq \gamma\mathbb{E}\Big[\big|\bar{x}(k)-x^*\big|\Big] + \frac{C\alpha}{n} + \frac{L\alpha}{n}\mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(k)\big\|\,\Big], \tag{4.13}$$

which by Eq. (4.7) implies

$$\begin{aligned}&\mathbb{E}\Big[\big|\bar{x}(k+1)-x^*\big|\Big]\\ &\quad\overset{(4.7)}{\leq} \gamma\mathbb{E}\Big[\big|\bar{x}(k)-x^*\big|\Big] + \frac{C\alpha}{n}\\ &\qquad + \frac{L\alpha}{n}\left(\sigma_2^k\mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(0)\big\|\,\Big] + \alpha\sum_{t=0}^{k-1}\sigma_2^{k-1-t}\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(t)\big\|\,\Big]\right).\end{aligned}$$

Using Eq. (4.8) into the last term on the righ-hand side of the preceding equation and iterating over $k$ further give

$$\begin{aligned}\mathbb{E}\Big[\big|\bar{x}(k+1)-x^*\big|\Big] \leq \mathbb{E}\Big[\big|\bar{x}(0)-x^*\big|\Big]\gamma^{k+1} + \frac{C\alpha}{n(1-\gamma)} + \frac{L\alpha\mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(0)\big\|\,\Big]}{n(\gamma-\sigma_2)}\gamma^{k+1}\\ + \frac{L\alpha^2}{n}\sum_{t=0}^{k}\gamma^{k-t}\sum_{\ell=0}^{t-1}\sigma_2^{t-1-\ell}(D\gamma^\ell+B).\end{aligned} \tag{4.14}$$

The last term of on the right-hand side of Eq. (4.14) is bounded by

$$\begin{aligned}\sum_{t=0}^{k}\gamma^{k-t}\sum_{\ell=0}^{t-1}\sigma_2^{t-1-\ell}(D\gamma^\ell+B) &\leq (D+B)\gamma^k\sum_{t=0}^{k}\gamma^{-t}\sum_{\ell=0}^{t-1}\sigma_2^{t-1-\ell}\\ &\leq \frac{D+B}{1-\sigma_2}\gamma^k\sum_{t=0}^{k}\gamma^{-t} \leq \frac{D+B}{(1-\gamma)(1-\sigma_2)},\end{aligned}$$

which substituting into Eq. (4.14) gives us

$$\mathbb{E}\Big[\big|\bar{x}(k+1) - x^*\big|\Big] \leq \left(\mathbb{E}\Big[\big|\bar{x}(0) - x^*\big|\Big] + \frac{L\alpha\mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(0)\big\|\,\Big]}{n(\gamma - \sigma_2)}\right)\gamma^{k+1}$$
$$+ \frac{C\alpha}{n(1-\gamma)} + \frac{L(B+D)\alpha^2}{n(1-\sigma_2)(1-\gamma)}.$$

This concludes our proof. □

**Remark 1.** *In Eq.* (4.11)*, the first term on the right-hand side is decaying linearly to zero. On the other hand, the second and the third terms are constantly depending on $\alpha$. Similarly, we have the same observation in Eq.* (4.7)*. Thus, given any accuracy level $\epsilon$, one can chose a sufficient small $\alpha$ such that $\mathbb{E}\Big[\big|x_i(k) - x^*\big|\Big] < \epsilon$ for sufficient large $k$.*

### 4.3.2 Achieving Gradient Reduction with a Linear Rate

We now show that the condition in Eq. (4.8) can be achieved by Algorithm 2. This will complete our analysis for the linear convergence rate of DASG. Note that the stepsizes considered in this section have to satisfy Eq. (4.10). We first consider the following sequence of lemmas, where their proofs are presented in Appendix C. For convenience, we introduce a bit more notation.

$$\begin{aligned}
\beta_1 &= \left(\mathbb{E}\Big[\big|\bar{x}(0) - x^*\big|\Big] + \frac{\mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(0)\big\|\,\Big]}{n(1+\tau)(\gamma - \sigma_2)}\right), & \beta_2 &= \frac{L\alpha^2}{n(\theta - \sigma_2)},\\
\beta_3 &= 2L\left(\beta_1 + \mathbb{E}\Big[\,\big\|\mathbf{x}^\dagger(0)\big\|\,\Big]\right), & \beta_4 &= 2C + \frac{2LC\alpha}{n(1-\sigma_2)},\\
\beta_5 &= \frac{2L(\beta_2 + \alpha)}{\sigma_2}, & \beta_6 &= \frac{\beta_5\theta}{\theta - \sigma_2 - L\alpha}.
\end{aligned} \tag{4.15}$$

In addition, we let $\theta$ be defined as

$$\theta = \sqrt{1 - \alpha\frac{2L\mu}{n(\mu + L)}} \in (\sigma_2, 1).$$

**Lemma 7.** *Suppose that Assumptions 1, 3, 5, and 6 hold. Let the sequences $\{x_i(k)\}$ and $y_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 2. Let $\alpha$ satisfy*

*Eq.* (4.10)*. Then for some positive constants* $\beta_1, \beta_2$ *we have*

$$\mathbb{E}\Big[\big\|\bar{x}(k+1)\mathbf{1} - x^*\mathbf{1}\big\|\Big] + \mathbb{E}\Big[\big\|\bar{x}(k)\mathbf{1} - x^*\mathbf{1}\big\|\Big] \\ \leq 2\beta_1\theta^k + \frac{2C\alpha}{n(1-\sigma_2)} + \frac{2\beta_2}{\theta}\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(t)\big\|\,\Big]. \tag{4.16}$$

**Lemma 8.** *Suppose that Assumptions 1, 3, 5, and 6 hold. Let the sequences* $\{x_i(k)\}$ *and* $y_i(k)\}$*, for all* $i \in \mathcal{V}$*, be generated by Algorithm 2. Let* $\alpha$ *satisfy Eq.* (4.10)*. Then for some positive constant* $\beta_3, \beta_4$ *we have*

$$\mathbb{E}\Big[\big\|\mathbf{g}(\mathbf{x}(k+1)) - \mathbf{g}(\mathbf{x}(k))\big\|\Big] \leq \beta_3\theta^k + \beta_4 + L\alpha\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k)\big\|\,\Big] \\ + \beta_5\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(t)\big\|\,\Big]. \tag{4.17}$$

We now show our main result in this section.

**Theorem 9.** *Suppose that Assumptions 1, 3, 5, and 6 hold. Let the sequences* $\{x_i(k)\}$ *and* $y_i(k)\}$*, for all* $i \in \mathcal{V}$*, be generated by Algorithm 2. In addition, given some constants* $\gamma \in [\theta, 1)$ *and* $\tau > 0$*, let* $\alpha$ *satisfy*

$$\alpha \in \left[\frac{(1-\gamma^2)(\mu+L)}{2L\mu}\,,\,\frac{(1-\sigma_2^2)}{2(1+\tau)L}\right]. \tag{4.18}$$

*Then, for some positive constants* $D$ *and* $B$ *we have*

$$\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k)\big\|\,\Big] \leq D\gamma^k + B. \tag{4.19}$$

*Proof.* First, the stepsize $\alpha$ given by Eq. (4.18) also satisfies the condition in Eq. (4.10), implying $\theta \in (\sigma_2, 1)$. Second, one can further show that $\sigma_2 + L\alpha < \theta < 1$ with some proper choice of $\tau$. Third, using Eq. (4.17) gives

$$\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k+1)\big\|\,\Big] \\ \leq \sigma_2\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k)\big\|\,\Big] + \mathbb{E}\Big[\,\|\mathbf{g}(\mathbf{x}(k+1)) - \mathbf{g}(\mathbf{x}(k))\mathbf{1}\|\,\Big] \\ \overset{(4.17)}{\leq} \sigma_2\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k)\big\|\,\Big] + \beta_3\theta^k + \beta_4 \\ + L\alpha\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(k)\big\|\,\Big] + \beta_5\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\,\big\|\mathbf{y}^\dagger(t)\big\|\,\Big]$$

$$
\begin{aligned}
&= (\sigma_2 + L\alpha)\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(k)\|\,\Big] + \beta_3\theta^k + \beta_4 + \beta_5\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(t)\|\,\Big]\\
&\leq (\sigma_2 + L\alpha)^{k+1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(0)\|\,\Big] + \frac{\beta_4}{1-\sigma_2-L\alpha} + \beta_3\sum_{t=0}^{k}(\sigma_2+L\alpha)^{k-t}\theta^t\\
&\qquad + \beta_5\sum_{t=0}^{k}(\sigma_2+L\alpha)^{k-t}\sum_{\ell=0}^{t-1}\theta^{t-\ell}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(\ell)\|\,\Big]\\
&\leq (\sigma_2 + L\alpha)^{k+1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(0)\|\,\Big] + \frac{\beta_4}{1-L\alpha-\sigma_2} + \frac{\beta_3}{\theta-\sigma_2-L\alpha}\\
&\qquad + \beta_6\sum_{t=0}^{k-1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(t)\|\,\Big]\theta^{k-t}, \qquad (4.20)
\end{aligned}
$$

where in the last inequality we use $\beta_6$ in Eq. (4.15) and

$$
\begin{aligned}
&\sum_{t=0}^{k}(\sigma_2+L\alpha)^{k-t}\sum_{\ell=0}^{t-1}\theta^{t-\ell}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(\ell)\|\,\Big]\\
&\qquad = (\sigma_2+L\alpha)^k\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(\ell)\|\,\Big]\theta^{-\ell}\sum_{t=\ell+1}^{k}\left(\frac{\theta}{\sigma_2+L\alpha}\right)^t\\
&\qquad \leq \frac{\theta}{\theta-\sigma_2-L\alpha}\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(\ell)\|\,\Big]\theta^{k-\ell}.
\end{aligned}
$$

Let $h(k-1) = \sum_{t=0}^{k-1}\theta^{-t}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(t)\|\,\Big]$. Then $h(k)$ is a non-decreasing non-negative function with $h(-1) = 0$. In addition, using Eq. (4.20) gives

$$
\begin{aligned}
\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(k+1)\|\,\Big] &\leq (\sigma_2+L\alpha)^{k+1}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(0)\|\,\Big]\\
&\qquad + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha} + \beta_6\theta^k h(k-1). \qquad (4.21)
\end{aligned}
$$

We now provide an upper bound for $h(k)$, i.e., consider

$$
h(k) - h(k-1) = \theta^{-k}\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(k)\|\,\Big],
$$

which by Eq. (4.21) gives

$$
\begin{aligned}
&h(k) - h(k-1)\\
&\qquad \leq \theta^{-k}\left((\sigma_2+L\alpha)^k\mathbb{E}\Big[\,\|\mathbf{y}^\dagger(0)\|\,\Big] + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha} + \beta_6\theta^{k-1}h(k-2)\right).
\end{aligned}
$$

Since $h(k)$ is non-decreasing and $\sigma_2 + L\alpha < \theta$, we have from above

$$h(k) \leq \mathbb{E}\Big[\,\|\mathbf{y}^{\dagger}(0)\|\,\Big] + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha}\theta^{-k} + \left(1+\frac{\beta_6}{\theta}\right)h(k-1).$$

Using the preceding relation, $h(-1) = 0$, and $\beta_6$ in Eq. (4.15) we obtain

$$\begin{aligned} h(k) &\leq \mathbb{E}\Big[\,\|\mathbf{y}^{\dagger}(0)\|\,\Big]\sum_{t=0}^{k-1}\left(1+\frac{\beta_5}{\theta-\sigma_2-L\alpha}\right)^t \\ &\quad + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha}\sum_{t=0}^{k-1}\theta^{-k+t}\left(1+\frac{\beta_5}{\theta-\sigma_2-L\alpha}\right)^t \\ &= \mathbb{E}\Big[\,\|\mathbf{y}^{\dagger}(0)\|\,\Big]\sum_{t=0}^{k-1}\eta^t + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha}\sum_{t=0}^{k-1}\theta^{-k+t}\eta^t \\ &= \frac{\mathbb{E}\Big[\,\|\mathbf{y}^{\dagger}(0)\|\,\Big]}{\eta-1}\eta^k + \frac{\beta_3+\beta_4}{\theta-\sigma_2-L\alpha}\theta^{-k}\frac{1-(\eta\theta)^k}{1-\eta\theta}, \end{aligned}$$

where $\eta$ is defined as

$$\eta = \left(1+\frac{\beta_5}{\theta-\sigma_2-L\alpha}\right).$$

Note that using Eq. (4.18) and $\beta_5$ in Eq. (4.15) we can show that

$$\theta\eta = \theta\left(1+\frac{\beta_5}{\theta-\sigma_2-L\alpha}\right) < 1.$$

Thus, using the previous relation into Eq. (4.21) we obtain Eq. (4.19), i.e., for some constants $B, D$ and $\gamma \in [\theta, 1]$ we obtain

$$\mathbb{E}\Big[\,\|\mathbf{y}^{\dagger}(k+1)\|\,\Big] \leq B\gamma^{k+1} + D.$$

□

**Remark 2.** *We note that Eq.* (4.18) *is well-defined, i.e., as* $\gamma$ *goes to* 1*, the lower bound goes to* 0 *while the upper bound is strictly greater than* 0*.*

## 4.4 Simulations

In this section, we compare the performance of DASG with distributed stochastic gradient (DSG) methods [26] for solving the linear regression problems, studied in Section 3.4. For both algorithms, we choose the same stepsizes $\alpha$, which satisfy Eq. (4.18). We simulate the two algorithms for two different sizes of the networks, namely, $n = 70$, and $n = 80$. In addition, we consider the problem when $d = 30$. We implement DSG and DASG for each network when the number of iteration is fixed at 400. A plot for the decaying of the error $\max_i \mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|\right]$ is shown in Fig. 4.1. As shown in the plots in Fig. 4.1, the performance of DASG is significantly better than DSG.

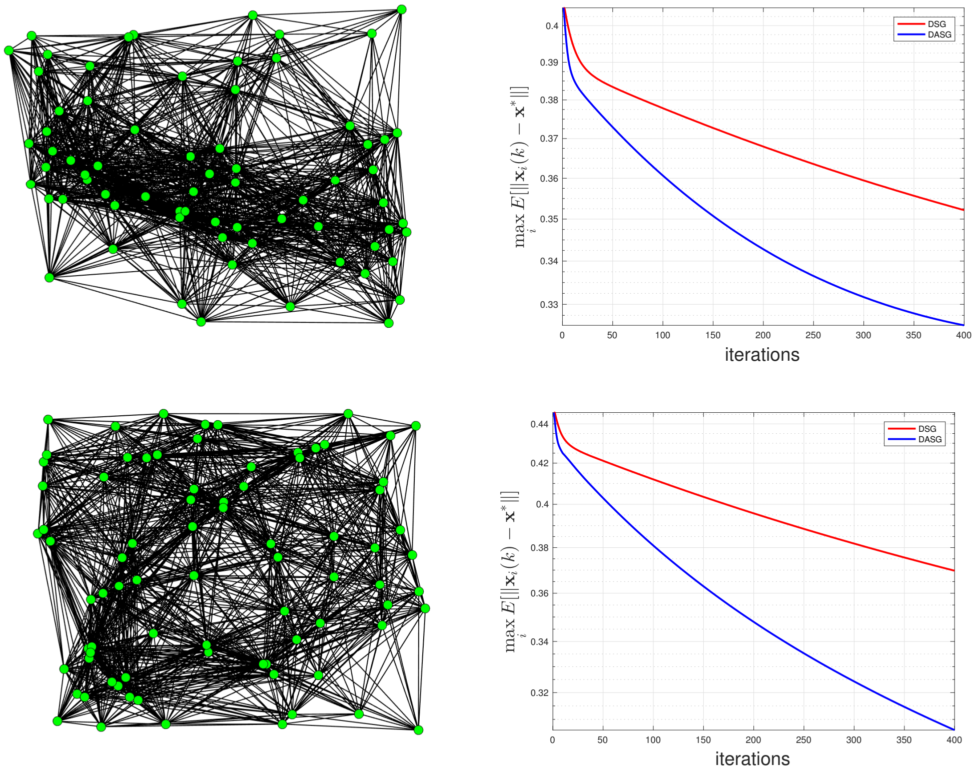

Figure 4.1: Performance of DSG and DASG methods over two networks with 70 and 80 nodes on the top and the bottom plots, respectively.

# Chapter 5

# Distributed Mirror Descent Methods

## 5.1 Motivation and Contribution

In previous chapters, we have studied distributed gradient methods for solving problem (2.1), where the performance of such methods are restricted to the Euclidean space. In this chapter, we consider mirror descent methods, which have been observed to have better performance than gradient descent methods. The method of Mirror Descent (MD), originally proposed by Nemirovski and Yudin [61], is a primal-dual method for solving constrained convex optimization problems. MD is fundamentally a (sub)gradient descent (GD) algorithm that exploits the geometry of problems through utilizing Bregman distances [62]. This method not only generalizes the standard (GD) method, but also achieves a better convergence rate. In addition, MD is applicable to optimization problems in Banach spaces where GD is not [60].

Mirror descent methods have been recently shown to be useful for efficiently solving large-scale optimization problems. In general, GD algorithms are simple to implement and achieve convergence rates, which are independent of the problem dimension under the Euclidean norm. However, such dimension-free convergence rates may not hold under other norms [60]. Alternatively, MD can bypass such limitation of GD, potentially improving its convergence; see [63] for an early example. Because of these notable benefits, MD has experienced significant recent attention for applications to large-scale optimization and machine learning problems in both the continuous and discrete time settings [64,65], both the deterministic and stochastic scenarios [66–69], and both the centralized and distributed contexts [68,70]. MD has also been applied to a variety of practical problems, e.g., game-theoretic applications [71], and multi-agent distributed learning problems [72–76].

Most of existing studies have focused on studying the convergence rate of

MD. In particular, if the stepsizes are properly selected then MD can achieve a convergence rate of $\mathcal{O}(1/k)$ or $\mathcal{O}(1/\sqrt{k})$ for strongly convex or convex objective functions, respectively, [61, 67]. However, the convergence of the objective function does not, in general, imply the convergence of iterates to an optimizer.[1] To the best of our knowledge, there has not been any prior work establishing the convergence of these variables to an optimizer. Our focus, therefore, is to provide such convergence analysis of MD.

We are motiaved by potential applications in Distributed Lagrangian (DL) methods and Game Theory where the convergence of iterates is required. Specifically, in the context of DL methods studied in Chapter 7, we can apply distributed subgradient methods, or preferably distributed MD methods, to find the solution to the dual problem. In this setting, convergence to the dual optimizer is needed to complete the convergence analysis of DL methods [11, 77]. To motivate our study from a game theoretic viewpoint, note that the dynamics of certain natural learning strategies in routing games have been identified as the dynamics of centralized mirror descent in the strategy space of the players; see [78] for an example. In that context, convergence of the learning dynamics to the Nash equilibria (the minimizers of a convex potential function of the routing game) is critical; convergence to the optimal function value is not enough.

**Main Contribution**. In this chapter, our main contribution is to show the asymptotic convergence of iterates to an optimizer in MD method for solving problem (2.1), where the objective function is convex and not necessarily differentiable. For the ease of exposition, we start our analysis with the centralized setting, which allows us to analyze the convergence of distributed MD methods. Finally, we provide simulations to show that MD outperforms GD for solving robust linear regression problems over simplex. The results in this chapter were presented in [79].

[1] One can only show the convergence of the sequence of these variables to the optimal set when this set is bounded.

## 5.2 Centralized Mirror Descent

The focus of this section is to study the convergence of centralized MD for solving problem (2.1), i.e., we consider the following minimization problem

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\ f(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(\mathbf{x}),$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set. We denote by $\Big\langle \mathbf{x}\, ,\, \mathbf{y} \Big\rangle$ the inner product of $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. In addition, let $\|\cdot\|$ be the norm induced by the inner product. In addition, we denote by $\|\cdot\|_*$ the dual norm of $\|\cdot\|$.

In MD, we consider a continuously differentiable $\mu$-strongly convex function with the induced norm, i.e. given $\|\mathbf{y}-\mathbf{x}\|^2 = \Big\langle \mathbf{x}\, ,\, \mathbf{y} \Big\rangle$, $\psi$ satisfies

$$\psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \Big\langle \psi(\mathbf{x})\, ,\, \mathbf{y}-\mathbf{x} \Big\rangle + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

Define the *Bregman divergence* $D_\psi(\cdot,\cdot)$ associated with $\psi$ as

$$D_\psi(\mathbf{y},\mathbf{x}) = \psi(\mathbf{y}) - \psi(\mathbf{x}) - \Big\langle \nabla\psi(\mathbf{x})\, ,\, \mathbf{y}-\mathbf{x} \Big\rangle, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{5.1}$$

The following two properties of Bregman divergence will be useful for our analysis given shortly, which are straightforward to derive from Eq. (5.1)

$$D_\psi(\mathbf{y},\mathbf{x}) - D_\psi(\mathbf{y},\mathbf{z}) - D_\psi(\mathbf{z},\mathbf{x}) = \Big\langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{x})\, ,\, \mathbf{y}-\mathbf{z} \Big\rangle \tag{5.2a}$$

$$D_\psi(\mathbf{z},\mathbf{x}) \geq \frac{\mu}{2}\left\|\mathbf{z}-\mathbf{x}\right\|^2, \tag{5.2b}$$

for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{X}$. Intuitively, MD iteratively minimizes the local linearization of $f$ regularized by $D_\psi$. In particular, MD updates the variable $\mathbf{x}$ as follows:

$$\mathbf{x}(k+1) = \arg\min_{\mathbf{z}\in\mathcal{X}} \left\{ \Big\langle \nabla f(\mathbf{x}(k))\, ,\, \mathbf{z}-\mathbf{x}(k) \Big\rangle + \frac{1}{\alpha(k)} D_\psi(\mathbf{z},\mathbf{x}(k)) \right\}, \tag{5.3}$$

where $\mathbf{x}$ is initialized arbitrarily in $\mathcal{X}$. Note that MD achieves a convergence rate $\mathcal{O}\left(1/\sqrt{k}\right)$ to the optimal value for $\alpha(k) = 1/\sqrt{k}$ [60, 61], i.e.,

$$f\left(\frac{1}{k}\sum_{t=1}^{k}\mathbf{x}(t)\right) - f^* \leq \mathcal{O}\left(\frac{1}{\sqrt{k}}\right),$$

where $f^*$ is the optimal value of $f$ over $\mathcal{X}$. Our interest is the asymptotic convergence of the problem variables themselves, i.e., whether $\mathbf{x}(k)$ converges to a minimizer for a suitable choice of stepsizes $\alpha(k)$. In the remainder of this section, we prove such a convergence result for centralized $\mathsf{MD}$, and extend this to a distributed setting in the next section.

**Theorem 10.** *Suppose that Assumption 7 holds. Let the sequence $\{\mathbf{x}(k)\}$ be generated by* $\mathsf{MD}$ *in Eq. (5.3). Let $\{\alpha(k)\}$ be the non-increasing positive sequence such that*

$$\sum_{k=0}^{\infty} \alpha(k) = \infty \qquad \textit{and} \qquad \sum_{k=0}^{\infty} \alpha^2(k) < \infty.$$

*Then for some minimizer $\mathbf{x}^*$ of problem (2.1) we have*

$$\lim_{k\to\infty} \mathbf{x}(k) = \mathbf{x}^*.$$

*Proof.* Let $C = \sum_{i=1}^{n} C_i$ where $C_i$ is given in Eq. (2.10). Our proof proceeds in two steps.

1. We first show that $\mathbf{x}(k)$ satisfies

$$\begin{aligned} &D_\psi(\mathbf{z}, \mathbf{x}(k+1)) - D_\psi(\mathbf{z}, \mathbf{x}(k)) \\ &\qquad \leq \alpha(k) \Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{z} - \mathbf{x}(k) \Big\rangle + \frac{\alpha^2(k) L^2}{2\mu}, \end{aligned} \tag{5.4}$$

for each $\mathbf{z} \in \mathcal{X}$. Indeed, the optimality of $\mathbf{x}(k+1)$ in Eq. (5.3) implies

$$0 \leq \Big\langle \alpha(k) \nabla f(\mathbf{x}(k)) + \nabla_1 D_\psi(\mathbf{x}(k+1), \mathbf{x}(k))\,,\, \mathbf{z} - \mathbf{x}(k+1) \Big\rangle, \tag{5.5}$$

where $\nabla_1 D_\psi$ denotes the gradient of $D_\psi$ with respect to the first coordinate. The properties of the divergence in Eqs. (5.2a) and (5.2b) yield

$$\begin{aligned} &\Big\langle \nabla_1 D_\psi(\mathbf{x}(k+1), \mathbf{x}(k))\,,\, \mathbf{z} - \mathbf{x}(k+1) \Big\rangle \\ &= \Big\langle \nabla \psi(\mathbf{x}(k+1)) - \nabla \psi(\mathbf{x}(k))\,,\, \mathbf{z} - \mathbf{x}(k+1) \Big\rangle \\ &\overset{(5.2a)}{=} D_\psi(\mathbf{z}, \mathbf{x}(k)) - D_\psi(\mathbf{z}, \mathbf{x}(k+1)) - D_\psi(\mathbf{x}(k), \mathbf{x}(k+1)) \\ &\overset{(5.2b)}{\leq} D_\psi(\mathbf{z}, \mathbf{x}(k)) - D_\psi(\mathbf{z}, \mathbf{x}(k+1)) - \frac{\mu}{2} \left\| \mathbf{x}(k+1) - \mathbf{x}(k) \right\|^2. \end{aligned}$$

Substituting the above relation in Eq. (5.5), we get

$$\begin{aligned}
&D_\psi(\mathbf{z},\mathbf{x}(k+1)) - D_\psi(\mathbf{z},\mathbf{x}(k)) \\
&\le \alpha(k)\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{z}-\mathbf{x}(k+1)\Big\rangle - \frac{\mu}{2}\left\|\mathbf{x}(k+1)-\mathbf{x}(k)\right\|^2 \\
&= \alpha(k)\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{z}-\mathbf{x}(k)\Big\rangle - \frac{\mu}{2}\left\|\mathbf{x}(k+1)-\mathbf{x}(k)\right\|^2 \\
&\qquad + \alpha(k)\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{x}(k)-\mathbf{x}(k+1)\Big\rangle. \qquad (5.6)
\end{aligned}$$

By the Cauchy-Schwarz inequality, we can upper bound the last term on the right-hand side of Eq. (5.6) as

$$\begin{aligned}
&\Big\langle \alpha(k)\nabla f(\mathbf{x}(k))\,,\, \mathbf{x}(k)-\mathbf{x}(k+1)\Big\rangle \\
&\qquad \le \frac{\alpha^2(k)}{2\mu}\left\|\nabla f(\mathbf{x}(k))\right\|^2 + \frac{\mu}{2}\left\|\mathbf{x}(k+1)-\mathbf{x}(k)\right\|^2.
\end{aligned}$$

Using the preceding relation into Eq. (5.6) we obtain Eq. (5.4), i.e.,

$$\begin{aligned}
&D_\psi(\mathbf{z},\mathbf{x}(k+1)) - D_\psi(\mathbf{z},\mathbf{x}(k)) \\
&\le \alpha(k)\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{z}-\mathbf{x}(k)\Big\rangle + \frac{\alpha^2(k)}{2\mu}\left\|\nabla f(\mathbf{x}(k))\right\|_*^2 \\
&\le \alpha(k)\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{z}-\mathbf{x}(k)\Big\rangle + \frac{C^2\alpha^2(k)}{2\mu},
\end{aligned}$$

where the last inequality is due to the Lipschitz continuity of $f_i$.

2. We now show the convergence of $\mathbf{x}(k)$ through utilizing Eq. (5.4). Let $\mathbf{x}^*$ be an optimizer of $f$ over $\mathcal{X}$. Then, the convexity of $f$ implies

$$\Big\langle \nabla f(\mathbf{x}(k))\,,\, \mathbf{x}^*-\mathbf{x}(k)\Big\rangle \le f^* - f(\mathbf{x}(k)).$$

Using the above relation in Eq. (5.4) with $\mathbf{z}=\mathbf{x}^*$ gives

$$D_\psi(\mathbf{x}^*,\mathbf{x}(k+1)) \le D_\psi(\mathbf{x}^*,\mathbf{x}(k)) - \alpha(k)\Big(f(\mathbf{x}(k)) - f^*\Big) + \frac{C^2\alpha^2(k)}{2\mu}\cdot \qquad (5.7)$$

Let $V(k)$ be defined as

$$V(k) = D_\psi(\mathbf{x}^*,\mathbf{x}(k)) + \frac{C^2}{2\mu}\sum_{t=k}^{\infty}\alpha^2(t).$$

First, due to the square summability of $\alpha(k)$ we have $V(0)$ is bounded. Adding $\frac{C^2}{2\mu}\sum_{t=k+1}^{\infty}\alpha^2(t)$ to both sides of Eq. (5.7) we obtain

$$V(k+1) \leq V(k),$$

implying the sequence $\{V(k)\}$ is non-increasing and bounded. Thus we have $V(k)$ is convergent, which implies that $D_\psi(\mathbf{x}^*, \mathbf{x}(k))$ is convergent for every solution $\mathbf{x}^*$ of problem (2.1). Second, summing both sides of Eq. (5.7) over $k$ from 0 to $K$, we get

$$\begin{aligned} D_\psi(\mathbf{x}^*, \mathbf{x}(K+1)) + \sum_{k=0}^{K}\alpha(k)\Big(f(\mathbf{x}(k)) - f^*\Big) \\ \leq D_\psi(\mathbf{x}^*, \mathbf{x}(0)) + \frac{C^2}{2\mu}\sum_{k=0}^{K}\alpha^2(k), \end{aligned}$$

which by letting $K \to \infty$ and using the square summability of $\alpha(k)$ gives

$$\sum_{k=0}^{\infty}\alpha(k)\Big[f(\mathbf{x}(k)) - f^*\Big] < \infty.$$

The non-summability of $\alpha(k)$ further yields

$$\liminf_{k\to\infty} f(\mathbf{x}(k)) = f^*.$$

The convergence of $D_\psi(\mathbf{x}^*, \mathbf{x}(k))$ for each $\mathbf{x}^*$ implies the boundedness of $\mathbf{x}(k)$. Let $\{\mathbf{x}(k_\ell)\}$ be the bounded subsequence of $\mathbf{x}(k)$ such that

$$\lim_{k_\ell\to\infty} f(\mathbf{x}(k_\ell)) = \liminf_{k\to\infty} f(\mathbf{x}(k)) = f^*. \tag{5.8}$$

This bounded sequence $\{\mathbf{x}(k_\ell)\}$ has a convergent subsequence. By Eq. (5.8) and the continuity of $f$, this subsequence converges to a point in $\mathcal{X}^*$. Call this point $\mathbf{x}^*$. Since $D_\psi(\mathbf{x}^*, \mathbf{x}(k))$ converges we get

$$\lim_{k\to\infty} D_\psi(\mathbf{x}^*, \mathbf{x}(k)) = 0,$$

implying $\lim\limits_{k\to\infty}\mathbf{x}(k) = \mathbf{x}^*$ due to Eq. (5.2b). This completes our proof.

□

## 5.3 Distributed Mirror Descent

In this section, we consider a distributed variant of MD for solving problem (2.1) over $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In particular, each node $i$ maintains a local copy $\mathbf{x}_i$ of $\mathbf{x}^*$, a solution of problem (2.1). The nodes then update their variables as

$$\begin{aligned} \mathbf{v}_i(k) &= \sum_{j \in \mathcal{N}_i} a_{ij} \mathbf{x}_j(k) \\ \mathbf{x}_i(k+1) &= \arg\min_{\mathbf{z} \in \mathcal{X}} \left\{ \Big\langle \nabla f(\mathbf{v}_i(k))\, ,\, \mathbf{z} - \mathbf{v}_i(k) \Big\rangle + \frac{1}{\alpha(k)} D_\psi(\mathbf{z}, \mathbf{v}_i(k)) \right\}, \end{aligned} \tag{5.9}$$

where $\mathbf{A} = [a_{ij}]$ satisfies Assumption 3. As mentioned, our focus is to show the asymptotic convergence of the iterates in the above *distributed mirror descent* (DMD) algorithm to a common optimizer $\mathbf{x}^*$ of $f$ over $\mathcal{X}$.

**Theorem 11.** *Suppose that Assumptions 1, 3, and 7 hold. Let the sequence* $\{\mathbf{x}_i(k)\}$, *for all* $i \in \mathcal{V}$, *be generated by* DMD *in Eq.* (5.9). *Let* $\{\alpha(k)\}$ *be the non-increasing positive sequence of stepsizes with* $\alpha(0) = 1$ *such that*

$$\sum_{k=0}^{\infty} \alpha(k) = \infty \qquad \textit{and} \qquad \sum_{k=0}^{\infty} \alpha^2(k) < \infty.$$

*In addition, we assume that* $D_\psi(\mathbf{x}, \mathbf{y})$ *is convex on* $\mathbf{y}$ *for fixed* $\mathbf{x}$. *Then for some minimizer* $\mathbf{x}^*$ *of problem* (2.1) *we get*

$$\lim_{k \to \infty} \mathbf{x}_i(k) = \mathbf{x}^*, \qquad \textit{for all } i \in \mathcal{V}.$$

*Proof.* By the optimality of $\mathbf{x}_i(k+1)$ in Eq. (5.9) we get

$$\Big\langle \alpha(k) \nabla f_i(\mathbf{v}_i(k)) + \nabla_1 D_\psi(\mathbf{x}_i(k+1), \mathbf{v}_i(k))\, ,\, \mathbf{z} - \mathbf{x}_i(k+1) \Big\rangle \geq 0, \tag{5.10}$$

for every $\mathbf{z} \in \mathcal{X}$. The property of Bregman divergence in Eq. (5.2a) yields

$$\begin{aligned} &\Big\langle \nabla_1 D_\psi(\mathbf{x}_i(k+1), \mathbf{v}_i(k))\, ,\, \mathbf{z} - \mathbf{x}_i(k+1) \Big\rangle \\ &\quad = \Big\langle \nabla \psi(\mathbf{x}_i(k+1)) - \nabla \psi(\mathbf{v}_k)\, ,\, \mathbf{z} - \mathbf{x}_i(k+1) \Big\rangle \\ &\quad = D_\psi(\mathbf{z}, \mathbf{v}_i(k)) - D_\psi(\mathbf{z}, \mathbf{x}_i(k+1)) - D_\psi(\mathbf{v}_i(k), \mathbf{x}_i(k+1)). \end{aligned} \tag{5.11}$$

Substituting the above equality into Eq. (5.10) and summing over $i \in \mathcal{V}$ give

$$\sum_{i\in\mathcal{V}} S_k^i(\mathbf{z}) + \sum_{i\in\mathcal{V}} T_k^i(\mathbf{z}) \geq 0, \tag{5.12}$$

where for each $\mathbf{z} \in \mathcal{X}$

$$\begin{aligned} S_k^i(\mathbf{z}) &:= \Big\langle \alpha(k)\nabla f_i(\mathbf{v}_i(k))\,,\, \mathbf{z} - \mathbf{x}_i(k+1)\Big\rangle \\ T_k^i(\mathbf{z}) &:= D_\psi(\mathbf{z}, \mathbf{v}_i(k)) - D_\psi(\mathbf{z}, \mathbf{x}_i(k+1)) - D_\psi(\mathbf{x}_i(k+1), \mathbf{v}_i(k)). \end{aligned}$$

We provide bounds on each term above separately.

1. We first consider $\sum_{i\in\mathcal{V}} S_k^i(\mathbf{z})$

$$\begin{aligned} \sum_{i\in\mathcal{V}} S_k^i(\mathbf{z}) = \sum_{i\in\mathcal{V}} &\Big\langle \alpha(k)\nabla f_i(\mathbf{v}_i(k))\,,\, \mathbf{z} - \mathbf{v}_i(k)\Big\rangle \\ &+ \sum_{i\in\mathcal{V}} \Big\langle \alpha(k)\nabla f_i(\mathbf{v}_i(k))\,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1)\Big\rangle. \end{aligned}$$

First, using the convexity and Lipschitz continuity of $f_i$ we get

$$\begin{aligned} \Big\langle \nabla f_i(\mathbf{v}_i(k))\,,\, \mathbf{z} - \mathbf{v}_i(k)\Big\rangle &\leq f_i(\mathbf{z}) - f_i(\mathbf{v}_i(k)) \\ &= f_i(\mathbf{z}) - f_i(\bar{\mathbf{x}}(k)) + f_i(\bar{\mathbf{x}}(k)) - f_i(\mathbf{v}_i(k)) \\ &\leq f_i(\mathbf{z}) - f_i(\bar{\mathbf{x}}(k)) + \Big\langle \nabla f_i(\mathbf{v}_i(k)), \bar{\mathbf{x}}(k) - \mathbf{v}_i(k)\Big\rangle \\ &\leq f_i(\mathbf{z}) - f_i(\bar{\mathbf{x}}(k)) + C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{v}_i(k)\right\|. \end{aligned} \tag{5.13}$$

Second, by the Cauchy-Schwarz inequality we have

$$\begin{aligned} &\Big\langle \alpha(k)\nabla f_i(\mathbf{v}_i(k))\,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1)\Big\rangle \\ &\leq \frac{\alpha^2(k)}{2\mu}\left\|\nabla f_i(\mathbf{v}_i(k))\right\|_*^2 + \frac{\mu}{2}\left\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\right\|^2 \\ &\leq \frac{\alpha^2(k)C_i^2}{2\mu} + \frac{\mu}{2}\left\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\right\|^2. \end{aligned} \tag{5.14}$$

Third, the double stochasticity of $\mathbf{A}$ and the convexity of $\|\cdot\|$ gives

$$\sum_{i\in\mathcal{V}} C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{v}_i(k)\right\| \leq \sum_{i\in\mathcal{V}} C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{x}_i(k)\right\|.$$

Combining the preceding equation with Eqs. (5.13) and (5.14) we obtain

$$\begin{aligned}\sum_{i\in\mathcal{V}} S_k^i(\mathbf{z}) &\le \alpha(k)\big[f(\mathbf{z}) - f(\bar{\mathbf{x}}(k))\big] + \alpha(k)\sum_{i\in\mathcal{V}} C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{v}_i(k)\right\| \\ &\quad + \frac{C^2}{2\mu}\alpha^2(k) + \frac{\mu}{2}\sum_{i\in\mathcal{V}} \left\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\right\|^2 \\ &\le \alpha(k)\big[f(\mathbf{z}) - f(\bar{\mathbf{x}}(k))\big] + \alpha(k)\sum_{i\in\mathcal{V}} C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{x}_i(k)\right\| \\ &\quad + \frac{C^2}{2\mu}\alpha^2(k) + \frac{\mu}{2}\sum_{i\in\mathcal{V}} \left\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\right\|^2. \end{aligned} \tag{5.15}$$

2. Second, we provied an upper bound for $\sum_{i\in\mathcal{V}} T_k^i(\mathbf{z})$

$$\begin{aligned}\sum_{i\in\mathcal{V}} T_k^i(\mathbf{z}) &= \sum_{i\in\mathcal{V}} \Big[D_\psi(\mathbf{z}, \mathbf{v}_i(k)) - D_\psi(\mathbf{z}, \mathbf{x}_i(k+1))\Big] \\ &\quad - \sum_{i\in\mathcal{V}} D_\psi(\mathbf{x}_i(k+1), \mathbf{v}_i(k)).\end{aligned}$$

Utilizing the convexity of $D_\psi$ in the second argument gives

$$\sum_{i\in\mathcal{V}} D_\psi(\mathbf{z}, \mathbf{v}_i(k)) = \sum_{i\in\mathcal{V}} D_\psi\left(\mathbf{z}, \sum_{j\in\mathcal{N}_i} a_{ij}\mathbf{x}_i(k)\right) \le \sum_{i\in\mathcal{V}} D_\psi\left(\mathbf{z}, \mathbf{x}_i(k)\right),$$

which when combining with Eq. (5.2b) gives

$$\begin{aligned}\sum_{i\in\mathcal{V}} T_k^i(\mathbf{z}) &\le \sum_{i\in\mathcal{V}} \Big[D_\psi(\mathbf{z}, \mathbf{x}_i(k)) - D_\psi(\mathbf{z}, \mathbf{x}_i(k+1))\Big] \\ &\quad - \frac{\mu}{2}\sum_{i\in\mathcal{V}} \left\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\right\|^2. \end{aligned} \tag{5.16}$$

We now utilize the bounds on $\sum_{i\in\mathcal{V}} S_k^i(\mathbf{z})$ and $\sum_{i\in\mathcal{V}} T_k^i(\mathbf{z})$ to show the convergence of $\mathbf{x}_i(k)$, for all $i \in \mathcal{V}$. Let $\mathbf{x}^*$ be an optimizer of $f$ over $\mathcal{X}$. Applying the bounds in Eqs. (5.15) and (5.16) to Eq. (5.12) with $\mathbf{z} = \mathbf{x}^*$ gives

$$\begin{aligned}&\sum_{i\in\mathcal{V}} \Big[D_\psi(\mathbf{x}^*, \mathbf{x}_i(k+1)) - D_\psi(\mathbf{x}^*, \mathbf{x}_i(k))\Big] + \alpha(k)\Big(f(\bar{\mathbf{x}}(k)) - f(\mathbf{x}^*)\Big) \\ &\qquad \le \alpha(k)\sum_{i\in\mathcal{V}} C_i \left\|\bar{\mathbf{x}}(k) - \mathbf{x}_i(k)\right\| + \frac{C^2}{2\mu}\alpha^2(k)\cdot\end{aligned}$$

Summing the above over $k = 0, \ldots, K$, for some $K \geq 0$ we obtain

$$\sum_{i \in \mathcal{V}} \Big[ D_\psi(\mathbf{x}^*, \mathbf{x}_i(K+1)) - D_\psi(\mathbf{x}^*, \mathbf{x}_i(0)) \Big] + \sum_{k=0}^{K} \alpha(k) \Big( f(\bar{\mathbf{x}}(k)) - f(\mathbf{x}^*) \Big)$$
$$\leq \sum_{k=0}^{K} \sum_{i \in \mathcal{V}} \alpha(k) C_i \left\| \bar{\mathbf{x}}(k) - \mathbf{x}_i(k) \right\| + \frac{C^2}{2\mu} \sum_{k=0}^{K} \alpha^2(k). \tag{5.17}$$

We mimic the proof of Theorem 10 to complete the derivation. We first provide a bound on the right-hand side of Eq. (5.17). For convenience, consider

$$\mathbf{W} := \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\intercal \in \mathbb{R}^{n \times n} \qquad \text{and} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

By the Cauchy-Schwarz inequality, we then have

$$\sum_{i \in \mathcal{V}} C_i \left\| \bar{\mathbf{x}}(k) - \mathbf{x}_i(k) \right\| \leq C \left\| \mathbf{W}\mathbf{X}(k) \right\|_F, \tag{5.18}$$

where $\left\| \cdot \right\|_F$ denotes the Frobenius norm. Next, we consider

$$\begin{aligned} \left\| \mathbf{W}\mathbf{X}(k+1) \right\|_F &= \left\| \mathbf{W} \left( \mathbf{A}\mathbf{X}(k) + \mathbf{X}(k+1) - \mathbf{V}(k) \right) \right\|_F \\ &\leq \left\| \mathbf{A}\mathbf{W}\mathbf{X}(k) \right\|_F + \left\| \mathbf{W} \left( \mathbf{X}(k+1) - \mathbf{V}(k) \right) \right\|_F \\ &\leq \sigma_2 \left\| \mathbf{W}\mathbf{X}(k) \right\|_F + \left\| \mathbf{X}(k+1) - \mathbf{V}(k) \right\|_F, \end{aligned} \tag{5.19}$$

where the last inequality is due to Eq. (2.5). To bound each term in Eq. (5.19), we utilize Eqs. (5.10) and (5.11) with $\mathbf{z} = \mathbf{v}_i(k)$ to obtain

$$\Big\langle \alpha(k) \nabla f_i(\mathbf{v}_i(k)) \,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \Big\rangle$$
$$\geq \Big\langle \nabla \psi(\mathbf{v}_i(k)) - \nabla \psi(\mathbf{x}_i(k+1)) \,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \Big\rangle. \tag{5.20}$$

Since $f_i$ is $C_i$-Lipschitz we have and $\psi$ is $\mu$-strongly convex, we have

$$\Big\langle \alpha(k) \nabla f_i(\mathbf{v}_i(k)) \,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \Big\rangle \leq \alpha(k) C_i \left\| \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \right\|$$
$$\Big\langle \nabla \psi(\mathbf{v}_i(k)) - \nabla \psi(\mathbf{x}_i(k+1)) \,,\, \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \Big\rangle \geq \mu \left\| \mathbf{v}_i(k) - \mathbf{x}_i(k+1) \right\|^2,$$

that together with Eq. (5.20) gives

$$\|\mathbf{v}_i(k) - \mathbf{x}_i(k+1)\| \leq \frac{C_i \alpha(k)}{\mu}\cdot$$

Summing the above over $i \in \mathcal{V}$ gives a bound of Eq. (5.19), i.e.,

$$\|\mathbf{WX}(k+1)\|_F \leq \sigma_2 \|\mathbf{WX}(k)\|_F + \frac{C}{\mu}\alpha(k)\cdot \tag{5.21}$$

Iterating the above inequality gives

$$\|\mathbf{WX}(k)\|_F \leq \sigma_2^k \|\mathbf{WX}(0)\|_F + \frac{C}{\mu}\sum_{\ell=0}^{k-1} \alpha(\ell)\sigma_2^{k-\ell-1},$$

which further yields

$$\begin{aligned} &\sum_{k=0}^{K} \alpha(k) \|\mathbf{WX}(k)\|_F \\ &\qquad \leq \sum_{k=0}^{K} \alpha(k)\sigma_2^k \|\mathbf{X}(0)\|_F + \frac{C}{\mu}\sum_{k=0}^{K}\alpha(k)\sum_{\ell=0}^{k-1}\alpha(\ell)\sigma_2^{k-\ell-1}. \end{aligned} \tag{5.22}$$

Since $\alpha(k)$ is non-increasing with $\alpha(0) = 1$ and $\sigma_2 < 1$ we get

$$\sum_{k=0}^{K} \alpha(k)\, \sigma_2^k \leq \frac{1}{1-\sigma_2}\cdot$$

Using the preceding inequality in Eq. (5.22) we have

$$\begin{aligned} &\sum_{k=0}^{K}\sum_{i\in\mathcal{V}} \alpha(k) \|\bar{\mathbf{x}}(k) - \mathbf{x}_i(k)\| \\ &\qquad \leq \frac{\|\mathbf{X}(0)\|_F}{1-\sigma_2} + \frac{C}{\mu}\sum_{k=0}^{K}\sum_{\ell=0}^{k-1}\alpha^2(\ell)\sigma_2^{k-\ell-1} \\ &\qquad = \frac{\|\mathbf{X}(0)\|_F}{1-\sigma_2} + \frac{C}{\mu}\sum_{\ell=0}^{K-1}\alpha^2(\ell)\sum_{k=\ell+1}^{K}\sigma_2^{k-\ell-1} \\ &\qquad \leq \frac{\|\mathbf{X}(0)\|_F}{1-\sigma_2} + \frac{C\sum_{k=0}^{K-1}\alpha^2(k)}{\mu(1-\sigma_2)}\cdot \end{aligned} \tag{5.23}$$

Utilizing Eq. (5.23) into Eq. (5.17) yields

$$\sum_{i\in\mathcal{V}}\Big[D_\psi(\mathbf{x}^*,\mathbf{x}_i(K+1))-D_\psi(\mathbf{x}^*,\mathbf{x}_i(0))\Big]+\sum_{k=0}^{K}\alpha(k)[f(\bar{\mathbf{x}}(k))-f(\mathbf{x}^*)]$$
$$\leq\frac{\|\mathbf{X}(0)\|_F}{1-\sigma_2}+\frac{C}{\mu(1-\sigma_2)}\sum_{k=0}^{K}\alpha^2(k)+\frac{C^2}{2\mu}\sum_{k=0}^{K}\alpha^2(k).$$

Repeating the same argument on the preceding relation as in the proof of Theorem 10 we achieve

$$\lim_{k\to\infty}\mathbf{x}_i(k)=\mathbf{x}^*,\qquad\text{for all } i\in\mathcal{V}.$$

□

**Remark.** *Note that the assumption on the convexity of* $D_\psi(\mathbf{x},\mathbf{y})$ *over* $\mathbf{y}$ *for fixed* $\mathbf{x}$ *is crucial to our proof. A sufficient condition is derived in [80], that requires* $\psi$ *to satisfy* $\mathbf{H}_\psi(\mathbf{x})$ *and* $\mathbf{H}_\psi(\mathbf{x})+\nabla\mathbf{H}_\psi(\mathbf{x})(\mathbf{x}-\mathbf{y})$ *being positive semi-definite, for all* $\mathbf{x},\mathbf{y}\in\mathcal{X}$*, where* $\mathbf{H}_\psi$ *is the Hessian of* $\psi$*.*

## 5.4 Simulations

Theorems 10 and 11 guarantee the convergence of the iterates to the optimizer, but do not provide convergence rates with non-summable but square summable stepsizes. Given the lack of rates, we empirically illustrate that mirror descent – both in centralized and distributed settings – often outperforms vanilla subgradient methods on simple examples with our stepsizes.

Consider the following robust linear regression problem over a simplex.

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\text{minimize}}\quad\|\mathbf{G}\mathbf{x}-\mathbf{h}\|_1,\quad\text{subject to}\ \ \mathbf{1}^\intercal\mathbf{x}=1,\ \mathbf{x}\geq 0.\tag{5.24}$$

Robust regression fits a linear model to the data $\mathbf{G}\in\mathbb{R}^{N\times d},\mathbf{h}\in\mathbb{R}^N$, which are chosen uniformly at random from $[0,1]$ in this simulation. It differs from ordinary least squares in that the objective function penalizes the entry-wise absolute deviation from the linear fit rather than the squared residue, and is known to be robust to outliers [81]. Consider two different Bregman divergences on the $d$-dimensional simplex $\mathcal{X}$ defined by the Euclidean distance

$\psi_1(\mathbf{x}) := \frac{1}{2}\left\|\mathbf{x}\right\|_2^2$, and negative entropy $\psi_2(\mathbf{x}) := \sum_{j=1}^{d} x_j \log x_j$.

Centralized mirror descent with $D_{\psi_1}$ amounts to a projected subgradient algorithm where each iteration is a subgradient step followed by a projection on $\mathcal{X}$. With $D_{\psi_2}$, the updates define an exponentiated gradient method, also known as the entropic mirror descent algorithm (cf. [62, 82])

$$x_j(k+1) := \frac{x_j(k)\exp\big(-\alpha(k) f_j'(\mathbf{x}(k))\big)}{\sum_{\ell=1}^{d} x_\ell(k)\exp\big(-\alpha(k) f_\ell'(\mathbf{x}(k))\big)},$$

where the objective in problem (5.24) is $f(\mathbf{x})$, and $f_j'(\mathbf{x})$ is the $j$-th entries of the gradient of $f(\mathbf{x})$

$$\nabla f(\mathbf{x}) = \sum_{j=1}^{n} \text{sgn}\ ([\mathbf{g}_i]^T \mathbf{x} - h_i)\mathbf{g}_i.$$

Here, sgn $(\cdot)$ denotes the sign of the argument, and $[\mathbf{g}_i]^T$ is the $i$-th row of $\mathbf{G}$. For solving problem (5.24), negative entropy being a 'natural' function over simplex, entropic mirror descent enjoys faster convergence than projected subgradient descent, as shown in Fig. 5.1 using stepsizes $\alpha(k) = \frac{1}{k+1}$.

Next, consider the case where each node $i \in \mathcal{V}$ in a graph only knows $\mathbf{g}^i$ and $h^i$, and their goal is to cooperatively solve the problem in Eq. (5.24). The plots in Fig. 5.2 show that the distributed variant of entropic mirror descent outperforms the projected subgradient method with stepsizes $\alpha(k) = \frac{1}{k+1}$. We choose $\mathbf{A}$ as the Metropolis-Hastings matrix corresponding to graphs $\mathcal{G}$, where each $\mathcal{G}$ is generated by using the same steps described in Section 3.4. Centralized algorithms converge faster than distributed algorithms; however, the denser the graph, the faster the convergence is of the distributed algorithms.

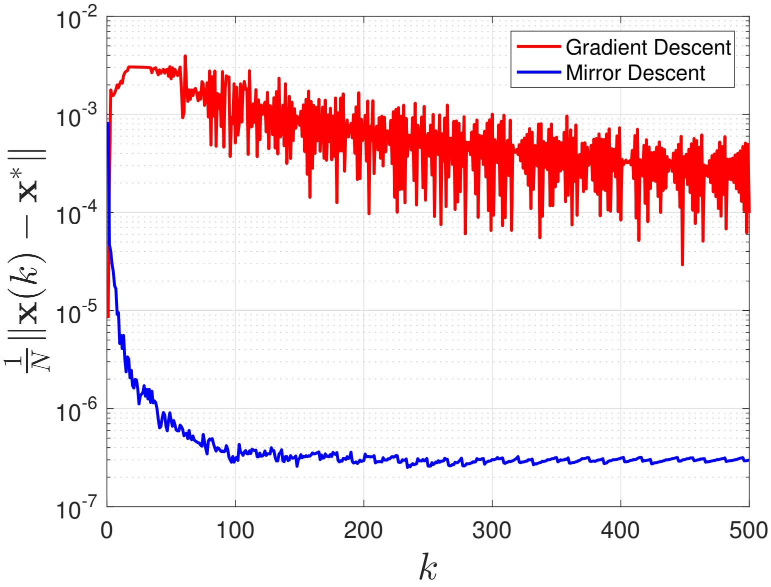

Figure 5.1: The convergence behavior of projected subgradient method (—) and entropic mirror descent (—) for solving problem (5.24).

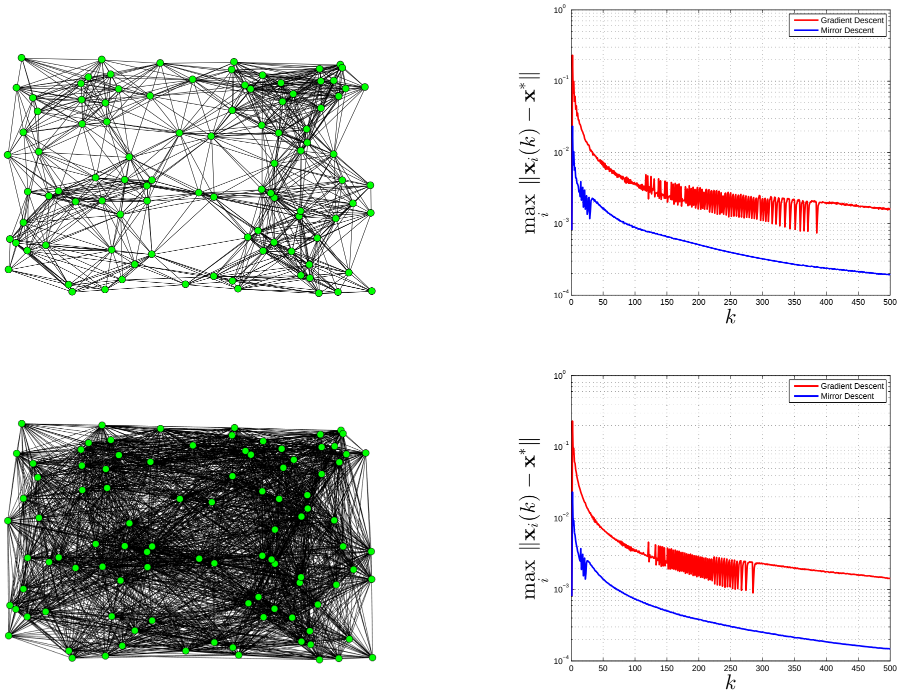

Figure 5.2: The convergence behavior of distributed projected subgradient method (—) and distributed entropic mirror descent (—) over a network with $n = 100$. Here, the top plot simulates for a graph with 939 edges, while the bottom plot simulates for a denser graph with 2678 edges.

# Chapter 6

# Distributed Random Projections

## 6.1 Problem Statement and Motivating Applications

In this chapter, we study problem (2.1) where the objective function and the constraint set are composed of local functions and local constraint sets, respectively. Due to the large number of these functions and constraint sets, we assume that they are distributed over a network of processors. Moreover, the communication structure between the processors is modeled by a star graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; see Fig. 1.1. For this structure, we are interested in the master-worker model, where there are $n$ worker nodes connected to a master. In particular, associated with each worker $i$ are a convex function $f_i : \mathbb{R}^d \to \mathbb{R}$ and a convex constraint set $\mathcal{X}_i \subset \mathbb{R}^d$. The goal of the master is to coordinate these nodes to solve the following problem

$$\text{minimize} \sum_{i=1}^{n} f_i(x) \tag{6.1a}$$

$$\text{such that } \mathbf{x} \in \mathcal{X} \triangleq \bigcap_{i=1}^{n} \mathcal{X}_i \subset \mathbb{R}^d. \tag{6.1b}$$

We further consider the case where the local constraint set $\mathcal{X}_i$ at the node $i$ is also expressed as an intersection of a large number of compact sets, i.e.,

$$\mathcal{X}_i = \bigcap_{j \in \mathcal{I}_i} \mathcal{X}_{ij}, \qquad \text{for all } i = 1, \ldots, n, \tag{6.2}$$

where each $\mathcal{X}_{ij}$ is assumed to have simple structure so that it is easy to implement projection. Since $|\mathcal{I}_i|$ is large, projecting onto $\mathcal{X}_i$ is costly, therefore, it is necessary to consider projections onto individual sets $\mathcal{X}_{ij}$.

A concrete motivating example for this problem is the distributed version of the well-known SVM problem [2] solved over a network of processors. SVM is one of the most popular methods for solving non-separable data classification problems in statistical learning theory. The goal of this problem is to find a hyper-plane with the largest margin to separate a collection of datasets. Mathematically, SVM can be formulated as the optimization problem

$$\begin{aligned} &\underset{\mathbf{w},\,b,\,\xi_i}{\text{minimize}} && \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} && \left.\begin{aligned} & v_i\Big(\mathbf{w}^T\mathbf{u}_i + b\Big) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}\right\} \forall i = 1,\ldots,n, \end{aligned}$$

where $\{(\mathbf{u}_i, v_i)\}$ are data points with very high dimension stored at processor $i$. The slack variables $\xi_i$ are used to handle the non-separable data. The goal is to find a hyper-plane represented by the normal vector $\mathbf{w}$ such that it has the largest margin (i.e., $1/\|\mathbf{w}\|$) on this dataset.

In solving problems (6.1), projected stochastic gradient descent (SGD) is a natural choice. However, such an approach would require one to project the result of each gradient descent step onto the feasible set, which can be computationally prohibitive. Therefore, there is a growing body of work which considers projections onto a random subset of the convex sets, see for example [83–86]. In [83–85], the focus is on centralized approaches, while [86] considers a fully distributed approach based on the so-called consensus algorithm. However, the consensus-based approach is not directly relevant for machine learning applications, since all processors that perform the computations are assumed to be peers, with no master node coordinating their actions. Here, we consider a distributed version where there is a master node (often called a parameter server) and a collection of other processors which are called workers. This model is a special case of the consensus model considered in [86]; however the simplicity of the parameter-server model allows one to derive convergence rates for the algorithm.

**Main Contributions**. The main contribution of this chapter is to derive the convergence rate of SGD with distributed random projections under the parameter-server model. In particular, we study the rate of convergence for two cases, namely, when the objective functions are convex and strongly convex, respectively. We provide an explicit formula for the convergence rate of the methods when the stepsizes are constant and time-varying. Under a mild regularity condition on the convex sets, we show that the rate of convergence of distributed SGD is unaffected by the presence of constraints, except for a factor which captures the regularity conditions. Since we perform random projections at each step, it is possible that the decision variables do not satisfy the constraints at each step of the SGD iteration. Our convergence rate results indicate that both the rate of convergence of the objective to its optimal value and the rate of convergence of the decision variables to the constraint set are the same.

In practice, the distributed algorithm when applied to real datasets can be implemented in multiple ways. In particular, each worker can select a subset of the data available to it and choose to implement each step of the algorithm on the chosen subset of data. Such a chosen subset is called a minibatch. In the simulations section, we study the impact of the number of workers and the mini-batch size on SGD with random projections.

**Relationship to Prior Work**. Our analysis is strongly motivated by similar analysis in [84,86]. The results in [84] are for the centralized case and the analysis in [86] is for a more general distributed consensus algorithm. However their assumptions on the (sub)gradient are different, and therefore, one cannot reach the main conclusion of our results from their results, i.e., that the convergence rates with and without projections in the parameter server model are essentially the same.

## 6.2 Distributed Random Projection Methods

To solve problem (6.1) we study distributed random projection methods under the parameter-server model, formally stated in Algorithm 3. Specifically, the master maintains a global variable $\bar{\mathbf{x}} \in \mathbb{R}^d$ used to estimate the minimizer $\mathbf{x}^*$ of problem (6.1), while each node maintains a variable $\mathbf{x}_i \in \mathbb{R}^d$. At each iteration $k \geq 0$, each node receives $\bar{x}(k)$ from the master and updates $\mathbf{x}_i(k+1)$ by a local stochastic subgradient step followed by a random projection step, as given in Eq. (6.3). In Eq. (6.3), $g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))$ is an unbiased estimate of node $i$'s subgradient where $\boldsymbol{\omega}_i$ is a random vector, i.e., $\mathbb{E}_{\omega_i}[g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))] \in \partial f_i(\bar{\mathbf{x}}(k))$, the subdifferential of $f_i$ at $\bar{\mathbf{x}}(k)$. In addition, at each iteration node $i$ only projects its value to one subset $\mathcal{X}_{i\zeta_i(k)}$ chosen randomly from its constraint set $\mathcal{I}_i$, i.e., $\zeta_i$ is a random variable taking values in $\mathcal{I}_i$. The nodes then send their values $\mathbf{x}_i(k+1)$ to the master to update $\bar{\mathbf{x}}(k+1)$ by taking the average of $\mathbf{x}_i(k+1)$, as given in Eq. (6.4).

**Algorithm 3** Distributed Random Projection Method

```
1. Intialize: $\bar{\mathbf{x}}(0) = \mathbf{x}_0 \in \mathbb{R}^d$ and $\epsilon > 0$
2. Repeat: For $k \geq 0$
   Each worker $i$ receives $\bar{\mathbf{x}}(k)$ and implements
       $\mathbf{x}_i(k+1) = \mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\Big[\bar{\mathbf{x}}(k) - \alpha(k) g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))\Big].$     (6.3)
   The master receives $\mathbf{x}_i(k+1)$ and updates $\bar{\mathbf{x}}$ as
       $\bar{\mathbf{x}}(k+1) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i(k+1).$     (6.4)
   Until: $\|\bar{\mathbf{x}}(k) - \mathbf{x}^*\| \leq \epsilon$.
```

## 6.3 Main Results

In this section we present the convergence analysis of Algorithm 3 when the functions $f_i$ are convex and strongly convex. Our focus is to establish the rate of its convergence. In particular, our main result basically states that under a mild regularity condition (due to Bauschke and Borwein [87, Definition 5.6], [88, (Definition 4.2.1]) on the constraint sets, the rate of convergence of distributed SGD with distributed random projections is the same as that of distributed SGD applied to a problem with no constraints, except for the appearance of a factor which captures the regularity assumption. When removing the constraint sets, our results reduce to those for distributed SGD.

We start our analysis by stating the following assumptions, which are necessary to our later analysis.

**Assumption 8.** *For all $\in \mathcal{V}$ and all $k \geq 0$ the sequences of random vectors $\{\boldsymbol{\omega}_i(k)\}$ and random variables $\{\zeta_i(k)\}$ are independent and identically distributed. In addition, $\mathbb{P}\{\zeta_i(k) = j\} > 0$, for all $j \in \mathcal{I}_i$ and all $i \in \mathcal{V}$.*

The condition $\mathbb{P}\{\zeta_i(k) = j\} > 0$, for all $j \in \mathcal{I}_i$, implies that each subset $\mathcal{X}_{ij}$ of $\mathcal{X}_i$ is chosen infinitely often.

**Assumption 9.** *For all $i \in \mathcal{V}$ there exists positive number $c_i$ such that $g_i(\mathbf{x}, \boldsymbol{\omega}_i)$ satisfies*

$$\mathbb{E}_{\boldsymbol{\omega}_i}\Big[\|g_i(\mathbf{x}, \boldsymbol{\omega}_i)\|^2\Big] \leq c_i^2 + \|\nabla f_i(x)\|^2. \tag{6.5}$$

We note that Assumption 9 is standard in the literature of stochastic gradient descent, i.e., it states that the variance of $g_i(\mathbf{x}, \boldsymbol{\omega}_i)$ is mildly restricted [89].

Let $\mathcal{F}(k)$ denote the filtration that contains all the information generated by Algorithm 3 up to time $k$, i.e., all the variables $\bar{\mathbf{x}}(t), \mathbf{x}_i(t), \boldsymbol{\omega}_i(t), \zeta_i(t)$ and so forth for $t = 0, \ldots, k$. We assume the following regularity condition of $\mathcal{X}$.

**Assumption 10.** *For all $i \in \mathcal{V}$ there exists a positive constant $D > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$*

$$dist^2(\mathbf{x}, \mathcal{X}) \leq D \max_{i \in \mathcal{V}} \mathbb{E}_{\zeta_i(k)}\Big[dist^2(\mathbf{x}, \mathcal{X}_{i\zeta_i(k)}) \mid \mathcal{F}(k)\Big]. \tag{6.6}$$

Assumption 10, can be referred as the linear regularity condition of convex sets and first introduced by Bauschke and Borwein, is essential in our analysis. It basically states that the distance of a point $\mathbf{x}$ to the feasible set $\mathcal{X}$ can be upper-bouneded by its distance to individual sets $\mathcal{X}_i$. Assumption 10 is quite general, i.e., it holds when the set $\mathcal{X}$ has a nonempty interior [90] or the sets $\mathcal{X}_i$ are half spaces [91], for example, the latter holds in the example of SVMs. Finally, this assumption has been recently considered in [83–86].

Since the sets $\mathcal{X}_{ij}$ are compact, $\mathcal{X}_i$ is compact. Hence, by Assumption 9, there exists a positive constant $C_i$ such that

$$\mathbb{E}\Big[\|g_i(\mathbf{x}, \boldsymbol{\omega}_i(k))\|^2\Big] \leq C_i^2, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{6.7}$$

For convenience, the local update in Eq. (6.3) can be rewritten as

$$\mathbf{v}_i(k) = \bar{\mathbf{x}}(k) - \alpha(k) g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k)) \tag{6.8}$$

$$\mathbf{x}_i(k+1) = \mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\Big[\mathbf{v}_i(k)\Big]. \tag{6.9}$$

Finally, we denote by $\mathbf{x}^*$ a solution of problem (6.1), and consider a bit more notation

$$\begin{aligned} &\bar{\mathbf{r}}(k) = \bar{\mathbf{x}}(k) - \mathbf{x}^*, \quad \mathbf{e}(k) = \bar{\mathbf{x}}(k) - \mathcal{P}_{\mathcal{X}}[\bar{\mathbf{x}}(k)] \\ &f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x}), \quad f^* = \sum_{i=1}^n f_i(\mathbf{x}^*), \quad C = \sum_{i=1}^n C_i, \quad \mu = \sum_{i=1}^n \mu_i. \end{aligned} \tag{6.10}$$

Here $\|\mathbf{e}(k)\|$ denotes the feasibility violation of $\bar{\mathbf{x}}(k)$ at time $k$.

We now present the main result of this section, which is the convergence rate of Algorithm 3. In particular, we provide the rate of convergence when $f_i$ are convex and strongly convex as well as when the stepsizes $\alpha$ are constant and time-varying. To do this, we first state two preliminary results in the following lemmas. In the first lemma, we use the property of the projection step in Eq. (6.9) to upper bound the distance $\|\bar{\mathbf{x}}(k+1)-\mathbf{x}^*\|^2$ by the average of $\|\mathbf{v}(k)-\mathbf{x}^*\|^2$ and the feasibility violation $\mathbf{e}(k)$ of $\bar{\mathbf{x}}(k)$. The second lemma provides an upper bound on the distance $\|\mathbf{v}(k)-\mathbf{x}^*\|^2$ via Eq. (6.8). Their proofs are presented in Appendix D.

**Lemma 9.** *Suppose that Assumptions 8–10 hold. Let the sequences* $\{\mathbf{x}_i(k)\}$, $\{\mathbf{v}(k)\}$, *and* $\{\bar{\mathbf{x}}(k)\}$ *satisfy Eqs.* (6.9), (6.8), *and* (6.4), *for all* $i\in\mathcal{V}$. *Denote* $\beta_1 = 8C^2/Dn^2$. *Then,*

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{v}(k)-\mathbf{x}^*\|^2\Big] + \beta_1\alpha^2(k) - \frac{\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{2Dn}\cdot \tag{6.11}$$

**Lemma 10.** *Suppose that Assumptions 8–10 hold. Let the sequences* $\{\mathbf{x}_i(k)\}$, $\{\mathbf{v}(k)\}$, *and* $\{\bar{\mathbf{x}}(k)\}$ *satisfy Eqs.* (6.9), (6.8), *and* (6.4), *for all* $i\in\mathcal{V}$. *Let* $\mathbf{y} = \mathcal{P}_{\mathcal{X}}\big[\bar{\mathbf{x}}(k)\big]\in\mathcal{X}$ *and denote* $\beta_2 = (8Dn+1)C^2$. *Then,*

1. *If Assumption 5 holds then*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{v}(k)-\mathbf{x}^*\|^2\Big] &\leq \Big(1-\frac{\mu\alpha}{n}\Big)\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta_2\alpha^2(k)\\ &\qquad - \frac{2\alpha(k)\mathbb{E}\Big[f(\mathbf{y})-f^*\Big]}{n} + \frac{\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{4Dn}\cdot\end{aligned} \tag{6.12}$$

2. *If the functions* $f_i$ *are convex for all* $i\in\mathcal{V}$ *then*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\|\mathbf{v}(k)-\mathbf{x}^*\|^2\Big] &\leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta_2\alpha^2(k)\\ &\qquad - \frac{2\alpha(k)}{n}\mathbb{E}\Big[f(\mathbf{y})-f^*\Big] + \frac{\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{4Dn}\cdot\end{aligned} \tag{6.13}$$

We are now ready to study the convergence rate of Algorithm 3. As noted previously, our results basically states that under Assumption 10 the rate of convergence of Algorithm 3 is the same as that of distributed SGD applied to a problem with no constraints, except for the scaling $D$. The following two theorems present these results, where their analyses are based on coupling the results in Lemmas 9 and 10. The key idea is to show that the random projection at each node does not cause the variables to oscillate between the local sets, but rather pushes them toward the global feasible set $\mathcal{X}$. We first derive the convergence rate of Algorithm 3 when the functions $f_i$ are strongly convex under two conditions, namely, constant and time-varying stepsizes.

**Theorem 12.** *Suppose that Assumptions 5 and 8–10 hold. Let the sequences* $\{\mathbf{x}_i(k)\}$, $\{\mathbf{v}(k)\}$, *and* $\{\bar{\mathbf{x}}(k)\}$ *satisfy Eqs.* (6.9), (6.8), *and* (6.4), *for all* $i \in \mathcal{V}$. *Let* $\beta = \beta_1 + \beta_2$ *where* $\beta_1 = 8C^2/Dn^2$ *and* $\beta_2 = (8Dn+1)C^2$. *Then*

1. *Consider* $\alpha(k) = \alpha$ *for all* $k \geq 0$ *where* $\alpha \in (0, n/\mu)$. *In addition, let* $\gamma = (1 - \mu\alpha/n) \in (0,1)$. *Then we have*

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] \leq \gamma^k \mathbb{E}[\|\bar{\mathbf{r}}(0)\|^2] + \frac{n\beta\alpha}{\mu}\cdot \tag{6.14}$$

2. *Let* $\alpha(k) = n/\mu(k+1)$, *then we have for* $k \geq 1$

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq \frac{4\mu^2\mathbb{E}\Big[\|\bar{\mathbf{r}}(1)\|^2\Big] + n^2\beta\ln(k+1)}{4\mu^2(k+1)}\cdot \tag{6.15}$$

*Proof.* We proceed the proof of Theorem 12 in two steps as follows:

1. Since each $f_i$ is strongly convex, $\mu > 0$. Then we have

$$\gamma = 1 - \frac{\mu\alpha}{n} \in (0,1).$$

By Eq. (6.11) with $\alpha(k) = \alpha$

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\|\mathbf{v}_i(k) - \mathbf{x}^*\|^2\Big] + \beta_1\alpha^2 - \frac{1}{2Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big],$$

which by Eq. (6.12) with $\mathbf{y} = \mathcal{P}_{\mathcal{X}}[\bar{\mathbf{x}}(k)]$ and $\beta = \beta_1 + \beta_2$ implies

$$\begin{aligned}\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] &\leq \gamma\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta_2\alpha^2 - \frac{2\alpha}{n}\mathbb{E}\Big[f(\mathcal{P}_{\mathcal{X}}[\bar{\mathbf{x}}(k)]) - f^*\Big]\\ &\quad + \frac{1}{4n}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] + \beta_1\alpha^2 - \frac{1}{2n}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]\\ &\leq \gamma\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta\alpha^2 - \frac{1}{4n}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big],\end{aligned} \tag{6.16}$$

where we use the fact that $f\left(\mathcal{P}_{\mathcal{X}}[\bar{\mathbf{x}}(k)]\right) - f^* \geq 0$. Recursively updating Eq. (6.16) we obtain Eq. (6.14), i.e.,

$$\frac{1}{4n}\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] \leq \gamma\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta\alpha^2 \leq \gamma^{k+1}\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \beta\alpha^2\sum_{t=0}^{k}\gamma^{k-t}$$
$$\leq \gamma^{k+1}\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \frac{\beta\alpha^2}{1-\gamma} \leq \left(1-\frac{\mu\alpha}{n}\right)^k\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \frac{n\beta\alpha}{\mu}.$$

2. We now consider $\alpha(k) = n/\mu(k+1)$ for all $k \geq 0$. Then by Eq. (6.16) with $\gamma = 1 - \mu\alpha(k)/n = 1 - 1/(k+1)$ we have

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq \frac{k}{k+1}\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \frac{n^2\beta}{4\mu^2(k+1)^2}.$$

Multiplying both sides of the preceding relation by $k+1$

$$(k+1)\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq k\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \frac{n^2\beta}{4\mu^2(k+1)},$$

which when summing both sides over $k$ from 1 to $K$ for some $K \geq 1$ gives

$$(K+1)\mathbb{E}\Big[\|\bar{\mathbf{r}}(K+1)\|^2\Big] \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(1)\|^2\Big] + \frac{n^2\beta\ln(K+1)}{4\mu^2}, \tag{6.17}$$

where we use the following condition

$$\sum_{k=1}^{K}\frac{1}{k+1} \leq \frac{1}{2} + \int_1^K \frac{dt}{t+1} \leq \ln(K+1).$$

Dividing both sides of Eq. (6.17) by $K+1$ we have

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(K+1)\|^2\Big] \leq \frac{4\mu^2\mathbb{E}\Big[\|\bar{\mathbf{r}}(1)\|^2\Big] + n^2\beta\ln(K+1)}{4\mu^2(K+1)}.$$

□

We now derive the rate of Algorithm 3 when the functions $f_i$ are convex. Given any sequence $\{\mathbf{x}(k)\}$ we denote by $\hat{\mathbf{x}}(k)$ its time $\alpha$-weighted average

$$\hat{\mathbf{x}}(k) = \frac{\sum_{t=0}^{k-1}\alpha(t)\mathbf{x}(t)}{\sum_{t=0}^{k-1}\alpha(t)}. \tag{6.18}$$

**Theorem 13.** *Suppose that Assumptions 8–10 hold. Let the sequences* $\{\mathbf{x}_i(k)\}$, $\{\mathbf{v}(k)\}$, *and* $\{\bar{\mathbf{x}}(k)\}$ *satisfy Eqs.* (6.9), (6.8), *and* (6.4), *for all* $i \in \mathcal{V}$. *Let* $\beta = \beta_1 + \beta_2$ *where* $\beta_1 = 8C^2/Dn^2$ *and* $\beta_2 = (8Dn+1)C^2$.

1. *Let* $\alpha$ *be some positive constant. Then for all* $k \geq 0$

$$\mathbb{E}\Big[\|\hat{\mathbf{e}}(k)\|^2\Big] \leq \frac{4Dn\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big]}{k} + 4Dn\beta\alpha^2 \tag{6.19a}$$

$$\mathbb{E}\Big[\big|f(\hat{\bar{\mathbf{x}}}(k)) - f^*\big|\Big] \leq \frac{n\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big]}{2\alpha k} + \frac{(n\beta + 4DC^2)\alpha}{2}\cdot \tag{6.19b}$$

2. *Let* $\alpha(k) = 1/\sqrt{k+1}$. *Then for all* $k \geq 0$

$$\mathbb{E}\Big[\|\hat{\mathbf{e}}(k)\|^2\Big] \leq \frac{4Dn\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + 4Dn\beta(\ln(k)+1)}{\sqrt{k+1}} \tag{6.20a}$$

$$\mathbb{E}\Big[\big|f(\hat{\bar{\mathbf{x}}}(k)) - f^*\big|\Big] \leq \frac{n\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big]}{2\sqrt{k+1}} + \frac{(n\beta + 4DC^2)(\ln(k)+1)}{2\sqrt{k+1}}\cdot \tag{6.20b}$$

*Proof.* We proceed the proof of this theorem as follows:

1. Denote by $\mathbf{y}(k) = \mathcal{P}_{\mathcal{X}}\big[\bar{\mathbf{x}}(k)\big]$ and let $\alpha(k) = \alpha$. Using Eqs. (6.11) and (6.13) gives

$$\mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big] \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] + \beta\alpha^2 - \frac{1}{4Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] - \frac{2\alpha}{n}\mathbb{E}\Big[f(\mathbf{y}(k)) - f^*\Big],$$

which when reorganizing both sides implies

$$\frac{1}{4Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] + \frac{2\alpha}{n}\mathbb{E}\Big[f(\mathbf{y}(k)) - f^*\Big] \leq \Big(\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big]\Big) + \beta\alpha^2. \tag{6.21}$$

Summing Eq. (6.21) over $k$ from 0 to $K-1$ for some $K \geq 1$ and since $f^* \leq f(\mathbf{y}(k))$ gives

$$\frac{1}{4Dn}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \beta\alpha^2 K.$$

Dividing both sides of above by $K$ and using the Jensen inequality give

$$\mathbb{E}\Big[\|\hat{\mathbf{e}}(K)\|^2\Big] \leq \frac{4Dn\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big]}{K} + 4Dn\beta\alpha^2,$$

which is Eq. (6.19a). In addition, by the triangle inequality consider

$$\begin{aligned}\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big] &\leq \mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f(\mathbf{y}(k))\big|\Big] + \mathbb{E}\Big[\big|f(\mathbf{y}(k)) - f^*\big|\Big]\\ &\overset{(6.7)}{\leq} C\mathbb{E}\Big[\|\mathbf{e}(k)\|\Big] + \mathbb{E}\Big[\big|f(\mathbf{y}(k)) - f^*\big|\Big]. \qquad (6.22)\end{aligned}$$

The preceding relation implies that

$$\begin{aligned}\frac{2\alpha}{n}\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big] &\leq \frac{2C\alpha}{n}\mathbb{E}[\|\mathbf{e}(k)\|] + \frac{2\alpha}{n}\mathbb{E}[|f(\mathbf{y}(k)) - f^*|\Big]\\ &\leq \frac{1}{4Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] + \frac{4DC^2\alpha^2}{n} + \frac{2\alpha}{n}\mathbb{E}\Big[\big|f(\mathbf{y}(k)) - f^*\big|\Big]\\ &\overset{(6.21)}{\leq} \left(\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big]\right) + \beta\alpha^2 + \frac{4DC^2\alpha^2}{n}.\end{aligned}$$

Summing both sides of above over $k$ from 0 to $K-1$ for $K \geq 1$ gives

$$\frac{2\alpha}{n}\sum_{k=0}^{K}\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big] \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \left(\beta\alpha^2 + \frac{4DC^2\alpha^2}{n}\right)K,$$

which by dividing by $K$ and using the Jensen inequality gives Eq. (6.19b).

2. Let $\alpha(k) = 1/\sqrt{k+1}$ for $k \geq 0$. By Eq. (6.21) we have

$$\begin{aligned}&\frac{1}{4Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] + \frac{2\alpha(k)}{n}\mathbb{E}\Big[f(\mathbf{y}(k)) - f^*\Big]\\ &\qquad \leq \left(\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big]\right) + \beta\alpha^2(k),\end{aligned}$$

which since $\alpha(k) \leq 1$ implies that

$$\begin{aligned}&\frac{\alpha(k)\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{4Dn} + \frac{2\alpha(k)\mathbb{E}\Big[f(\mathbf{y}(k)) - f^*\Big]}{n}\\ &\qquad \leq \left(\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big]\right) + \beta\alpha^2(k). \qquad (6.23)\end{aligned}$$

Summing Eq. (6.23) over $k$ from 0 to $K-1$ for $K \geq 1$ and since $f^* \leq f(\mathbf{y}(k))$ gives

$$\sum_{k=0}^{K-1} \frac{\alpha(k)\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{4Dn} \leq \left(\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(K)\|^2\Big]\right) + \sum_{k=0}^{K-1} \frac{\beta}{k+1}$$
$$\leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \beta(\ln(K)+1), \qquad (6.24)$$

where the last inequality is due to

$$\sum_{k=0}^{K-1} \frac{1}{k+1} \leq ln(K) + 1,$$

as given earlier. We further have the following inequality since $K \geq 1$

$$\sum_{k=0}^{K-1} \frac{1}{\sqrt{k+1}} \geq 1 + \int_1^K \frac{du}{\sqrt{u+1}} \geq \sqrt{K+1}. \qquad (6.25)$$

Dividing both sides of Eq. (6.24) by $\sum\limits_{k=0}^{K-1} \alpha(k)$ and using Eq. (6.25) we have

$$\frac{\sum\limits_{k=0}^{K-1} \alpha(k)\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big]}{\sum\limits_{k=0}^{K-1} \alpha(k)} \leq \frac{4Dn\mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + 4Dn\beta(\ln(K)+1)}{\sqrt{K+1}}, \qquad (6.26)$$

which is Eq. (6.20a) due to the Jensen inequality. By Eq. (6.22) we obtain

$$\frac{2\alpha(k)}{n}\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big]$$
$$\leq \frac{2C\alpha(k)}{n}\mathbb{E}\Big[\|\mathbf{e}(k)\|\Big] + \frac{2\alpha(k)}{n}\mathbb{E}\Big[\big|f(\mathbf{y}(k)) - f^*\big|\Big]$$
$$\leq \frac{1}{4Dn}\mathbb{E}\Big[\|\mathbf{e}(k)\|^2\Big] + \frac{4DC^2\alpha^2(k)}{n} + \frac{2\alpha(k)}{n}\mathbb{E}\Big[\big|f(\mathbf{y}(k)) - f^*\big|\Big].$$

Using Eq. (6.23) into above gives

$$\frac{2\alpha(k)}{n}\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big]$$
$$\leq \left(\mathbb{E}\Big[\|\bar{\mathbf{r}}(k)\|^2\Big] - \mathbb{E}\Big[\|\bar{\mathbf{r}}(k+1)\|^2\Big]\right) + \frac{n\beta + 4DC^2}{n}\alpha^2(k),$$

which by summing both sides over $k$ from 0 to $K-1$ for $K \geq 1$ implies

$$\begin{aligned}
&\frac{2}{n}\sum_{k=0}^{K-1}\alpha(k)\mathbb{E}\Big[\big|f(\bar{\mathbf{x}}(k)) - f^*\big|\Big] \\
&\qquad \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \frac{n\beta + 4DC^2}{n}\sum_{k=0}^{K-1}\alpha^2(k) \\
&\qquad \leq \mathbb{E}\Big[\|\bar{\mathbf{r}}(0)\|^2\Big] + \frac{n\beta + 4DC^2}{n}\left(\ln(K) + 1\right).
\end{aligned}$$

Dividing both sides of the preceding relation by $\sum_{k=0}^{K-1}\alpha(k)$ and by the Jensen inequality we obtain Eq. (6.20b).

□

## 6.4 Simulations

In this section, we perform experiments to demonstrate the convergence of our algorithm in distributed SVM. We suppose that SVM is applied to a dataset consisting of $N$ points and we assume that the dataset is divided into $n$ equal parts and stored at $n$ different workers. Let $\mathcal{D}_i$ denote the subset of data points stored at worker $i$. We take the formulation of SVM described in Section 6.1, and define $f_i$ and $\mathcal{X}_i$ for each data point so that the original objective can be decomposed according to Eqs. (6.1a) and (6.1b)

$$\begin{aligned}
f_i(\mathbf{w}, b, \mathcal{X}_i) &= \frac{1}{2n}\|\mathbf{w}\|^2 + C\sum_{j\in\mathcal{D}_i}\mathcal{X}_{ij} \\
\mathcal{X}_i &= \{(\mathbf{w}, b, \mathcal{X}_i) : v_j(\langle\mathbf{w}, \mathbf{u}_j\rangle + b) \geq 1 - \mathcal{X}_{ij}, \mathcal{X}_{ij} \geq 0, \forall j \in \mathcal{D}_i\}.
\end{aligned}$$

Here $(\mathbf{u}_j, v_j)$ denotes a single data point. For all $k \geq 0$, each worker $i$ runs in parallel, has access to $\mathcal{D}_i$, receives an estimate $\bar{\mathbf{w}}(k)$, and produces an update of the form

$$\mathbf{w}_i(k+1) = \hat{\mathcal{P}}_{\mathcal{X}_i}\left[\bar{\mathbf{w}}(k) - \alpha(k)\nabla f_i(\bar{\mathbf{w}}(k))\right],$$

where $\hat{\mathcal{P}}_{\mathcal{X}_i}$ is an estimate of the projection onto the set $\mathcal{X}_i$ at time $k$.

In this simulation, $\hat{\mathcal{P}}_{\mathcal{X}_i}(\mathbf{z})$ is computed as follows:

$$\hat{\mathcal{P}}_{\mathcal{X}_i}(\mathbf{z}) = \frac{1}{B} \sum_{\ell \in \mathcal{I}_i(k)} \mathcal{P}_{\mathcal{X}_{i\ell}}(\mathbf{z}), \qquad \text{for all } \mathbf{z},$$

where $\mathcal{I}_i(k)$ is a randomly selected subset of $\mathcal{D}_i$ with cardinality $B$. The master receives each $\mathbf{w}_i(k+1)$ and updates $\bar{\mathbf{w}}$ as:

$$\bar{\mathbf{w}}(k+1) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i(k+1).$$

For our tests, we repeat this process for 100,000 iterations and examine the relative error $\left| f(\bar{\mathbf{w}}(k)) - f(\mathbf{w}^*) \right| / \left| f(\mathbf{w}^*) \right|$, where $\mathbf{w}^*$ is found by running the LIBLINEAR SVM solver until convergence.

Our simulations use the "a1a" dataset from the LIBSVM dataset collection [92]. This dataset attempts to predict whether income exceeds 50k/yr, based on census data. It contains 1605 labeled data points and 123 features. For testing, we split the data into $n$ roughly equal portions and run the Distributed Stochastic Subgradient Descent with Random Projection for $100,000$ iterations, with a batch size of $B$ per worker. The stepsizes $\alpha(k)$ are set to $1/\sqrt{k+1}$ and $C$ is set to 1.

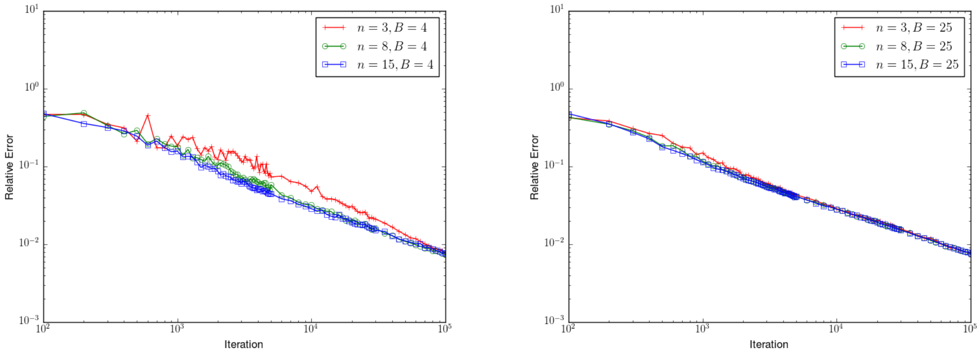

Figure 6.1: Relative error, with two different values of $B$ and varying numbers of workers

First, we fix the minibatch size $B$ and vary the number of workers $n$, and plot the relative error over the run of the algorithm in Figure 6.1. Second, we fix the number of workers and vary the minibatch size, as seen in Figure 6.2. The relative error decreases faster as $B$ increases. In both figures, as the theory suggests, the relative errors asymptotically become small for large $k$, regardless of $n$ or $B$.

Our findings for Distributed Stochastic Subgradient Descent with Random Projection appear to reflect classic results for Minibatch SGD. Although both algorithms are proven to converge even for minibatch sizes of 1, there is a significant practical benefit from increasing the minibatch size. This may be due to the increased number of total data points per iteration contributing to a similar averaging effect. We have observed that increasing the number of workers has a similar effect.

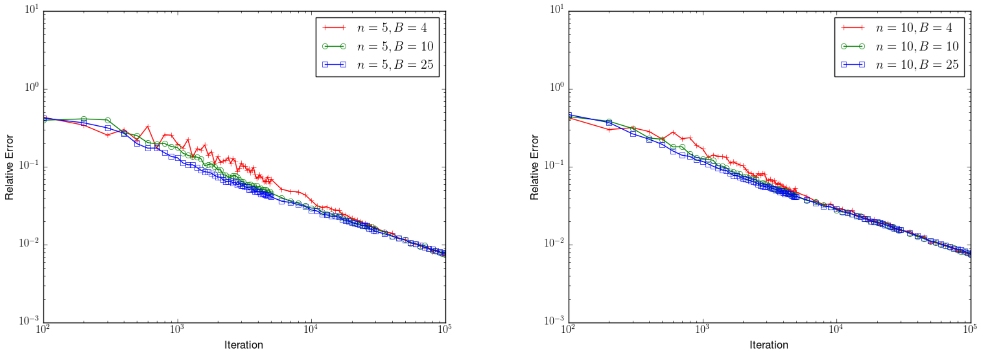

Figure 6.2: Relative error, with two different values of $n$ and varying minibatch sizes

# Chapter 7

# Distributed Lagrangian Methods for Network Resource Allocation

## 7.1 Problem Statement, Motivation, and Contribution

In this chapter, we consider an optimization problem, defined over time-varying graphs $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k))$, of the form

$$
\mathsf{P}\ :\ \begin{cases} \underset{x_1,x_2,\ldots,x_n}{\text{minimize}} & \displaystyle\sum_{i=1}^{n} f_i(x_i) \\ \text{subject to} & x_i \in \mathcal{X}_i, \qquad \forall i \in \mathcal{V} \qquad\qquad (7.1\text{a}) \\ & \displaystyle\sum_{i=1}^{n} x_i = b, \qquad\qquad (7.1\text{b}) \end{cases}
$$

where each node $i \in \mathcal{V}$ stores a variable $x_i \in \mathbb{R}$. We assume that each $f_i : \mathbb{R} \to \mathbb{R}$ is a convex function, and $\mathcal{X}_i \subset \mathbb{R}$ is a compact set, where both are known by node $i$.

Problem $\mathsf{P}$ is often referred to as a network resource allocation problem, where the goal is to optimally allocate a fixed quantity of resource $b$ over a network of nodes. Each node $i$ suffers a cost given by function $f_i$ of the amount of resource $x_i$ allocated to it. The goal of this problem is to seek an optimal allocation such that the total cost $\sum_{i\in\mathcal{V}} f_i(x_i)$ incurred over the network is minimized while satisfying the nodes' local constraints, i.e., $x_i \in \mathcal{X}_i$. Often problem $\mathsf{P}$ is described in terms of utility functions, where each function is the nodes' utility and the goal is to maximize the total utility.

Network resource allocation is a fundamental and important problem that arises in a variety of application domains within engineering. One standard example is the problem of congestion control where the global objective is to route and schedule information in a large-scale Internet network such that a fair resource allocation between users is achieved [4]. Another example is coverage control problems in wireless sensor networks, where the goal is to optimally allocate a large number of sensors to an unknown environment such that the coverage area is maximized [12,13]. Furthermore, resource allocation may be viewed as a simplification of the important economic dispatch problem in power systems, wherein geographically distributed generators of electricity must coordinate to meet a fixed demand while maintaining the stability of the system [93,94].

Due to its broad applications, especially, in power engineering, there has been a great interest in studying distributed methods for problem $\mathsf{P}$. In particular, the authors in [95–99] design distributed algorithms for solving economic dispatch problems, an application of $\mathsf{P}$, where objective functions are assumed to be quadratic. The authors in [100,101] relax the assumption on quadratic costs to convex cost functions with Lipschitz continuous gradients, and consider relaxed problems by using appropriate penalty functions for the nodes' local constraints. In a similar approach, the authors in [102] consider problem (7.1) with general non-smooth convex cost functions and propose a method with a convergence rate $o(1/k)$ where $k$ is the number of iterations.

In this chapter, we provide a distributed algorithm, namely a distributed Lagrangian method, for problem $\mathsf{P}$. The hallmark of this approach is the eliminatation of the need for a central coordinator to update the dual variables, where these authors employ the distributed gradient method presented in Section 2.2.2 to solve the dual problem of $\mathsf{P}$. The results in this chapter are based on our recent work in [77,103]. Distributed Lagrangian methods have been also considered in [104], which requires an assumption of strict convexity of the objective functions. Our focus in this chapter is to consider problem $\mathsf{P}$ when the objective functions are convex and the number of resources is uncertain.

**Main Contribution**. Previous approaches that have been proposed to solve problem $\mathsf{P}$ assume the number of resources $b$ is constant. This critical assumption is impractical in most applications. For example, in power systems load demands are typically time-varying and the data defining such load demands may be uncertain [105]. For this reason, any solution to network resource allocation problems should be robust to uncertainty. This issue has not been addressed in the literature. Therefore, our main contribution in this chapter is to address this question. Specifically, our primary contributions are summarized as follows:

1. We first study a distributed Lagrangian method for problem $\mathsf{P}$, where the total quantity of resource $b$ is assumed to be constant over time-varying networks. The development and analysis in this case allows for an extension to the case where $b$ is uncertain.

2. We then propose a distributed randomized Lagrangian approach for problem $\mathsf{P}$ for the case where $b$ is unknown and may be time-varying. We show that our approach is robust to this uncertainty, that is, our randomized method achieves an asymptotic convergence in expectation to the optimal value. Moreover, we show that our method converges with rate $\mathcal{O}(n\ln(k)/\delta\sqrt{k})$, where $\delta$ is a parameter representing spectral properties of the graph structure underlying the connectivity of the nodes, $n$ is the number of nodes, and $k$ is the number of iterations.

3. To illustrate the effectiveness of the proposed methods, we present numerical results from applications to economic dispatch problems using the benchmark IEEE-14 and IEEE-118 bus test systems for three case studies.

For ease of exposition, we put all the proofs of the main results in this chapter to Appendix E.

## 7.2 Distributed Lagrangian Methods

Lagrangian methods have been widely used to construct a decentralized framework for problem $\mathsf{P}$, where each node in the network only has partial knowledge of the objective function and the constraints; this approach requires a central coordinator to update and distribute the Lagrange multiplier to the nodes. In this section, we present an alternative approach that allows us to bypass the need for a central coordinator, that is, our proposed approach allows for a truly distributed implementation, thus leading to a more efficient algorithm. The development in this section also informs our approach on distributed randomized Lagrangian methods for network resource allocation problems, which is presented in Section 7.3.

**Algorithm 4** Distributed Lagrangian Method for Solving $\mathsf{P}$.

```
1. Initialize: Each node $i \in \mathcal{V}$ initializes $\lambda_i(0) \in \mathbb{R}$.
2. Iteration: Each node $i \in \mathcal{V}$, executes
       $v_i(k+1) = \sum_{j \in \mathcal{N}_i(k)} a_{ij}(k) \lambda_j(k)$   (7.2)
       $x_i(k+1) \in \arg\min_{x_i \in \mathcal{X}_i} f_i(x_i) + v_i(k+1)(x_i - b_i)$   (7.3)
       $\lambda_i(k+1) = v_i(k+1) + \alpha(k)\left(x_i(k+1) - b_i\right).$   (7.4)
```

### 7.2.1 Main Algorithm

Without loss of generality, we assume that each node $i$ knows the constant $b_i$ such that $\sum_{i=1}^{n} b_i = b$. Here $b_i$ can be interpreted as the initial resource allocation at node $i$. One specific choice is to initially distribute $b$ equally to all nodes, i.e., $b_i = b/n$ for all $i \in \mathcal{V}$. We note that the design of our algorithm as well as our analysis given later does not depend on the choice of these $b_i$ since they are only used for notational convenience.

We now explain the mechanics of our approach. Consider the Lagrangian function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ of $\mathsf{P}$ given as

$$\mathcal{L}(\mathbf{x}, \lambda) \triangleq \sum_{i \in \mathcal{V}} f_i(x_i) + \lambda \left( \sum_{i \in \mathcal{V}} (x_i - b_i) \right), \tag{7.5}$$

where $\lambda \in \mathbb{R}$ is the Lagrangian multiplier associated with Eq. (7.1b). The dual function $d : \mathbb{R} \to \mathbb{R}$ of problem $\mathsf{P}$ for some $\lambda$ value is then defined as

$$d(\lambda) := \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \sum_{i \in \mathcal{V}} f_i(x_i) + \lambda \left( \sum_{i \in \mathcal{V}} (x_i - b_i) \right) \right\} = \sum_{i \in \mathcal{V}} \{ -f_i^*(-\lambda) - \lambda b_i \},$$

where $f_i^*$ is the Fenchel conjugate of $f_i$, defined by

$$f_i^*(u) = \sup_{x \in \mathbb{R}} \{ ux - f_i(x) \}.$$

The dual problem of $\mathsf{P}$, denoted by $\mathsf{DP}$, is given by

$$\mathsf{DP} \ : \ \max_{\lambda \in \mathbb{R}} \sum_{i \in \mathcal{V}} \{ -f_i^*(-\lambda) - \lambda b_i \},$$

which is then equivalent to solving

$$\min_{\lambda\in\mathbb{R}} \sum_{i\in\mathcal{V}} \underbrace{f_i^*(-\lambda) \; + \; \lambda b_i}_{=q_i(\lambda)}, \tag{7.6}$$

where $q_i \;:\; \mathbb{R}\to\mathbb{R}$ is a convex function since $f_i^*$ is a convex function [106]. We assume the following Slater's condition to guarantee for the strong duality.

**Assumption 11** (Slater's condition [40])**.** *There exists a point $\tilde{\mathbf{x}}$ that belongs to the relative interior of $\mathcal{X}$ and satisfies $\sum_{i\in\mathcal{V}}\tilde{x}_i = b$.*

As remarked, Lagrangian methods [107] require a central coordinator to update and distribute the multiplier $\lambda$ to the nodes. The key idea of our approach is to eliminate this requirement by utilizing the distributed bgradient method in Section 2.2.2 to compute the solution of problem (7.6). In particular, we have each node $i$ stores a local copy $\lambda_i$ of the Lagrange multiplier $\lambda$, and then iteratively update $\lambda_i$ upon communicating with its neighbors. The update of $\lambda_i$ in Eq. (7.4) is the distributed gradient update in Eq. (2.16), where the subgradient of $q_i$ at $v_i(k+1)$ is given by

$$b_i - x_i(k+1) \in \partial q_i(v_i(k+1)). \tag{7.7}$$

These two steps, coupled with the primal update in the Lagrangian approach, results in our distributed Lagrangian algorithm as presented in Algorithm 4.

### 7.2.2 Convergence Analysis

We now present our main result on the convergence results of Algorithm 4. For the ease of exposition, we present their proofs in the appendix. We denote by $\mathcal{L}_i : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ the local Lagrangian function at node $i$

$$\mathcal{L}_i(x_i, v_i) = f_i(x_i) + v_i(x_i - b_i). \tag{7.8}$$

We show that the sequence of each dual variable $\{\lambda_i(k)\}$, for all $i\in\mathcal{V}$, converges to a dual solution of DP under some proper choice of the sequence of stepsizes $\{\alpha(k)\}$. This result is formally stated in the following theorem.

**Theorem 14.** *Suppose that Assumptions 2, 4, and 11 hold. Let the sequences $\{x_i(k)\}$ and $\{\lambda_i(k)\}$, for all $i\in\mathcal{V}$, be generated by Algorithm 4. Assume that* *the stepsizes $\alpha(k)$ are non-increasing, $\alpha(0) = 1$, and satisfies the following conditions*

$$\sum_{k=1}^{\infty} \alpha(k) = \infty, \qquad \sum_{k=1}^{\infty} \alpha^2(k) < \infty. \tag{7.9}$$

*Then we have*

*(a)* $\lim\limits_{k\to\infty} \lambda_i(k) = \lambda^*$, *for all* $i \in \mathcal{V}$, *where* $\lambda^*$ *is an optimizer of* $\mathsf{DP}$.

*(b)* $\lim\limits_{k\to\infty} \sum_{i\in\mathcal{V}} \mathcal{L}_i(x_i(k), \lambda_i(k))$ *is the optimal value of* $\mathsf{P}$.

A specific choice for the sequence of stepsizes is $\alpha(k) = 1/k$ for $k \geq 1$, which obviously satisfies the conditions in Eq. (7.9). Part (a) of Theorem 14 is a consequence of Theorem 5, while part (b) can be derived using the strong duality of $\mathsf{P}$ and $\mathsf{DP}$. A key step in showing part $(a)$ requires that $\partial q_i(v_i(k))$ remains bounded for all $k \geq 0$. Given the compactness of $\mathcal{X}_i$ and by Eq. (7.7), this boundedness condition is satisfied; this is formally stated in the following.

**Proposition 2.** *Let the sequences $\{x_i(k)\}$ and $\{\lambda_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 4. Then there exists a positive constant $C_i$ such that*

$$\left|\partial q_i(v_i(k))\right| \leq C_i, \qquad \forall\, k \geq 0,\ \forall\, i \in \mathcal{V}. \tag{7.10}$$

Note that while the local copies of the dual variable $\lambda_i(k)$ tend to a dual optimizer $\lambda^*$ of $\mathsf{DP}$, Theorem 14 does not automatically imply that

$$\mathbf{x}(k) := \left(x_1^T(k), \ldots, x_n^T(k)\right)^T$$

converges to an optimal solution of $\mathsf{P}$. Such a convergence is guaranteed, however, when the functions $f_i$ are strongly convex.

On the other hand, if every node $i$ maintains a variable to track a time-weighted average of its dual variable, the distributed Lagrangian method converges at a rate $\mathcal{O}(n\ln(k)/\delta\sqrt{k})$ when the stepsizes decay as[1] $\alpha(k) = 1/\sqrt{k}$ and $\delta$ is a parameter representing the spectral properties of the network graph. This is a consequence of Theorem 6.

[1] Note that the choice of $\alpha(k) = 1/\sqrt{k}$ does not satisfy Eq. (7.9). Hence, we only establish the rate of convergence to the optimal value.

## 7.3 Distributed Randomized Lagrangian Methods

We now study problem $\mathsf{P}$ under uncertainty, where the portion of resources $b$ is unknown. Our goal in this section is to design a distributed randomized Lagrangian method, and demonstrate that this method is robust to resource uncertainty. Motivated by the analysis in Section 7.2.2, we also provide an upper bound for the rate of convergence of this method in expectation on the size and the topology of the underlying networks.

### 7.3.1 Main Algorithm

We assume the exact allotment of some resource, $b$, is unknown and we can only estimate it from noisy data. For example, power generation levels in power systems at any time are predicted from hourly day-ahead energy consumption data, which may not be accurate. Therefore, we assume that at any time $k \geq 0$ each node $i$ is able to access only a partial noisy measurement of $b$, i.e., each node $i$ can sample $\ell_i(k)$ from the data, where

$$\ell_i(k) = b_i + \eta_i(k), \quad k \geq 0, \tag{7.11}$$

and the random variables $\eta_i$ represent random fluctuations in the allocations of the resources at the nodes; the sum of constants $b_i$ represents the expected resource shared by the nodes. We note that we do not assume the constants $b_i$ are known by the nodes, thus our model is general enough to cover the case of time-varying resources. We do, however, assume that the random variables $\eta_i$ satisfy the following assumption.

**Assumption 12.** *The random variables $\eta_i$ are independent with zero mean, i.e., $\mathbb{E}[\eta_i] = 0$, for all $i \in \mathcal{V}$. Moreover, they are almost surely bounded, i.e., there exists a scalar $c_i > 0$ such that $|\eta_i| \leq c_i$, for all $i \in \mathcal{V}$, almost surely.*

This assumption implies that we only allow finite, but possibly arbitrarily large, perturbations limited by the constant $c$, in the nodes' measurements. This condition is reasonable, for example, in actual power systems the hourly day-ahead data is often approximately accurate with respect to the current consumption. Moreover, small fluctuations in loads are often seen in practice since large fluctuations may lead to a blackout condition. The condition on

**Algorithm 5** Distributed Randomized Lagrangian Method (DRLM) for Solving $\mathsf{P}$ under Uncertainty.

```
1. Initialize: Each node $i \in \mathcal{V}$ initializes $\lambda_i(0) \in \mathbb{R}$.
2. Iteration: Each node $i \in \mathcal{V}$ executes
```

$$v_i(k+1) = \sum_{j \in \mathcal{N}_i} a_{ij}(k)\lambda_j(k) \tag{7.12}$$

$$x_i(k+1) \in \arg\min_{x_i \in \mathcal{X}_i} f_i(x_i) + v_i(k+1)(x_i - \ell_i(k)) \tag{7.13}$$

$$\lambda_i(k+1) = v_i(k+1) + \alpha(k)\left(x_i(k+1) - \ell_i(k)\right). \tag{7.14}$$

zero mean implies that while being robust to the noisy measurements of $b_i$, the goal is to meet the expected number of loads defined by $\sum_{i \in \mathcal{V}} b_i$.

We now proceed to present our distributed randomized Lagrangian method for solving problem $\mathsf{P}$ under the uncertainty described above. Recall that in the distributed Lagrangian method we utilize the distributed subgradient algorithm to solve the dual problem $\mathsf{DP}$ of $\mathsf{P}$. However, due to the uncertainty of $b$, at any time $k \geq 0$ each node $i$ only has access to a partial noisy measurement of $b$ represented by $\ell_i(k)$. The nodes $i$, therefore, have to use these measurements to update their dual variables $\lambda_i(k)$, resulting in a distributed noisy subgradient update for problem $\mathsf{DP}$. The proposed distributed randomized Lagrangian algorithm is formally presented in Algorithm 5.

Algorithm 5 shares similar mechanics to Algorithm 4. A notable difference is in the step in Eq. (7.14) of Algorithm 5 where $\lambda_i$ is updated by using a noisy subgradient of $q_i$ at $v_i(k+1)$, given by

$$g_i(k+1) \triangleq \ell_i(k) - x_i(k+1), \tag{7.15}$$

that is $g_i(k)$ is a noisy measurement of the subgradient of $q_i$ at $v_i(k)$. We note that the variables $v_i$, $x_i$, and $\lambda_i$ are now random variables.

### 7.3.2 Convergence analysis

We now present our main results on the convergence of Algorithm 4. We first show that the sequence of every local copy $\{\lambda_i(k)\}$ converges to a dual solution of $\mathsf{DP}$ almost surely (*a.s.*). This result, which can be viewed as a stochastic version of Theorem 14, is formally stated in the following theorem.

**Theorem 15.** *Suppose that Assumptions 2, 4, 11, and 12 hold. Let the sequences $\{x_i(k)\}, \{\lambda_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 5. Assume that the stepsizes $\alpha(k)$ is non-increasing, $\alpha(0) = 1$, and satisfy*

$$\sum_{k=1}^{\infty} \alpha(k) = \infty, \quad \sum_{k=1}^{\infty} \alpha^2(k) < \infty. \tag{7.16}$$

*Then the sequences $\{x_i(k)\}$ and $\{\lambda_i(k)\}$ satisfy*

*(a)* $\lim_{k\to\infty} \lambda_i(k) = \lambda^*$ *a.s., for all $i \in \mathcal{V}$, where $\lambda^*$ is an optimizer of* $\mathsf{DP}$.

*(b)* $\lim_{k\to\infty} \mathbb{E}\Big[ \sum_{i\in\mathcal{V}} \mathcal{L}_i(x_i(k), \lambda_i(k))\Big]$ *is the optimal value of* $\mathsf{P}$.

Similar to Theorem 14, the key step is to show part $(a)$, which again requires that $g_i(v_i(k))$ remains bounded almost surely for all $k \geq 0$. Thank to the compactness of $\mathcal{X}_i$ and Assumption 12, this condition is guaranteed.

**Proposition 3.** *Suppose that Assumption 12 holds. Let the sequences $\{x_i(k)\}$ and $\{\lambda_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 5. Then there exists a constant $D_i > 0$, for all $i \in \mathcal{V}$, such that*

$$\Big|g_i(v_i(k+1))\Big| \leq D_i \ \ a.s., \qquad \forall\, k \geq 0,\ \forall\, i \in \mathcal{V}. \tag{7.17}$$

Finally, similar to the results in Theorem 6 we show that Algorithm 5 converges at a rate $\mathcal{O}(n\ln(k)/\delta\sqrt{k})$ to the optimal value in expectation.

**Theorem 16.** *Suppose that Assumptions 2, 4, 11, and 12 hold. Let the sequences $\{x_i(k)\}$ and $\{\lambda_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 5. Let $\alpha(k) = 1/\sqrt{k}$ for $k \geq 1$ and $\alpha(0) = 1$. Moreover, suppose that every node $i$ stores the variable $y_i(k) \in \mathbb{R}$, which is initialized arbitrarily and updated by*

$$y_i(k) = \frac{\sum_{t=0}^{k} \alpha(t)\lambda_i(t)}{\sum_{t=0}^{k}\alpha(t)}, \qquad \forall k \geq 0. \tag{7.18}$$

*Then, let $\delta \leq \min\{(1-\frac{1}{2n^3})^{1/B}, \max_{k\geq 0}\sigma_2(A(k))\}$ we have for all $i \in \mathcal{V}$*

$$\begin{aligned}\mathbb{E}\Big[q(y_i(k+1))\Big] - q^* \\ \leq \frac{\mathbb{E}\Big[\|\boldsymbol{\lambda}(0) - \lambda^*\mathbf{1}\|^2\Big]}{2\sqrt{k+1}} + \frac{D\mathbb{E}\Big[\|\boldsymbol{\lambda}(0)\|\Big]}{2(1-\delta)\sqrt{k+1}} + \frac{D^2(2+\ln(k))}{2(1-\delta)\sqrt{k+1}}\cdot\end{aligned} \tag{7.19}$$

## 7.4 Simulations

In this section, we consider case studies that demonstrate the effectiveness of the two methods proposed in Sections 4 and 5, for solving economic dispatch problems in power systems, where power flow equations between buses are ignored. We first consider the IEEE-14 bus test system [108] where we consider two cases, constant loads and uncertain loads. We then apply our method to the IEEE-118 bus test system [109], assuming a constant load. In all cases, we model the communication between nodes by a sequence of time-varying graphs. Specifically, we assume that at any iteration $k \geq 0$, a graph $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k))$ is generated randomly such that $\mathcal{G}(k)$ is undirected and connected, i.e., in Assumption 2 the constant $B = 1$. The sequence of communication matrices $\{\mathbf{A}(k)\}$ is then set equal to the sequence of lazy Metropolis matrices corresponding to $\mathcal{G}(k)$, i.e., for all $k \geq 0$,

$$\mathbf{A}(k) = [a_{ij}(k)] = \begin{cases} \frac{1}{2(\max\{|\mathcal{N}_i(k), \mathcal{N}_j(k)|\})}, & \text{if } (i,j) \in \mathcal{E}(k) \\ 0, & \text{if } (i,j) \notin \mathcal{E} \text{ and } i \neq j \\ 1 - \sum_{j \in \mathcal{N}_i(k)} a_{ij}(k), & \text{if } i = j. \end{cases} \tag{7.20}$$

Since $\mathcal{G}(k)$ are undirected and connected graphs, $\mathbf{A}(k)$ satisfy Assumption 4. Finally, for all studies the simulations are terminated when the relative errors are less than 5%, i.e., $|\lambda_i(k) - \lambda^*| < 0.05\, \lambda^*$ for all $i \in \mathcal{V}$.

### 7.4.1 Economic dispatch for IEEE 14-bus test systems

We now consider economic dispatch problems on the IEEE 14-bus test system [108]. Each generator $i$ suffers a quadratic cost as a function of the amount of its generated power $x_i$, i.e., $f_i(x_i) = c_i x_i^2 + d_i x_i$ where $c_i, d_i$ are cost coefficients of generators $i$ and $x_i \in [0, P_i^{\max}]$. The coefficients of the generators are listed in Table 7.1 which are adopted from [98]. The expected load addressed by the network is assumed to be $P = 300$(MW).

We first consider the case of constant loads, in which we initialize the generator power levels such that $\sum_{i \in \mathcal{V}} x_i(0) = P = 300$. We apply the distributed Lagrangian method to solve this dispatch problem. Simulations of this case are shown on the left of Fig. 7.1, where the top plot shows that our method achieves the optimal cost, while the bottom plot shows that the total generated power of the network, $\sum_{i \in \mathcal{V}} x_i$, meets the load demand $P = 300(MW)$.

Table 7.1: Node Parameters (MU= Monetary Units).

| Gen. | Bus | $c_i[MU/MW^2]$ | $d_i[MU/MW]$ | $P_i^{\max}[MW]$ |
|---|---|---|---|---|
| 1 | 1 | 0.04 | 2.0 | 80 |
| 2 | 2 | 0.03 | 3.0 | 90 |
| 3 | 3 | 0.035 | 4.0 | 70 |
| 4 | 6 | 0.03 | 4.0 | 70 |
| 5 | 8 | 0.04 | 2.5 | 80 |

We then consider the case of uncertain loads, where we assume that at any iteration $k \geq 0$, each node $i$ has access to a noisy measurement $b_i + \eta_i(k)$. At each iteration $k \geq 0$, $\eta_i(k)$ are generated as independent zero-mean random variables. We apply the distributed randomized Lagrangian method to this case. Similar to the deterministic case, simulations also demonstrate the convergence of our algorithm; see the plots on the right in Fig. 7.1.

### 7.4.2 Economic dispatch for IEEE 118-bus test systems

We now consider economic dispatch problems on a larger system, the IEEE-118 bus test system [109]. This system has 54 generators connected by bus lines. Each generator $i$ suffers a quadratic cost as a function of generated power $x_i$, i.e., $f_i(x_i) = c_i + d_i P_i + q_i P_i^2$. The coefficients of functions $f_i$ belong to the ranges $c_i \in [6.78, 74.33]$, $d_i \in [8.3391, 37.6968]$, and $q_i \in [0.0024, 0.0697]$. The units of $c_i, d_i$, are $MBtu$, $MBtu/MW$ and $MBtu/MW^2$, respectively.

In addition, each $x_i$ is constrained on some interval $[P_i^{\min}, P_i^{\max}]$ where these values vary as $P_i^{\min} \in [5, 150]$ and $P_i^{\max} \in [150, 400]$. The unit of power in this system is $MW$. The total load required from the system is assumed to be $P = 6000(MW)$, which is initially distributed equally to the nodes, i.e., $x_i(0) = P/54 \; \forall i \in \mathcal{V}$. We apply the distributed Lagrangian method for this study, with resulting simulations shown in Fig. 7.2. The plots in Fig. 7.2 suggest that our method is applicable to large-scale systems.

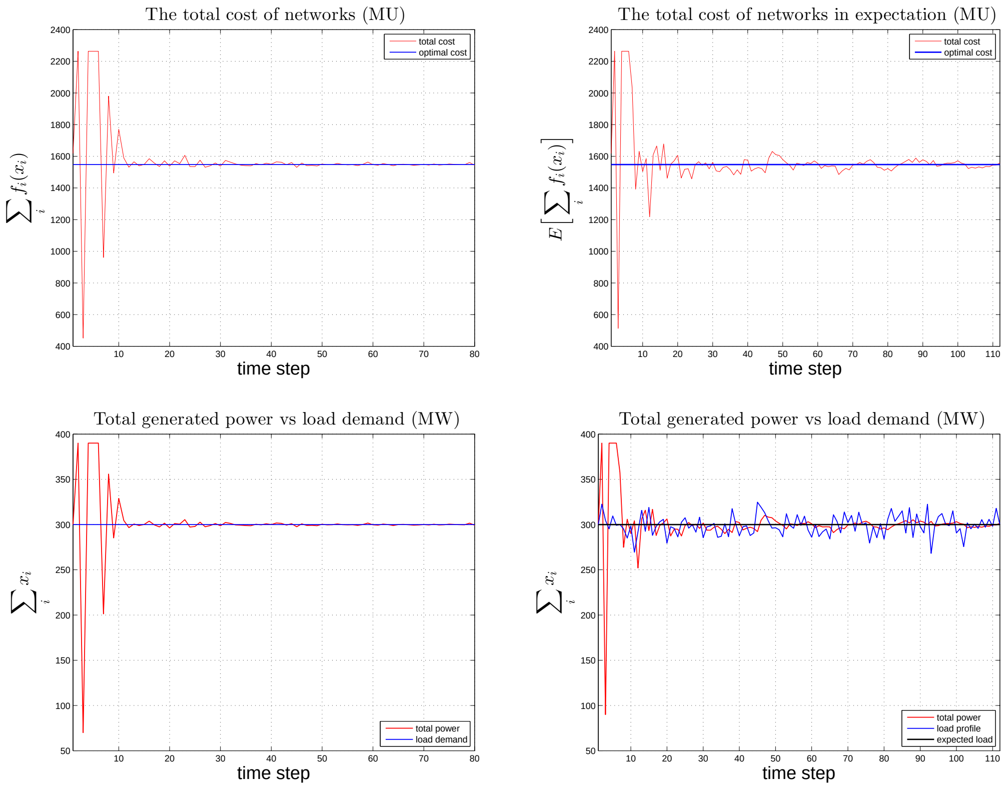

Figure 7.1: Simulations on IEEE 14 bus test system with constant loads on the left and load uncertainty on the right.

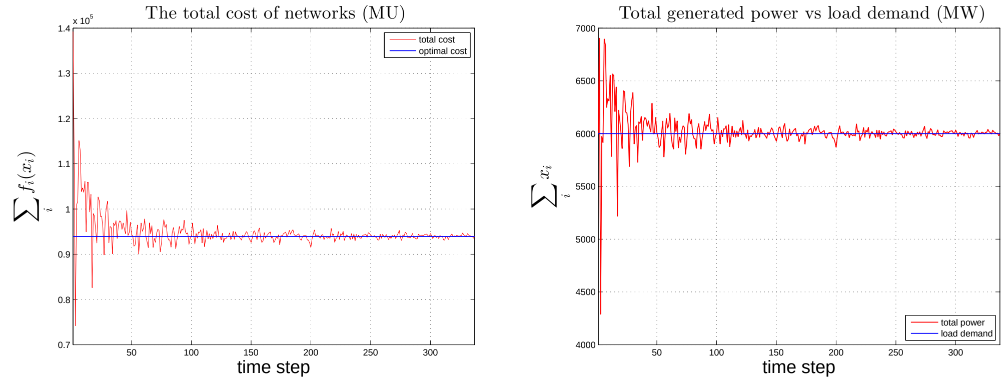

Figure 7.2: Simulations on IEEE 118 bus test system.

# Chapter 8

# Distributed Resource Allocation on Dynamic Networks in Quadratic Time

## 8.1 Problem Statement and Contribution

In this chapter, we consider problem $\mathsf{P}$ studied in Chapter 7 when $\mathcal{X}_i = \mathbb{R}$. We refer this problem to as a relaxed resource allocation problem, given as

$$\underset{x_1,\ldots,x_n}{\text{minimize}} \quad \sum_{i=1}^{n} f_i(x_i) \tag{8.1a}$$

$$\text{subject to} \quad \sum_{i=1}^{n} x_i = b. \tag{8.1b}$$

We are interested in studying distributed algorithms for solving problem (8.1) over a sequence of time-varying undirected graphs $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k))$. For simplicity, we denote by $f$ the objective function and by $\mathcal{X}$ the feasible set of problem (8.1), i.e.,

$$f(\mathbf{x}) = \sum_{i\in\mathcal{V}} f_i(x_i), \qquad \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i\in\mathcal{V}} x_i = b\}.$$

We will be assuming that there exists at least one optimal solution.

**Assumption 13.** *There exists a vector* $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ *with* $\mathbf{x}^* \in \mathcal{X}$ *which achieves the minimum of problem* (8.1).

We will use $\mathcal{X}^*$ to denote the set of optimal solutions to problem (8.1); the previous assumption ensures that $\mathcal{X}^*$ is not empty.

As will be seen shortly, the constraint relaxation, $\mathcal{X}_i = \mathbb{R}$, will allow us to exploit more about the optimality condition of problem (8.1), leading to a faster convergence algorithm. Our focus in this chapter is on designing protocols with good convergence speed. Specifically, we are interested at how the gap to the optimal objective value scales in the worst case with iteration $k$ and the number of nodes $n$ in the system.

The best previously known results were provided in the antecedent papers [101, 110]. Both papers considered the class of costs which have Lipschitz-continuous derivatives. The paper [110] considers schemes which randomly pick pairs of neighbors to perform a center-free update; if the pairs are chosen uniformly at random the convergence time implied by the results of [110] is $\mathcal{O}(Ln^4/k)$ in expectation[1] on fixed graphs; here $L$ is the largest of the Lipschitz constants of the derivatives of the cost functions. However, we note here that it is possible to shave off a factor of $n$ off this bound by adjusting the probabilities in a graph-dependent way. The paper [101] does not give an explicit convergence rate for the objective, but gives a worst-case $\mathcal{O}(LBn^3/k)$ rate for the decay of the average of *squared* gradient differences in the graph; here $B$ is a constant which measures how long it takes for the time-varying graph sequence to reach connectivity. Improved rates were obtained in [111] and in [112] for a more general problem, but under the assumption that the graph is a fixed complete graph.

**Main Contribution**. In this chapter, we show a convergence rate of $\mathcal{O}\left(LBn^2/k\right)$ for the objective under the same assumptions of Lipschitz-continuous derivatives in the more general setting of time-varying graphs. Additionally, when the costs are strongly convex, we demonstrate a geometric rate of $\mathcal{O}\left(\left(1-\mu/(4Ln^2)\right)^{k/B}\right)$ where $\mu$ is the parameter of strong convexity. For both of these rates, the number of iterations until the objective is within $\epsilon$ of its optimal value scales quadratically with the number of nodes $n$. This is an improvement over the results described above, though we note that our protocol involves every node contacting its neighbors and performing an update at every step (which involves $\mathcal{O}(|\mathcal{E}(k)|)$ messages exchanged, where $E(k)$ is the set of edges at time $k$, and $\mathcal{O}(n)$ updates); whereas [110] relied on only a pair of randomly chosen nodes updating at each step.

[1] The convergence rate in [110] is given in terms of the eigenvalues of a certain matrix; the quartic bound above follows by putting [110] together with the well-known fact that the smallest eigenvalue of the Laplacian of a connected, undirected graph on $n$ is $\Omega(1/n^2)$.

## 8.2 Gradient Balancing Protocol

In this section, we will introduce a distributed protocol, which we call the *gradient balancing protocol*, to solve problem (8.1). Before giving a statement of the algorithm, we provide some brief motivation for its form.

Previous protocols for problem (8.1) tended to be "center-free" updates [101, 110, 113, 114] where each node $i$, for all $i \in \mathcal{V}$, updates its variable as

$$x_i(k+1) = x_i(k) - \sum_{i \in \mathcal{N}_i(k)} a_{ij} \Big( f_i'(x_i(k)) - f_j'(x_j(k)) \Big), \tag{8.2}$$

where $a_{ij}$ is a collection of non-negative weights. The protocol of [110] had a different form but proceeded in the same spirit; in that protocol, edges were repeatedly chosen according to some probability distribution and a form of the above update was performed by the incident nodes.

The protocol we propose in this section speeds up this update by employing some local "pruning" wherein each node tries to perform a version of Eq. (8.2), but only with the two nodes whose derivative is largest and smallest in its neighborhood. Thus nodes essentially ignore neighbors whose derivatives are close to their own. Intuitively, by focusing on nodes whose derivatives are far apart we increase the speed at which information propagates through the network. The idea has been previously used in [45] and is inspired by an algorithm from [16, Chapter 7].

We now describe the steps node $i$ executes at step $k$ to update its value from $x_i(k)$ to $x_i(k+1)$. We assume that all nodes execute these steps synchronously, and furthermore that all four steps of the protocol given below can be executed before the graph changes from $\mathcal{G}(k)$ to $\mathcal{G}(k+1)$. Speaking informally, the protocol consists of each node repeatedly trying to "match" itself to the node in its neighborhood whose derivative is smallest and smaller than its own in order to perform a center-free update. Finally, we will be assuming that our algorithm starts from a feasible point.

**Assumption 14.** $\mathbf{x}(0) \in \mathcal{X}$.

### The Gradient Balancing Protocol

1. Node $i$ broadcasts the value $f_i'(x_i(k))$ and Lipschitz constant $L_i$ to its neighbors.

2. Going through the messages it has received from neighbors as a result of step 1, node $i$ finds the neighbor with the smallest derivative that is strictly less than its own. Let $p$ be a neighbor with this derivative; ties can be broken arbitrarily. Formally,

$$p \in \arg\min_j \Big\{ f_j'(x_j(k)) \ \mid \ j \in N_i(k), \quad f_j'(x_j(k)) < f_i'(x_i(k)) \Big\}.$$

   Node $i$ then sends a message to node $p$ the quantity

$$\Delta_{ip}(k) = \frac{1}{2} \frac{f_i'(x_i(k)) - f_p'(x_p(k))}{L_i + L_p}.$$

   If no neighbor of $i$ has a derivative strictly less than $f_i'(x_i(k))$, node $i$ does nothing during this step.

3. Node $i$ goes through any $\Delta_{ji}(k)$ it has just received from its neighbors $j \in N_i(k)$ as a result of step 2, and finds the largest among them; ties can be broken arbitrarily. Let us suppose this is $\Delta_{qi}(k)$. Node $i$ then sets

$$y_i(k) = x_i(k) + \Delta_{qi}(k).$$

   Furthermore, node $i$ sends an "accept" message to node $q$ and a "reject" message to any other neighbor $j$ that sent it a $\Delta_{ji}(k)$ in step 2.
   If node $i$ did not receive any $\Delta_{ji}(k)$ in step 2, it sets $y_i(k) = x_i(k)$.

4. If node $i$ did not send out $\Delta_{ip}(k)$ during step 2, or if it received a "reject" from the node $p$ to whom it sent $\Delta_{ip}(k)$, it sets $x_i(k+1) = y_i(k)$.
   If node $i$ has received an "accept" from node $p$, it sets

$$x_i(k+1) = y_i(k) - \Delta_{ip}(k).$$

Informally, we will refer to the numbers $\Delta_{ij}(k)$ as "offers." We may summarize the gradient balancing protocol as follows. Each node $i$ makes an offer to the node with the smallest derivative (below its own) in its neighborhood, and the size of the offer is proportional to the difference of the derivatives normalized by the sum of the respective Lipschitz constants. Each node then accepts the largest offer it has received and rejects the rest. Note that each node "accepts" at most one offer and "makes" at most one offer. The final result is something like Eq. (8.2), except that the graph has been pruned to be of degree at most two and contain only edges between nodes whose derivatives are "far apart."

We remark that an immediate consequence of Assumption 14 is that $\mathbf{x}(k) \in \mathcal{S}$ for all $k \geq 0$, since every accepted offer involves an increase at the receiving node and a decrease at the offering node of the same magnitude.

For concreteness, we provide an example of our protocol; see Fig. 8.1. The top part of the figure shows $x_i(k)$ and $f_i'(x_i(k))$ for each node in parenthesis, respectively. We assume that $L_i = 1/2$ for all $i \in \mathcal{V}$. The bottom part of the figure shows the new values $x_i(k+1)$. As we can see that node $B$ and node $C$ send offers to node $D$ but node $D$ only accepts node $B$'s offer. Node $D$ also sends an offer to node $E$ and node $E$ accepts since it is the only offer it receives. Nodes $A$ and $C$ do not end up participating in any accepted offers and consequently for those nodes $x_i(k+1) = x_i(k)$.

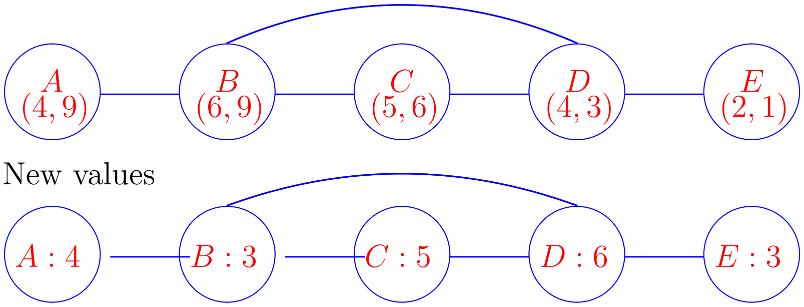

Figure 8.1: A step of the gradient balancing protocol.

## 8.3 Convergence Analysis

We now turn to the convergence analysis of the gradient balancing protocol. We will prove upper bounds on $f(\mathbf{x}(k)) - f(\mathbf{x}^*)$ which imply that the time until this quantity shrinks below $\epsilon$ is quadratic in the number of nodes $n$.

For the remainder of this chapter, we will be assuming that Assumptions 2, 4, 6, 13, and 14 hold without mention. We start this section with a characterization of the points in the optimal set $\mathcal{X}^*$; the proof is immediate.

**Proposition 4.** *We have that* $\mathbf{x} \in \mathcal{X}^*$ *if and only if* $\mathbf{x} \in \mathcal{X}$ *and* $f_i'(x_i) = f_j'(x_j)$ *for all* $i, j \in \mathcal{V}$.

Second, observe that the gradient balancing protocol may be rewritten in a particularly convenient way. Denote by $\overline{\mathcal{E}}(k)$ the set of pairs $(i,j)$ such that either $i$ accepts an offer from $j$ at time $k$ or vice versa. We can then write

$$x_i(k+1) = x_i(k) - \sum_{j \ | \ (i,j) \in \overline{\mathcal{E}}(k)} \frac{f_i'(x_i(k)) - f_j'(x_j(k))}{2(L_i + L_j)}. \tag{8.3}$$

We now begin with a series of lemmas which lead the way to our main convergence result. Our first lemma shows the monotonicity of the largest and smallest derivatives in the network.

**Lemma 11.** *The function* $\min_{i \in \mathcal{V}} f_i'(x_i(k))$ *is non-decreasing in* $k$ *and the function* $\max_{i \in \mathcal{V}} f_i'(x_i(k))$ *is non-increasing in* $k$.

*Proof.* Consider node $j$. We show that there is always some node $q$ such that

$$f_j'(x_j(k+1)) \geq f_q'(x_q(k)).$$

This will prove the monotonicity of the smallest derivative; the monotonicity of the largest derivative is proved analogously. Indeed, if $j$ does not make any offers during step 2 of the gradient balancing protocols, or if it makes an offer which is rejected, then we must have $x_j(k+1) \geq x_j(k)$. Since $f_j(\cdot)$ is convex this implies that

$$f_j'(x_j(k+1)) \geq f_j'(x_j(k)).$$

Thus we may take $q = j$ in this case. On the other hand, suppose $j$ makes an offer during step 2 which is accepted, say by node $m$. From Eq. (8.3),

$$x_j(k+1) \geq x_j(k) - \frac{f_j'(x_j(k)) - f_m'(x_m(k))}{2(L_m + L_j)},$$

and since the function $f_j(\cdot)$ is convex and $L_j$-smooth,

$$\begin{aligned} f_j'(x_j(k+1)) &\geq f_j'\left(x_j(k) - \frac{f_j'(x_j(k)) - f_m'(x_m(k))}{2(L_m + L_j)}\right) \\ &\geq f_j'(x_j(k)) - \frac{L_j\Big(f_j'(x_j(k)) - f_m'(x_m(k))\Big)}{2(L_m + L_j)} > f_m'(x_m(k)), \end{aligned}$$

so that we may take $q = m$ in this case. □

We will often need to be making statements about the $d$ largest derivatives at time $k$. To avoid overburdening the reader with notation, we will begin many of our lemmas with a variation on the words "let us relabel the vertices so that the sequence $f_1'(x_1(k)), f_2'(x_2(k), \ldots, f_n'(x_n(k))$ is non-increasing." Under this assumption, the $d$ largest derivatives may be taken to be $f_1'(x_1(k)), \ldots, f_d'(x_d(k))$.

Furthermore, assuming the nodes have been relabeled as above, we will say that *edge* $(i, j)$ *crosses the cut* $d$ if one of $i, j$ belongs to $\{1, \ldots, d\}$ while the other belongs to $\{d+1, \ldots, n\}$.

An example of the use of these definitions is in the following corollary, which is an immediate consequence of Lemma 11.

**Corollary 1.** *Let us relabel the nodes so that the sequence $f_1'(x_1(k)), f_2'(x_2(k)), \ldots, f_n'(x_n(k))$ is non-increasing. Suppose that during times $t = k, k+1, \ldots, k+T$ we have that $\overline{\mathcal{E}}(t)$ did not include any edges crossing the cut $d$. Then for $i = 1, \ldots, d$, we have that $f_i'(x_i(t+T+1)) \geq f_d'(x_d(t))$, while for $i = d+1, \ldots, n$, $f_i'(x_i(t+T+1)) \leq f_{d+1}'(x_{d+1}(t))$.*

Our next lemma essentially says that cuts in the graph which separate larger derivatives from smaller derivatives must have edges in $\overline{\mathcal{E}}(k)$ which cross them eventually. The proof follows from Assumption 4 on $B$-connectivity and is the same as the proof of Lemma 3 in [45], so we omit it.

**Lemma 12.** *Let $\ell \geq 0$ and let us relabel the nodes so that the sequence $f_1'(x_1(\ell B)), f_2'(x_2(\ell B)), \ldots, f_n'(x_n(\ell B))$ is non-increasing. Then for every $d \in \{1, \ldots, n-1\}$, either $f_d'(x_d(\ell B)) = f_{d+1}'(x_{d+1}(\ell B))$, or there exist some time $k \in \{\ell B, \ldots, (\ell+1)B - 1\}$ when an edge in $\overline{\mathcal{E}}(k)$ crosses the cut $d$.*

We now proceed to our first substantial lemma, which shows that the gradient balancing protocol is a descent protocol, i.e., $f(\mathbf{x}(k))$ is non-increasing.

**Lemma 13.**

$$f(\mathbf{x}(k+1)) \leq f(\mathbf{x}(k)) - \sum_{(i,j) \in \overline{\mathcal{E}}(k)} \frac{\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{4(L_i + L_j)}. \qquad (8.4)$$

*Proof.* Assumption 2 immediately implies that for all $x_i, y_i \in \mathbb{R}$

$$f_i(y_i) \leq f_i(x_i) + f_i'(x_i)(y_i - x_i) + \frac{L_i}{2}(y_i - x_i)^2.$$

Summing up both sides over $i \in \mathcal{V}$, we obtain

$$f(\mathbf{y}) \le f(\mathbf{x}) + \sum_{i=1}^{n} f_i'(x_i)(y_i - x_i) + \sum_{i=1}^{n} \frac{L_i}{2}(y_i - x_i)^2.$$

Replacing $\mathbf{x}$ by $\mathbf{x}(k)$, $\mathbf{y}$ by $\mathbf{x}(k+1)$, we obtain

$$\begin{aligned} f(\mathbf{x}(k+1)) \le f(\mathbf{x}(k)) &+ \sum_{i=1}^{n} f_i'(x_i(k))(x_i(k+1) - x_i(k)) \\ &+ \sum_{i=1}^{n} \frac{L_i}{2}(x_i(k+1) - x_i(k))^2. \end{aligned} \tag{8.5}$$

On the other hand, one consequence of Eq. (8.3) is that

$$\begin{aligned} &\sum_{i=1}^{n} f_i'(x_i(k))(x_i(k+1) - x_i(k)) \\ &= -\sum_{i=1}^{n} \sum_{j \;|\; (i,j)\in\overline{\mathcal{E}}(k)} \frac{f_i'(x_i(k))}{2(L_i + L_j)}\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big) \\ &= -\sum_{(i,j)\in\overline{\mathcal{E}}(k)} \frac{\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{2(L_i + L_j)}. \end{aligned} \tag{8.6}$$

Furthermore, another consequence of Eq. (8.3) is that

$$\begin{aligned} \sum_{i=1}^{n} \frac{L_i}{2}(x_i(k+1) - x_i(k))^2 &= \sum_{i=1}^{n} \frac{L_i}{2}\left( \sum_{j \;|\; (i,j)\in\overline{\mathcal{E}}(k)} \frac{f_i'(x_i(k)) - f_j'(x_j(k))}{2(L_i + L_j)} \right)^2 \\ &\le \sum_{i=1}^{n} \sum_{j \;|\; (i,j)\in\overline{\mathcal{E}}(k)} \frac{2L_i\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{8(L_i + L_j)^2} \\ &= \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \frac{\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{4(L_i + L_j)}, \end{aligned} \tag{8.7}$$

where the inequality follows because node $i$ is incident to at most two edges in $\overline{\mathcal{E}}(k)$, which allows us to use the inequality $(a+b)^2 \le 2(a^2 + b^2)$. Finally, we substitute Eqs. (8.6) and (8.7) into Eq. (8.5) gives Eq. (8.4). □

Glancing at Eq. (8.4), we might guess that the second term on the right is ultimately going to determine how fast the gradient balancing protocol will converge. Our next two lemmas provide useful lower bounds for this quantity over the time interval $k = \ell B, \ldots, (\ell+1)B - 1$.

**Lemma 14.** *Let us relabel the nodes so that the sequence $f_1'(x_1(\ell B)), \ldots, f_n'(x_n(\ell B))$, is in non-increasing order. Then,*

$$\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2 \geq \sum_{d=1}^{n-1} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2.$$

*Proof.* We begin the proof by introducing a bit more notation. For all $k \in \{\ell B, \ell B+1, \ldots, (\ell+1)B-1\}$, we use $D(k)$ to denote the set of $d \in \{1, \ldots, n-1\}$ such that time $k$ is the first time in $\{\ell B, \ell B+1, \ldots, (l+1)B-1\}$ with an edge $(i,j) \in \overline{\mathcal{E}}(k)$ crossing the cut $d$. Note that $D(k)$ may be empty. Furthermore, given the edge $(i,j) \in \overline{\mathcal{E}}(k)$ we will use $F_{ij}(k)$ to denote all the cuts $d \in D(k)$ crossed by $(i,j)$ at time $k$. Likewise, it may be the case that $F_{ij}(k)$ is empty.

We begin with the following observation. Suppose $F_{ij}(k) = \{d_1, \ldots, d_q\}$ where $d_1 < d_2 < \cdots < d_q$. Then

$$\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2 \geq \sum_{d\in F_{ij}(k)} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2. \quad (8.8)$$

We now justify Eq. (8.8). Indeed, since $d_1 \in F_{ij}(k)$, we have $d_1 \in D(k)$. By definition of $D(k)$ there were no edges $(i,j)$ during times $\ell B, \ldots, k-1$ which crossed the cut $d_1$. Applying Corollary 1, we have that $f_i'(x_i(k))) \geq f_{d_1}'(x_{d_1}(\ell B))$ and that $f_j'(x_j(k)) \leq f_{d_q+1}'(x_{d_q+1}(\ell B))$. Therefore,

$$\begin{aligned} f_i'(x_i(k)) - f_j'(x_j(k)) &\geq f_{d_1}'(x_{d_1}(\ell B)) - f_{d_q+1}'(x_{d_q+1}(\ell B)) \\ &\geq \sum_{d\in F_{ij}(k)} f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B)). \end{aligned}$$

This implies that

$$\Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2 \geq \sum_{d\in F_{ij}(k)} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2.$$

A consequence of this last inequality is that

$$\begin{aligned}
&\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2 \\
&\geq \sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \sum_{d\in F_{ij}(k)} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2 \\
&= \sum_{k=\ell B}^{(\ell+1)B-1} \sum_{d\in D(k)} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2 \\
&= \sum_{d=1}^{n-1} \Big(f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B))\Big)^2,
\end{aligned}$$

where the final equality used the fact that every $d \in \{1,\ldots,n-1\}$ such that $f_d'(x_d(\ell B)) - f_{d+1}'(x_{d+1}(\ell B)) \neq 0$ appears in some $D(k)$, which is a restatement of Lemma 12. □

We provide below more convenient bounds than those in Lemma 14.

**Lemma 15.**

$$\begin{aligned}
\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2& \\
\geq \frac{1}{n^2}\sum_{i=1}^{n} \Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2.& \qquad (8.9)
\end{aligned}$$

*Proof.* By Lemma 14, if we relabel the nodes so that $f_1'(x_1(\ell B))$, $f_2'(x_2(\ell B)),\ldots,f_n'(x_n(\ell B))$ is non-increasing,

$$\begin{aligned}
&\frac{\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{\sum_{i=1}^{n} \Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2} \\
&\geq \frac{\sum_{i=1}^{n-1} \Big(f_i'(x_i(\ell B)) - f_{i+1}'(x_{i+1}(\ell B))\Big)^2}{\sum_{i=1}^{n} \Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2}.
\end{aligned}$$

Let $q = f_1'(x_1^*)$; by Proposition 4, we have that $q = f_i'(x_i^*)$ for all $i \in \mathcal{V}$.

Define $g_i(z) = f_i(z) - qz$. We can then rewrite the above inequality as

$$\frac{\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{\sum_{i=1}^{n} \Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2}$$
$$\geq \frac{\sum_{i=1}^{n-1} \Big(g_i'(x_i(\ell B)) - g_{i+1}'(x_{i+1}(\ell B))\Big)^2}{\sum_{i=1}^{n} g_i'^2(x_i(\ell B))}.$$

Clearly, the sequence $g_1'(x_1(\ell B)), \ldots, g_n'(x_n(\ell B)$ is in non-increasing order. It therefore follows that

$$\frac{\sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(t)} \Big(f_i'(x_i(k)) - f_j'(x_j(k))\Big)^2}{\sum_{i=1}^{n} \Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2}$$
$$\geq \min_{s_1\geq s_2\geq \ldots \geq s_n} \frac{\sum_{i=1}^{n-1} (s_i - s_{i+1})^2}{\sum_{i=1}^{n} s_i^2}. \tag{8.10}$$

Lemma 5 of [45] shows that the right-hand side is at least $1/n^2$. This immediately implies the lemma. □

We now turn to the statement and proof of our main result. We will use $R_0$ to denote a measure of initial distance to an optimal solution defined as

$$R_0 = \sup_{\mathbf{x}\in\mathcal{X}:f(\mathbf{x})\leq f(\mathbf{x}(0))} \sup_{\mathbf{x}^*\in\mathcal{X}^*} \mathbf{x} - \mathbf{x}^*\|.$$

In words, $R_0$ is the largest distance to the set of optimal solutions from any point whose objective not larger than the objective at $\mathbf{x}(0)$. Note that $R_0$ may not be finite, in which case part of our result below will be vacuously true. Our main result is then the following theorem.

**Theorem 17.**

$$f(\mathbf{x}(k)) - f(\mathbf{x}^*) \leq \frac{8LR_0^2 n^2}{\lfloor k/B \rfloor}, \tag{8.11}$$

*where $L = \max_{i\in\mathcal{V}} L_i$ and $\lfloor z \rfloor$ denotes the largest integer which is at most $z$. Furthermore, if all $f_i(\cdot)$ are $\mu$-strongly convex, i.e., Assumption 5 holds for $\mu_i = \mu \geq 0$ for all $i \in \mathcal{V}$, then we also have*

$$f(\mathbf{x}(k)) - f(\mathbf{x}^*) \leq \Big(1 - \frac{\mu}{4Ln^2}\Big)^{\lfloor k/B \rfloor} \Big(f(\mathbf{x}(0)) - f(\mathbf{x}^*)\Big). \tag{8.12}$$

*Proof.* By Lemma 13 we have

$$\begin{aligned} f(\mathbf{x}((\ell+1)B)) &\leq f(\mathbf{x}(\ell B)) - \sum_{k=\ell B}^{(\ell+1)B-1} \sum_{(i,j)\in\overline{\mathcal{E}}(k)} \frac{\Big(f_j'(x_j(k)) - f_i'(x_i(k))\Big)^2}{4(L_i+L_j)} \\ &\leq f(\mathbf{x}(\ell B)) - \sum_{i=1}^{n} \frac{\Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2}{8Ln^2}, \qquad (8.13) \end{aligned}$$

where the last step is due to Lemma 15 and the inequality $L_i + L_j \leq 2L$. Next, since $f$ is convex we have

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}(\ell B)) &\geq \Big(\mathbf{x}^* - \mathbf{x}(\ell B)\Big)^T \nabla f(\mathbf{x}(\ell B)) \\ &= \Big(\mathbf{x}^* - \mathbf{x}(\ell B)\Big)^T \Big(\nabla f(\mathbf{x}(\ell B)) - \nabla f(\mathbf{x}^*)\Big), \end{aligned}$$

where the last equality follows since, by Proposition 4, the components of $\nabla f(\mathbf{x}^*)$ are identical and since $\mathbf{x}(\ell B),\ \mathbf{x}^* \in S$, we have that the entries of $\mathbf{x}^* - \mathbf{x}(\ell B)$ sum to zero. Next, negating both sides of the above equation and using the Cauchy-Schwarz inequality

$$f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*) \leq R_0 \Big\|\nabla f(\mathbf{x}(\ell B)) - \nabla f(\mathbf{x}^*)\Big\|, \qquad (8.14)$$

where we used that $f(\mathbf{x}(k))$ is non-increasing. Combining Eqs. (8.9) and (8.14) we have

$$\begin{aligned} f(\mathbf{x}((\ell+1)B)) - f(\mathbf{x}^*) &\leq f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*) - \sum_{i=1}^{n} \frac{\Big(f_i'(x_i(\ell B)) - f_i'(x_i^*)\Big)^2}{8Ln^2} \\ &\leq f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*) - \frac{\Big(f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*)\Big)^2}{8Ln^2R_0^2}. \qquad (8.15) \end{aligned}$$

We now show the last inequality implies Eq. (8.11) via some standard equation manipulations. Letting $\Delta(k) = f(\mathbf{x}(k)) - f(\mathbf{x}^*)$, note that $\Delta(k)$ is non-increasing by Lemma 13. We have just shown

$$\Delta((\ell+1)B) \leq \Delta(\ell B) - \frac{\Delta^2(\ell B)}{8LR_0^2n^2}.$$

Dividing both sides of this by $\Delta((\ell+1)B)\Delta(\ell B)$ and rearranging, we obtain

$$\frac{1}{\Delta((\ell+1)B)} \geq \frac{1}{\Delta(\ell B)} + \frac{1}{8LR_0^2n^2}\frac{\Delta(\ell B)}{\Delta((\ell+1)B)} \geq \frac{1}{\Delta(\ell B)} + \frac{1}{8LR_0^2n^2},$$

where we used the monotonicity of $\Delta(k)$. Summing this inequality over $\ell = 0, \ldots, k-1$, we obtain

$$f(\mathbf{x}(kB)) - f(\mathbf{x}^*) \leq \frac{8LR_0^2n^2}{k},$$

and using the monotonicity of $f(\mathbf{x}(k))$ we obtain Eq. (8.11).

Turning now to Eq. 8.12, let us define as before $g_i(x) = f_i(x) - qx$ where $q = f_1'(x_1^*)$, and further let $G(\mathbf{x}) = \sum_{i=1}^n g_i(x_i)$. Observe that $G(\mathbf{x})$ is $\mu$-strongly convex and has global minimizer at $\mathbf{x}^*$. Consequently if $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}\sum_{i=1}^n \Big(f_i'(x_i) - f_i'(x_i^*)\Big)^2 = ||\nabla G(\mathbf{x})||^2 &\geq 2\mu(G(\mathbf{x}) - G(\mathbf{x}^*)) \\ &= 2\mu(f(\mathbf{x}) - f(\mathbf{x}^*)),\end{aligned}$$

where the final equality used the fact that the sum of the entries of $\mathbf{x}$ and $\mathbf{x}^*$ is the same since both are in $\mathcal{X}$. Thus from Eq. (8.15),

$$f(\mathbf{x}((\ell+1)B)) - f(\mathbf{x}^*) \leq f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*) - \sum_{i=1}^n \frac{2\mu\left(f(\mathbf{x}(\ell B)) - f(\mathbf{x}^*)\right)}{8Ln^2},$$

which immediately implies Eq. (8.12). □

**Remark 3.** *Note that although Eq.* (8.1) *does not have constraints on the variables $x_i$, for certain functions $f_i(x_i)$ our algorithm automatically solves a constrained version of the problem. For example, if the initial conditions $x_i(0)$ are all non-negative and $f_i'(0) = f_j'(0)$ for all $i, j$, then by Lemma 11 the constraint $x_i \geq 0$ will automatically be satisfied throughout the execution of the gradient balancing method. In other words, the constraints $x_i \geq 0$ can be added "for free." The condition on the functions $f_i(x)$ is somewhat restrictive, but admissible $f_i$ include, for example, all polynomials with non-negative coefficients whose linear coefficient is zero.*

## 8.4 Simulations

We now describe a simulation of the gradient balancing protocol on some particular graphs. We consider local objective functions

$$f_i(x_i) = w_i(x_i - a_i)^4,$$

where the non-negative coefficient $w_i$ and the coefficient $a_i$ are chosen uniformly on $[0, 1]$. We set $b = 0$. We show simulations of the line and lollipop graphs in Fig. 8.2, where we plot the first time $f(\mathbf{x}(k)) - f(\mathbf{x}^*) < 1/100$ on the y-axis. The figures appear to be broadly consistent with the quadratic bound of Theorem 17.

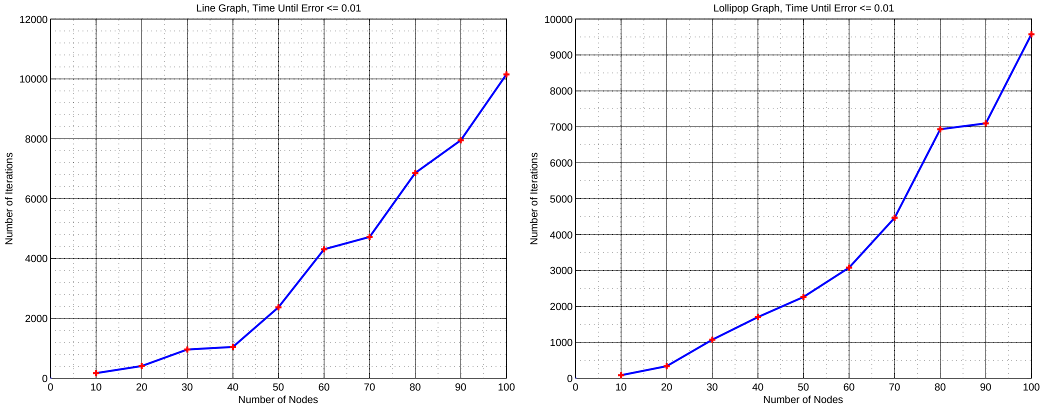

Figure 8.2: Convergence time for gradient balancing as a function of the number of nodes for the line graph on the left and the lollipop graph on the right.

# Chapter 9

# Concluding Remarks

In this thesis, we have studied distributed algorithms for solving network optimization problems, where the focus is on understanding the performance of distributed gradient methods under practical considerations, such as communication delays, random projections, and resource uncertainty. The main contributions of the thesis is summarized as below.

1. In Chapter 3, we provide an explicit formula for the rate of convergence of distributed gradient methods under communication delays, a critical issues in distributed systems. We studied both continuous-time and discrete-time variants of these methods.

2. To improve the convergence of distributed gradient methods, we study distributed aggregated gradient methods and distributed mirror descent methods in Chapters 4 and 5, respectively. We analyze the convergence of these two methods and provide numerical simulations to show that they outperform distributed gradient methods.

3. In Chapter 6 our focus is to study distributed random projection approaches for master-worker architectures. Our main observation is that distributed random projection shares the same convergence rate as distributed stochastic gradient methods, except for a constant factor capturing the regularity condition of the constraint sets.

4. In Chapter 7, we consider network resource allocation problems where we propose distributed Lagrangian methods for solving these problems through utilizing distributed gradient methods studied in Chapter 2. On the other hand, we study the relaxed variant of network resource allocation problems in Chapter 8, where our main contribution is to design the distributed gradient balancing protocol for solving this relaxed problem. In addition, we show that our algorithm achieves a quadratic convergence

time, which is an improvement over the existing results by a factor of $n$, the number of nodes in the network.

Although we have been able to address some important questions in the area of distributed algorithms, many practical challenges remain unsolved. We provide here a list of such challenges, which we leave for our future studies.

1. A natural question left by Chapter 3 is whether the results of uniform delays can be generalized when delays are heterogeneous or time-varying, conditions which often arise in highly noisy environments, for example, in mobile sensor networks.

   A second challenge is to investigate the impact of packet drops in interprocessor communication. Can we distinguish between packet drops and communication delays among the processors?

2. A question left by Chapter 4 is the convergence rate of distributed aggregated gradient methods when the noise is multiplicative, that is, can we achieve an asymptotic convergence with a linear rate? In addition, what is the convergence rate when there are constraints distributed over the network? Can we utilize the condition of set regularity in Chapter 6 to achieve the same rate as in unconstrained problems?

3. In Chapter 6, we have observed through simulations that increasing batch size does improve the performance of distributed random projection, which is similar to observations in distributed SGD. However, it has been observed that SGD suffers a phenomenon known as speedup saturation, that is, the algorithms converge slower when the batch size is greater than a certain threshold. Thus, an interesting question left by our work is to characterize the impact of batch sizes on the performance of distributed random projection?

4. Can we improve the convergence rate of distributed Lagrangian method for solving network resource allocation problems when the objective problems are strongly convex? The sublinear convergence rate established in Chapter 7 is much slower than the linear rate of distributed gradient balancing protocols studied in Chapter 8 for the relaxed problem.

5. Finally, recall that all the results in this thesis are designed based on critical assumptions regarding convexity of the problems. A natural question to ask is whether we can generalize these results for nonconvex problems, since practical problems, such as, optimal power flow problems in power networks and distributed estimations in machine learning are purely nonconvex. Nonconvexity of optimization problems implies a combinatorial structure, which often makes the computation fundamentally intractable for general problems, as compared to its convex counterpart. However, we may be able to utilize the geometric structure of individual problems through duality theory, leading to a good approximation to the global optimal solution with some complexity guarantees.

# Appendix A

# Proofs of Section 2.2 in Chapter 2

We provide here the proofs of Lemma 1, and Theorems 3 and 4 in Section 2.2. In the sequel, let $\mathbf{y} = \mathbf{x} - \bar{x}\mathbf{1}$ and consider the following notation,

$$F(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(x_i), \quad \nabla F(\mathbf{x}) \triangleq [f_1'(x_1), \ldots, f_n'(x_n)]^T, \quad C \triangleq \sum_{i=1}^{n} C_i.$$

## A.1 Proof of Lemma 1

1. By Eq. (2.17) we have

$$\dot{\bar{x}}(t) = -\bar{x}(t) + +\frac{\alpha(t)}{n}\sum_{i=1}^{n} f_i'(x_i(t)).$$

Thus, using $\mathbf{L1} = 0$ and let $\mathbf{g}(t) = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t))$ we have

$$\dot{\mathbf{y}}(t) = -L\mathbf{y}(t) - \alpha(t)\mathbf{g}(t). \tag{A.1}$$

Using the Cauchy-Schwarz inequality and Assumption 7 gives

$$\|\mathbf{g}(t)\|^2 = \left\|\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla f(\mathbf{x}(t))\right\|^2 \leq \|\nabla f(\mathbf{x}(t))\|^2 \overset{(2.11)}{\leq} C^2, \tag{A.2}$$

which implies that $\|\mathbf{g}(t)\| \leq C$. Second, Eq. (A.1) gives

$$\mathbf{y}(t) = e^{-\mathbf{L}t}\mathbf{y}(0) - \int_0^t \alpha(u)e^{-\mathbf{L}(t-u)}\mathbf{g}(u)du.$$

Applying the 2-norm to the preceding relation gives

$$\|\mathbf{y}(t)\| \leq \|e^{-\mathbf{L}t}\mathbf{y}(0)\| + \int_0^t \|\alpha(u)e^{-\mathbf{L}(t-u)}\mathbf{g}(u)\|du,$$

where using Eq. (A.2) and $\|\mathbf{y}(0)\| \leq \|\mathbf{x}(0)\|$ yields

$$\|\mathbf{y}(t)\| \overset{(2.6)}{\leq} e^{-\lambda_2 t}\|\mathbf{x}(0)\| + C\int_0^t \alpha(u)e^{-\lambda_2(t-u)}du. \tag{A.3}$$

This completes our proof of Eq. (2.18).

2. Suppose that $\{\alpha(t)\}$ is a non-increasing positive scalar sequence such that $\lim_{t\to\infty}\alpha(t) = 0$. We first show that

$$\lim_{t\to\infty}\int_0^t \alpha(u)e^{-\lambda_2(t-u)}du = 0. \tag{A.4}$$

Indeed, since $\alpha(t)$ is non-increasing with $\alpha(0) = 1$ we have

$$\begin{aligned}\lim_{t\to\infty}\int_0^t \alpha(u)e^{-\lambda_2(t-u)}du &= \lim_{t\to\infty}\Big(\int_0^{t/2}\alpha(u)e^{-\lambda_2(t-u)}du + \int_{t/2}^t \alpha(u)e^{-\lambda_2(t-u)}du\Big)\\ &\leq \lim_{t\to\infty}\Big(\int_0^{t/2}e^{-\lambda_2(t-u)}du + \alpha(t/2)\int_{t/2}^t e^{-\lambda_2(t-u)}du\Big)\\ &= \lim_{t\to\infty}\frac{e^{-\lambda_2(t-t/2)} - e^{-\lambda_2 t}}{\lambda_2} + \lim_{t\to\infty}\alpha(t/2)\frac{1-e^{-\lambda_2(t-t/2)}}{\lambda_2} = 0. \end{aligned} \tag{A.5}$$

Using Eqs. (A.3) and (A.5) gives Eq. (2.19), i.e.,

$$\lim_{t\to\infty}\|\mathbf{y}(t)\| \leq \lim_{t\to\infty}e^{-\lambda_2 t}\|\mathbf{x}(0)\| + \lim_{t\to\infty}C\int_0^t \alpha(u)e^{-\lambda_2(t-u)}du = 0.$$

3. Suppose further that $\int_0^\infty \alpha^2(t)dt < \infty$. Integrating Eq. (A.3) gives

$$\begin{aligned}&\int_0^t \alpha(t)\|\mathbf{y}(t)\|dt\\ &\qquad\leq \int_0^t \alpha(t)e^{-\lambda_2 t}\|\mathbf{x}(0)\|dt + C\int_0^t \alpha(u)\int_0^u \alpha(s)e^{-\lambda_2(u-s)}dsdu.\end{aligned} \tag{A.6}$$

First, the first term on the right-hand side of above is bounded by

$$\int_0^t \alpha(t)e^{-\lambda_2 t}\|\mathbf{x}(0)\|dt \leq \|\mathbf{x}(0)\|\int_0^t e^{-\lambda_2 t}dt \leq \frac{\|\mathbf{x}(0)\|}{\lambda_2}\cdot \tag{A.7}$$

Second, using $\lambda_2 \in (0,1)$ we obtain a bound for the second term on the right-hand side of Eq. (A.6)

$$\begin{aligned}
&C\int_0^t \alpha(u)\int_0^u \alpha(s)e^{-\lambda_2(u-s)}dsdu \leq C\int_0^t\int_0^u \alpha^2(s)e^{-\lambda_2(u-s)}dsdu\\
&= C\int_0^t\int_0^{u/2} \alpha^2(s)e^{-\lambda_2(u-s)}dsdu + C\int_0^t\int_{u/2}^u \alpha^2(s)e^{-\lambda_2(u-s)}dsdu\\
&\leq C\int_0^t\int_0^{u/2} e^{-\lambda_2(u-s)}dsdu + C\int_0^t \alpha^2(u/2)\int_{u/2}^u e^{-\lambda_2(u-s)}dsdu\\
&= C\int_0^t e^{-\lambda_2 u}\frac{e^{\lambda_2 u/2}-1}{\lambda_2}du + C\int_0^t \alpha^2(u/2)e^{-\lambda_2 u}\frac{e^{\lambda_2 u}-e^{\lambda_2 u/2}}{\lambda_2}du\\
&\leq C\int_0^t \Big(\frac{e^{-\lambda_2 u/2}}{\lambda_2}+\frac{\alpha^2(u/2)}{\lambda_2}\Big)du \leq \frac{2C}{(\lambda_2)^2}+\frac{C}{\lambda_2}\int_0^t \alpha^2(u/2)du. \qquad \text{(A.8)}
\end{aligned}$$

Substituting Eqs. (A.7) and (A.8) into Eq. (A.6), and letting $t \to \infty$ give

$$\begin{aligned}
\int_0^\infty \alpha(t)\|\mathbf{y}(t)\|dt &\leq \frac{\|\mathbf{x}(0)\|}{\lambda_2}+\frac{2C}{(\lambda_2)^2}+\frac{C}{\lambda_2}\lim_{t\to\infty}\int_0^t \alpha^2(u/2)du\\
&= \frac{\|\mathbf{x}(0)\|}{\lambda_2}+\frac{2C}{(\lambda_2)^2}+\frac{2C}{\lambda_2}\lim_{t\to\infty}\int_0^{t/2} \alpha^2(u)du\\
&= \frac{\|\mathbf{x}(0)\|}{\lambda_2}+\frac{2C}{(\lambda_2)^2}+\frac{2C}{\lambda_2}\int_0^\infty \alpha^2(t)dt < \infty. \qquad \text{(A.9)}
\end{aligned}$$

This completes the proof of Lemma 1.

## A.2 Proof of Theorem 3

Let $x^* \in \mathcal{X}^*$. Consider a candidate Lyapunov function $V : \mathbb{R} \to \mathbb{R}$ given as,

$$V(\bar{x}(t)) = \frac{1}{2}(\bar{x}(t)-x^*)^2. \qquad \text{(A.10)}$$

Recall that $f(x) = \sum_{i\in\mathcal{V}} f_i(x)$. The derivative of $V$ along Eq. (2.17) is

$$\begin{aligned}
\dot{V}(\bar{x}(t)) &= (\bar{x}(t)-x^*)\dot{\bar{x}} = -\frac{\alpha(t)}{n}\sum_{i=1}^n(\bar{x}(t)-x^*)f_i'(x_i(t))\\
&= -\frac{\alpha(t)}{n}\sum_{i=1}^n(\bar{x}(t)-x_i(t))f_i'(x_i(t)) - \frac{\alpha(t)}{n}\sum_{i=1}^n(x_i(t)-x^*)f_i'(x_i(t)).
\end{aligned}$$

which by using the Cauchy-Schwarz inequality on the first term and the convexity of $F$ on the second term gives

$$\begin{aligned}\dot{V}(\bar{x}(t)) &\leq \frac{\alpha(t)}{n}\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\|\sqrt{\sum_{i-1}^{n}|f_i'(x_i(t))|^2} - \frac{\alpha(t)}{n}(F(\mathbf{x}(t)) - f^*)\\ &\overset{(2.11)}{\leq} \frac{C\alpha(t)}{n}\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\| - \frac{\alpha(t)}{n}(F(\mathbf{x}(t)) - f^*)\\ &= \frac{C\alpha(t)}{n}\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\| - \frac{\alpha(t)}{n}(F(\mathbf{x}(t)) - F(\bar{x}(t)\mathbf{1}))\\ &\qquad - \frac{\alpha(t)}{n}(F(\bar{x}(t)\mathbf{1}) - f^*)\\ &\overset{(2.11)}{\leq} \frac{2C\alpha(t)}{n}\|\mathbf{x}(t) - \bar{x}(t)\mathbf{1}\| - \frac{\alpha(t)}{n}(f(\bar{x}(t)) - f^*), \end{aligned} \tag{A.11}$$

where the last inequality is due to Eq. (2.10) and we use $f(\bar{x}(t)) = F(\bar{x}(t)\mathbf{1})$. Integrating both sides of Eq. (A.11) from $t_1$ to $t_2$, for $0 \leq t_1 \leq t_2$, gives

$$V(\bar{x}(t_2)) \leq V(\bar{x}(t_1)) + \int_{t_1}^{t_2} \alpha(u)\|\mathbf{x}(u) - \bar{\mathbf{x}}(u)\|du.$$

Let $h(t) \triangleq V(\bar{x}(t)) + \int_t^\infty \alpha(u)\|\mathbf{x}(u) - \bar{\mathbf{x}}(u)\|du$. Adding both sides of the inequality above by $\int_{t_2}^\infty \alpha(u)\|\mathbf{x}(u) - \bar{\mathbf{x}}(u)\|du$ gives $h(t_2) \leq h(t_1)$. This implies that $h(t)$ is non-increasing and bounded since $h(0)$ is bounded due to Eq. (2.20). Thus, we have $h(t)$ is convergent, which gives

$$\lim_{t\to\infty} V(\bar{x}(t)) \text{ exists implying } (\bar{x}(t) - x^*)^2 \text{ converges.} \tag{A.12}$$

Integrating both sides of Eq. (A.11) again and rearranging the terms give,

$$0 \leq \int_0^\infty \frac{\alpha(t)}{n}(f(\bar{x}(t)) - f^*)dt \leq \gamma + V(\bar{x}(0)) < \infty,$$

implying $\liminf_{t\to\infty} f(\bar{x}(t)) = f^*$ since $\int_0^\infty \alpha(t)dt = \infty$. Let $\bar{x}(t_\ell)$ be a subsequence of $\bar{x}(t)$ such that

$$\lim_{t_\ell\to\infty} f(\bar{x}(t_\ell)) = \liminf_{t\to\infty} f(\bar{x}(t)) = f^*. \tag{A.13}$$

By Eq. (A.12) $\bar{x}(t)$ is bounded, which without loss of generality implies that $\bar{x}(t_\ell)$ is converging to some $\tilde{x}$ (otherwise we can in turn select a convergent subsequence of $\bar{x}(t_\ell)$). Therefore, $\lim_{t_\ell\to\infty} f(\bar{x}(t_\ell)) = f(\tilde{x})$. This implies that $\tilde{x}$ is a solution of problem (2.1) due to Eq. (A.13). By letting $x^* = \tilde{x}$ in Eq. (A.12) we obtain that $\lim_{t\to\infty} \bar{x}(t) = \tilde{x}$, which concludes our proof.

## A.3 Proof of Theorem 4

Using $\alpha(t) = 1/\sqrt{t}$ into Eq. (A.9) and since $\lambda_2 \in (0,1)$ we have

$$\begin{aligned}
&\int_0^t \alpha(t)\|\mathbf{x}(t) - \bar{x}\mathbf{1}\|dt\\
&\qquad\leq \frac{\|\mathbf{x}(0)\|}{\lambda_2} + \frac{2C}{(\lambda_2)^2} + \frac{2C}{\lambda_2}\int_0^{t/2}\alpha^2(u)du\\
&\qquad= \frac{\|\mathbf{x}(0)\|}{\lambda_2} + \frac{2C}{(\lambda_2)^2} + \frac{2C}{\lambda_2}\Big(\int_0^1\alpha^2(u)du + \int_1^{t/2}\alpha^2(u)du\Big)\\
&\qquad\leq \frac{\|\mathbf{x}(0)\|}{\lambda_2} + \frac{4C}{(\lambda_2)^2} + \frac{2C\ln(t)}{\lambda_2}. \qquad (A.14)
\end{aligned}$$

Taking the integration of Eq. (A.11) and using Eq. (A.14) gives

$$\begin{aligned}
&V(\bar{x}(t)) - V(\bar{x}(0))\\
&\leq \frac{2C}{n}\int_0^t \alpha(u)\|\mathbf{x}(u) - \bar{x}(u)\mathbf{1}\|du - \frac{1}{n}\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du\\
&\leq \frac{2C\|\mathbf{x}(0)\|}{n\lambda_2} + \frac{8C^2}{(\lambda_2)^2} + \frac{4C^2\ln(t)}{\lambda_2} - \frac{1}{n}\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du. \qquad (A.15)
\end{aligned}$$

Rearranging Eq. (A.15), dropping $V(\bar{x}(t))$, and using $\lambda_2 \in (0,1)$ we have

$$\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du \leq \frac{2C\|\mathbf{x}(0)\|}{\lambda_2} + \frac{12C^2\ln(t)}{(\lambda_2)^2} + nV(\bar{x}(0)),$$

which when diving both sides by

$$\int_0^t \alpha(u)du = \int_0^t \frac{1}{\sqrt{u}}du = 2\sqrt{t}$$

implies

$$\frac{\int_0^t \alpha(u)(f(\bar{x}(u)) - f^*)du}{\int_0^t \alpha(u)du} \leq \frac{C\|\mathbf{x}(0)\|}{\lambda_2\sqrt{t}} + \frac{6C^2\ln(t)}{(\lambda_2)^2\sqrt{t}} + \frac{nV(\bar{x}(0))}{2\sqrt{t}}.$$

By the Jensen inequality the preceding relation implies

$$f\left(\frac{\int_0^t \alpha(u)\bar{x}(u)du}{\int_0^t \alpha(u)du}\right) - f^* \leq \frac{C\|\mathbf{x}(0)\|}{\lambda_2\sqrt{t}} + \frac{6C^2\ln(t)}{(\lambda_2)^2\sqrt{t}} + \frac{nV(\bar{x}(0))}{2\sqrt{t}}\cdot \tag{A.16}$$

Recall that $S(0) = 0$, $\dot{S}(t) = \alpha(t)$ for $t > 0$, and by Eq. (2.23) we have

$$\begin{aligned}&\frac{d}{dt}(S(t)z_i(t)) = \dot{S}(t)z_i(t) + S(t)\dot{z}_i(t) \overset{(2.23)}{=} \alpha(t)x_i(t)\\ &\Rightarrow z_i(t) = \frac{\int_0^t \alpha(u)x_i(u)du}{\int_0^t \alpha(u)du} \quad \forall i \in \mathcal{V},\end{aligned}$$

which by the Lipschitz continuity of $f_i$ implies

$$\begin{aligned}f(z_i(t)) - f\left(\tfrac{\int_0^t \alpha(u)\bar{x}(u)du}{\int_0^t \alpha(u)du}\right) &\leq C\left|\tfrac{\int_0^t \alpha(u)(x_i(u)-\bar{x}(u))du}{\int_0^t \alpha(u)du}\right|\\ &\overset{(A.14)}{\leq} \frac{C\|\mathbf{x}(0)\|}{2\lambda_2\sqrt{t}} + \frac{2C^2}{(\lambda_2)^2\sqrt{t}} + \frac{C^2\ln(t)}{\lambda_2\sqrt{t}}\cdot\end{aligned} \tag{A.17}$$

Adding Eq. (A.17) into Eq. (A.16) we obtain Eq. (2.24).

# Appendix B

# Extensions of Chapter 3

We use uppercase letters in boldface for matrices. Let $\mathbf{x}_i \in \mathbb{R}^d$, for all $i \in \mathcal{V}$, and $f_i : \mathbb{R}^d \to \mathbb{R}$. We define the following notation

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \dots \\ \mathbf{x}_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \in \mathbb{R}^d, \quad \mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \dots \\ \mathbf{a}_n^T \end{pmatrix} \in \mathbb{R}^{n \times d},$$

$$\mathbf{W} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T, \quad \mathbf{Y}(t) = \mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T = \mathbf{W}\mathbf{X}(t). \tag{B.1}$$

$$F(\mathbf{X}) \triangleq \sum_{i=1}^n f_i(\mathbf{x}_i), \quad \nabla F(\mathbf{X}) = \begin{pmatrix} \nabla f_1^T(\mathbf{x}_1) \\ \dots \\ \nabla f_n^T(\mathbf{x}_n) \end{pmatrix}, \quad \mathbf{G}(t) = \mathbf{W}\nabla F(\mathbf{X}(t)).$$

Moreover, we write $\|\mathbf{A}\|_F$ as the Frobenius norm of $\mathbf{A}$. Given a matrix $\mathbf{X}$ and a set $\mathcal{X}$, we denote by $\mathcal{P}_{\mathcal{X}}[\mathbf{X}]$ the row-wise projection of $\mathbf{X}$ on $\mathcal{X}$.

## B.1 Extension to $\mathbb{R}^d$ for Continuous-Time Distributed Gradient Methods with Delays

We present here a sketch of key steps to extend our analysis for the case $d \geq 1$. We rewrite the updates in Eqs. (3.3)–(3.8) in matrix form as

$$\begin{aligned} \mathbf{V}(t) &= -\beta\mathbf{X}(t) + \beta\mathbf{A}\mathbf{X}(t-\tau) - \alpha(t)\nabla F(\mathbf{X}(t)) \\ \dot{\mathbf{X}}(t) &= \mathcal{P}_{\mathcal{T}_{\mathcal{X}(\mathbf{X}(t))}}[\mathbf{X}(t)] = \mathbf{V}(t) - \zeta(\mathbf{V}(t)) \\ \bar{\mathbf{v}}(t) &= -\beta\bar{\mathbf{x}}(t) + \beta\bar{\mathbf{x}}(t-\tau) - \frac{\alpha(t)}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(t)) \\ \dot{\bar{\mathbf{x}}}(t) &= \bar{\mathbf{v}}(t) - \bar{\zeta}(\bar{\mathbf{v}}(t)). \end{aligned}$$

We make use of the following result studied in [18], which is a general version of Lemma 4, to analyze the impact of the projection.

**Lemma 16** (Lemma 1 [18])**.** *Let $\mathcal{X}$ be a nonempty closed convex set in $\mathbb{R}^d$. Then, we have for any $\mathbf{x} \in \mathbb{R}^d$*

*(a)* $(\mathcal{P}_{\mathcal{X}}[\mathbf{x}] - \mathbf{x})^T(\mathbf{x} - \mathbf{y}) \leq -\|\mathcal{P}_{\mathcal{X}}[\mathbf{x}] - \mathbf{x}\|^2$ *for all* $\mathbf{y} \in \mathcal{X}$*.*

*(b)* $\left\|\mathcal{P}_{\mathcal{X}}[\mathbf{x}] - \mathbf{y}\right\|^2 \leq \left\|\mathbf{x} - \mathbf{y}\right\|^2 - \left\|\mathcal{P}_{\mathcal{X}}[\mathbf{x}] - \mathbf{x}\right\|^2$ *for all* $\mathbf{y} \in \mathcal{X}$*.*

We now present the analysis for the general versions of Lemma 3 and Theorem 7, which are given in the following two lemmas.

**Lemma 17.** *Suppose that Assumptions 1 and 3 hold. Let the trajectories of $\mathbf{x}_i(t)$, for all $i \in \mathcal{V}$, be updated by Algorithm 1. Let $\{\alpha(t)\}$ be a non-increasing positive scalar sequence with $\alpha(t) = 1$ for $0 \leq t \leq 1$. Moreover, let*

$$\beta \in \left(0\,, \frac{\ln(1/\sigma_2)}{\tau}\right) \qquad \textit{and} \qquad \gamma = \sigma_2 e^{\beta\tau} \in (0\,,1).$$

*Then*

1. *For all $t \geq 0$ we have*

$$\left\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\right\|_F \leq \mu(t) + \beta\sigma_2 \int_0^t e^{-\beta(1-\gamma)(t-u)}\mu(u-\tau)du, \tag{B.2}$$

*where*

$$\mu(t) = e\frac{\|\mathbf{X}(0)\|_F + 2C}{\beta}e^{-\beta t/2} + \frac{2C\alpha(t/2)}{\beta}\cdot \tag{B.3}$$

2. *If $\lim_{t\to\infty}\alpha(t) = 0$ then we have*

$$\lim_{t\to\infty}\left\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\right\| = 0, \quad \textit{for all } i \in \mathcal{V}. \tag{B.4}$$

3. *Further we have*

$$\int_0^t \alpha(u)\left\|\mathbf{X}(u) - \mathbf{1}\bar{\mathbf{x}}(u)^T\right\|_F du \leq \frac{8\big(\|\mathbf{X}(0)\|_F + 2C\big)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{4C}{\beta^2(1-\gamma)}\int_0^t \alpha^2(\gamma u/4 - \tau)du. \tag{B.5}$$

*Proof sketch.* As mentioned, the key step in the proof of Lemma 17 is to show Eq. (B.2). The analysis of Eqs. (B.4) and (B.5) are consequences of Eq. (B.2). Consider the following notation:

$$\mathbf{H}(t) = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\zeta(\mathbf{V}(t)) = \mathbf{W}\zeta(\mathbf{V}(t)).$$

We first consider

$$\begin{aligned}\dot{\mathbf{y}}_i(t) &= \dot{\mathbf{x}}_i(t) - \dot{\bar{\mathbf{x}}}(t)\\ &= -\beta\mathbf{x}_i(t) + \beta\sum_{j=1}^n a_{ij}\mathbf{x}_j(t-\tau) - \alpha(t)\nabla f_i(\mathbf{x}_i(t)) - \zeta_i(\mathbf{v}_i(t))\\ &\qquad + \beta\bar{\mathbf{x}}(t) - \beta\bar{\mathbf{x}}(t-\tau) + \frac{\alpha(t)}{n}\sum_{j=1}^n \nabla f_j(\mathbf{x}_j(t)) + \bar{\zeta}(\bar{\mathbf{v}}(t))\\ &= -\mathbf{y}_i(t) + \beta\sum_{j=1}^n a_{ij}\mathbf{y}_j(t-\tau) - \alpha(t)\mathbf{g}_i(t) - \mathbf{h}_i(t),\end{aligned}$$

which implies

$$\mathbf{y}_i(t) = e^{-t}\mathbf{y}_i(0) + \int_0^t e^{-(t-u)}\left(\beta\sum_{j=1}^n a_{ij}\mathbf{y}_j(u-\tau) - \alpha(t)\mathbf{g}_i(u) - \mathbf{h}_i(u)\right)du.$$

Thus we obtain

$$\begin{aligned}\mathbf{Y}(t) = e^{-\beta t}\mathbf{Y}(0) + \beta\int_0^t e^{-\beta(t-u)}\mathbf{A}\mathbf{Y}(u-\tau)du\\ - \int_0^t e^{-\beta(t-u)}\left(\alpha(u)\mathbf{G}(u) + \mathbf{H}(u)\right)du.\end{aligned} \tag{B.6}$$

In addition, note that $\mathbf{1}^T\mathbf{Y}(t) = \mathbf{1}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}(t) = \mathbf{0}$, implying that each column of $\mathbf{Y}(t) \notin \mathit{span}\{\mathbf{1}\}$. Indeed, if there exists at least one column of $\mathbf{Y}(t)$, namely, $\mathbf{p}_\ell(t)$, such that $\mathbf{p}_\ell(t) \in \mathit{span}\{\mathbf{1}\}$ then $\mathbf{1}^T\mathbf{p}_\ell(t) \neq 0$, but $\mathbf{1}^T\mathbf{Y}(t) = \mathbf{0}$, a contradiction. The previous observation implies that

$$\|\mathbf{A}\mathbf{Y}(t)\|_F^2 = \sum_{i=1}^n \|\mathbf{A}\mathbf{p}_i(t)\|^2 \leq \sum_{i=1}^n \sigma_2\|\mathbf{p}_i(t)\|^2 = \sigma_2\|\mathbf{Y}\|_F^2, \tag{B.7}$$

where $\mathbf{p}_i(t)$ are columns of $\mathbf{Y}(t)$. Taking the Frobenius norm on both sides of Eq. (B.6), and using Eqs. (3.24) and (B.7) we have

$$\|\mathbf{Y}(t)\|_F \leq e^{-\beta t}\|\mathbf{Y}(0)\|_F + \beta\sigma_2 \int_0^t e^{-\beta(t-u)}\|\mathbf{Y}(u-\tau)\|_F du \\ + C\int_0^t e^{-\beta(t-u)}\alpha(u)du + \int_0^t e^{-\beta(t-u)}\|\zeta(\mathbf{V}(u))\|_F du. \quad \text{(B.8)}$$

We now use Lemma 16 to construct an upper bound for the last term on the righ-hand side of Eq. (B.8). First, since $\mathbf{A}$ is doubly stochastic and $\mathbf{x}_j(t-\tau) \in \mathcal{X}$, for all $j \in \mathcal{V}$, we have

$$\sum_{j\in\mathcal{N}_i} a_{ij}\mathbf{x}_j(t-\tau) \in \mathcal{X}.$$

Thus, by Eq. (3.2) with $\theta = \beta^{-1}$ we have

$$r_i(t) = -\beta\mathbf{x}_i(t) + \beta\sum_{j\in\mathcal{N}_i} a_{ij}\mathbf{x}_j(t-\tau) \in \mathcal{D}_{\mathcal{X}}(\mathbf{x}_i(t)).$$

Hence, by Proposition 1 we have $r_i(t) \in \mathcal{T}_{\mathcal{X}}(\mathbf{x}_i(t))$. Using Lemma 16(b) gives

$$\|\mathcal{P}_{\mathcal{T}_{\mathcal{X}}(\mathbf{x}_i(t))}[\mathbf{v}_i(t)] - \mathbf{r}_i(t)\|^2 \leq \|\mathbf{v}_i(t) - \mathbf{r}_i(t)\|^2 - \|\mathcal{P}_{\mathcal{T}_{\mathcal{X}}(\mathbf{x}_i(t))}[\mathbf{v}_i(t)] - \mathbf{v}_i(t)\|^2,$$

which since $\zeta_i(\mathbf{v}_i(t)) = \mathbf{v}_i(t) - \mathcal{P}_{\mathcal{T}_{\mathcal{X}}(\mathbf{x}_i(t))}[\mathbf{v}_i(t)]$ implies

$$\|\zeta_i(\mathbf{v}_i(t))\| \leq \|\mathbf{v}_i(t) - \mathbf{r}_i(t)\| = \|\alpha(t)\nabla f_i(\mathbf{x}_i(t))\| \leq C_i\alpha(t). \quad \text{(B.9)}$$

Thus we obtain $\|\zeta(\mathbf{V}(t)) - \mathbf{1}\bar{\zeta}(\bar{\mathbf{v}}(t))^T\|_F \leq \|\zeta(\mathbf{V}(t))\|_F \leq C\alpha(t)$. Substituting the previous relation into Eq. (B.8) and using Eq. (3.29) we obtain Eq. (B.2).

□

**Lemma 18.** *Suppose that Assumptions 1 and 3 hold. Let the trajectories of $\mathbf{x}_i(t)$, for all $i \in \mathcal{V}$, be updated by Algorithm 1. Suppose that*

$$\beta \in \left(0, \frac{\ln(1/\sigma_2)}{\tau}\right) \qquad \textit{and} \qquad \gamma = \sigma_2 e^{\beta\tau} \in (0,1).$$

*Let $\alpha(t) = 1/\sqrt{t}$ for $t \geq 1$ and $\alpha(t) = 1$ for $t \leq 1$. Then for each $i \in \mathcal{V}$*

$$f\left(\frac{\int_0^t \alpha(u)\mathbf{x}_i(u)du}{\int_0^t \alpha(u)du}\right) - f^* \leq \frac{2\Gamma_0(t) + nV(\bar{x}(0))}{2(\sqrt{t}-1)}, \quad \text{(B.10)}$$

*where*

$$\Gamma_0(t) \triangleq \frac{24C\left(\|\mathbf{X}(0)\|_F + 2C\right)e^{\beta\tau/2}}{\beta^3(1-\gamma)^2} + \frac{48C^2(1+\tau)}{\beta^2\gamma(1-\gamma)} + C^2\ln(t) + \frac{48C^2\ln(\gamma t - 4\tau)}{\beta^2\gamma(1-\gamma)}\cdot \tag{B.11}$$

*Proof Sketch.* Let $\mathbf{x}^*$ be a solution of problem (2.1). Consider the candidate Krasovskii Lyapunov function given in Eq. (3.40), where its derivative is given as

$$\begin{aligned}\dot{V}(\bar{\mathbf{x}}(t)) &\leq \underbrace{-\frac{\alpha(t)}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}(t)-\mathbf{x}^*)^T\nabla f_i(\mathbf{x}_i(t))}_{W_1}\underbrace{-\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}(t)-\mathbf{x}^*)^T\zeta_i(\mathbf{v}_i(t))}_{W_2}\\ &\leq W_1 + W_2.\end{aligned} \tag{B.12}$$

The term $W_1$ can be upper bounded by Eq. (3.42). We focus on delivering the upper bound of $W_2$. Recall that $\zeta_i(\mathbf{v}_i(t)) = \mathbf{v}_i(t) - \mathcal{P}_{\mathcal{T}_{\mathcal{X}(\mathbf{x}_i(t))}}[\mathbf{v}_i(t)]$. Consider

$$\begin{aligned}W_2 &= -(\bar{\mathbf{x}}(t)-\mathbf{x}^*)^T\bar{\zeta}(\bar{\mathbf{v}}(t))\\ &= -\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\mathbf{x}}(t) - (1+\beta)\mathbf{x}_i(t) + \beta\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(t-\tau) - \mathbf{v}_i(t)\right)^T\zeta_i(\mathbf{v}_i(t))\\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_i(t) + (1+\beta)\mathbf{x}_i(t) - \beta\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(t-\tau) - \mathbf{x}^*\right)^T\zeta_i(\mathbf{v}_i(t)),\end{aligned} \tag{B.13}$$

where by Eq. (3.3) the first sum on the right-hand side is equivalent to

$$\begin{aligned}&-\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\mathbf{x}}(t) - (1+\beta)\mathbf{x}_i(t) + \beta\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(t-\tau) - \mathbf{v}_i(t)\right)^T\zeta_i(\mathbf{v}_i(t))\\ &\quad = -\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\mathbf{x}}(t) - x_i(t) + \alpha(t)\nabla f_i(\mathbf{x}_i(t))\right)^T\zeta_i(\mathbf{v}_i(t))\\ &\quad \leq \frac{1}{n}\sum_{i=1}^{n}\|\bar{\mathbf{x}}(t)-\mathbf{x}_i(t)\|\|\zeta_i(\mathbf{v}_i(t))\| + \frac{1}{n}\sum_{i=1}^{n}\alpha(t)\|\nabla f_i(\mathbf{x}_i(t))\|\|\zeta_i(\mathbf{v}_i(t))\|\\ &\quad \overset{(B.9)}{\leq} \frac{C\alpha(t)}{n}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F + \frac{C^2\alpha^2(t)}{n}\cdot\end{aligned}$$

On the other hand, let $\mathbf{r}_i(t)$ be defined as

$$\mathbf{r}_i(t) = \mathbf{x}^* - (1+\beta)\mathbf{x}_i(t) + \beta\sum_{j=1}^{n} a_{ij}\mathbf{x}_j(t-\tau).$$

Consider

$$\mathbf{x}_i(t) + \frac{1}{2}\mathbf{r}_i(t) = \frac{1-\beta}{2}\mathbf{x}_i(t) + \frac{1}{2}\mathbf{x}^* + \frac{\beta}{2}\sum_{j=1}^{n} a_{ij}\mathbf{x}_j(t-\tau) \in \mathcal{X},$$

which by Eq. (3.2) with $\theta = 1/2$ implies $\mathbf{r}_i(t) \in \mathcal{D}_{\mathcal{X}}(\mathbf{x}_i(t))$. In addition, by Proposition 1 we have $\mathbf{r}_i(t) \in \mathcal{T}_{\mathcal{X}(\mathbf{x}_i(t))}$. Thus, by applying Lemma 16(1a) into the second term on the right-hand side of Eq. (B.13) we obtain

$$\begin{aligned} &-\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_i(t) - \mathbf{r}_i(t)\right)^T \zeta_i(\mathbf{v}_i(t)) \\ &\qquad \leq -\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{v}_i(t) - \mathcal{P}_{\mathcal{T}_{\mathcal{X}(\mathbf{x}_i(t))}}[\mathbf{v}_i(t)]\right\|^2 = -\frac{1}{n}\|\zeta(\mathbf{V}(t))\|_F^2. \end{aligned}$$

Applying the preceding two relations into Eq. (B.13) we obtain

$$\begin{aligned} W_2 &\leq \frac{C\alpha(t)}{n}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F + \frac{C^2\alpha^2(t)}{n} - \frac{1}{n}\|\zeta(\mathbf{V}(t))\|_F^2 \\ &\leq \frac{C\alpha(t)}{n}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F + \frac{C^2\alpha^2(t)}{n}. \end{aligned} \tag{B.14}$$

Thus we obtain the same result as in Eq. (3.46), i.e.,

$$\dot{V}(\bar{x}(t)) \leq \frac{3\alpha(t)C}{n}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F + \frac{C^2\alpha^2(t)}{n} - \frac{\alpha(t)}{n}(f(\bar{\mathbf{x}}(t)) - f^*).$$

The rest of this proof is the same as the one of Theorem 7. □

## B.2 Discrete-Time Distributed Gradient Methods with Communication Delays

Here, we consider the discrete-time version of Algorithm 1 with a positive constant stepsize $\alpha$, presented in Algorithm 6. Our focus is to establish the convergence rate of this algorithm. Specifically, if each node $i$ maintains a variable $\mathbf{v}_i$ to compute the time-weighted average of its estimate $\mathbf{x}_i$, then the objective function of problem (2.1), estimated at any variable $\mathbf{v}_i$, converges to the neighborhood of the optimal value with a rate $\mathcal{O}\left(\frac{n}{k(1-\eta)}\right)$, where $\eta$ is a positive constant depending on $\sigma_2$ in Eq. (2.5) and the delay constant $\tau$. An explicit formula for $\eta$ will be given later. We first rewrite Eq. (B.21)

$$\mathbf{Z}(k) = (1-\beta)\mathbf{x}(k) + \beta\mathbf{A}\mathbf{x}(k-\tau) - \alpha\nabla F(\mathbf{x}(k)) \tag{B.15}$$

$$\mathbf{x}(k+1) = \mathbf{Z}(k) - \zeta(\mathbf{Z}(k)). \tag{B.16}$$

Moreover, since $\mathbf{A}$ is doubly stochastic we have

$$\bar{\mathbf{z}}(k) = (1-\beta)\bar{\mathbf{x}}(k) + \beta\bar{\mathbf{x}}(k-\tau) - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k)) \tag{B.17}$$

$$\bar{\mathbf{x}}(k+1) = \bar{\mathbf{z}}(k) - \bar{\zeta}(\bar{\mathbf{z}}(k)). \tag{B.18}$$

We now proceed with our analysis. The proofs presented here will hold the same merit with those in Chapter 3. The first step is to show that $\|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F$ converges to the neighborhood of zero.

**Lemma 19.** *Suppose that Assumptions 1 and 3 hold. Let $\mathbf{x}_i(k)$, for all $i \in \mathcal{V}$, be updated by Algorithm 6. Suppose that $\beta \in (0,1)$.*

1. *Then for all $k \geq 0$*

$$\begin{aligned}\|\mathbf{Y}(k)\|_F \leq (1-\beta)^{k+1}\|\mathbf{Y}(0)\|_F + \frac{2\alpha C}{\beta}\\ + \sigma_2\beta\sum_{t=0}^{k}(1-\beta)^{k-t}\|\mathbf{Y}(t-\tau)\|_F.\end{aligned} \tag{B.19}$$

2. *In addition, if $\beta \in \left(0, 1-e^{-\ln(1/\sigma_2)/\tau}\right)$ then*

$$\|\mathbf{Y}(k)\|_F \leq 2\|\mathbf{Y}(0)\|_F\eta^{k+1} + \frac{4C\alpha}{1-\eta}, \tag{B.20}$$

*where*

$$\eta = 1-\beta+\frac{\sigma_2\beta}{(1-\beta)^{\tau}} \in (1-\beta, 1)\cdot$$

**Algorithm 6** Distributed Gradient Algorithm with Communication Delays

```
1. Initialize: Each node $i$ initializes arbitrarily $x_i(k) \in \mathcal{X}$ for $k = -\tau, \ldots, 0$.
2. Iteration: For $k \geq 0$ each node $i \in \mathcal{V}$ implements
    $\mathbf{x}_i(k+1) = \mathcal{P}_{\mathcal{X}}\left[(1-\beta)\mathbf{x}_i(k) + \beta\sum_{j=1}^{n} a_{ij}\mathbf{x}_j(k-\tau) - \alpha\nabla f_i(\mathbf{x}_i(k))\right].$   (B.21)
```

*Proof.* Recall that

$$\mathbf{W} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \qquad \text{and} \qquad \mathbf{G}(t) = \mathbf{W}\nabla F(\mathbf{X}(t)).$$

1. By Eqs. (B.15) and (B.18) we have

$$\begin{aligned}\bar{\mathbf{x}}(k+1)^T &= (1-\beta)\mathbf{1}\bar{\mathbf{x}}(k)^T + \beta\mathbf{1}\bar{\mathbf{x}}(k-\tau)^T \\ &\qquad - \frac{\alpha}{n}\mathbf{1}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k))^T - \mathbf{1}\bar{\zeta}(\bar{\mathbf{z}}(k))^T \\ &= (1-\beta)\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x}(k) + \beta\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x}(k-\tau) \\ &\qquad - \frac{\alpha}{n}\mathbf{1}\mathbf{1}^T\nabla F(\mathbf{X}(k)) - \frac{1}{n}\mathbf{1}\mathbf{1}^T\zeta(\mathbf{Z}(k)).\end{aligned}$$

Thus, using Eq. (B.16) and notation in Eq. (B.1) we have

$$\mathbf{Y}(k+1) = (1-\beta)\mathbf{Y}(k) + \beta\mathbf{A}\mathbf{Y}(k-\tau) - \mathbf{G}(k) - \mathbf{W}\zeta(\mathbf{Z}(k)),$$

which when updating iteratively until $\mathbf{Y}(0)$ we obtain

$$\begin{aligned}\mathbf{Y}(k+1) &= (1-\beta)^{k+1}\mathbf{Y}(0) + \beta\sum_{t=0}^{k}(1-\beta)^{k-t}\mathbf{A}\mathbf{Y}(t-\tau) \\ &\qquad - \sum_{t=0}^{k}(1-\beta)^{k-t}\left[\mathbf{G}(t) + \mathbf{W}\zeta(\mathbf{Z}(t))\right].\end{aligned} \tag{B.22}$$

Using Eq. (B.7) gives

$$\|\mathbf{A}\mathbf{Y}(k)\|_F^2 \leq \sum_{i=1}^{n}\sigma_2\|\mathbf{y}_i(k)\|^2 = \sigma_2\|\mathbf{Y}(k)\|_F. \tag{B.23}$$

Second, Eq. (3.24) implies

$$\|\mathbf{G}(k)\|_F = \alpha\|\mathbf{W}\nabla F(\mathbf{x}(k))\|_F \leq \alpha C.$$

Third, let $\zeta_i(k) = (1-\beta)\mathbf{x}_i(k) + \beta\sum_{j=1}^n a_{ij}\mathbf{x}_j(k-\tau)$. We have $\zeta_i(k) \in \mathcal{X}$ since $\mathbf{x}_j(k-\tau) \in \mathcal{X}$, for all $j \in \mathcal{V}$, $k \geq 0$, and $\mathbf{A}$ is doubly stochastic. Thus, by Lemma 16(b), Eq. (B.21), and Eq. (B.15) we have

$$\|\mathbf{x}_i(k+1) - \zeta_i(\mathbf{z}_i(k))\|^2 \leq \|\mathbf{z}_i(k) - \zeta_i(\mathbf{z}_i(k))\|^2 - \|\mathbf{x}_i(k+1) - \mathbf{z}_i(k)\|^2.$$

Since $\zeta_i(\mathbf{z}_i(k)) = \mathbf{z}_i(k) - \mathbf{x}_i(k+1)$ the preceding relation gives

$$\|\zeta_i(\mathbf{z}_i(k))\| \leq \|\mathbf{z}_i(k) - \zeta_i(k)\| \leq C_i\alpha, \tag{B.24}$$

which implies that $\|\zeta(\mathbf{Z}(k))\|_F \leq C\alpha$. Thus, we obtain

$$\|\mathbf{W}\zeta(\mathbf{Z}(k))\|_F \leq \|\zeta(\mathbf{Z}(k))\|_F \leq C\alpha. \tag{B.25}$$

Applying the 2-norm to Eq. (B.22), using Eqs. (B.23)–(B.25) gives

$$\begin{aligned}
&\|\mathbf{Y}(k+1)\|_F \\
&\leq (1-\beta)^{k+1}\|\mathbf{Y}(0)\|_F + 2\alpha C\sum_{t=0}^k (1-\beta)^{k-t} \\
&\qquad + \sigma_2\beta\sum_{t=0}^k (1-\beta)^{k-t}\|\mathbf{Y}(t-\tau)\|_F \\
&\leq (1-\beta)^{k+1}\|\mathbf{Y}(0)\|_F + 2\alpha C(1-\beta)^k\frac{1-\left(\frac{1}{1-\beta}\right)^{k+1}}{1-\frac{1}{1-\beta}} \\
&\qquad + \sigma_2\beta\sum_{t=0}^k (1-\beta)^{k-t}\|\mathbf{Y}(t-\tau)\|_F \\
&\leq \mu(k+1) + \sigma_2\beta\sum_{t=0}^k (1-\beta)^{k-t}\|\mathbf{Y}(t-\tau)\|_F,
\end{aligned} \tag{B.26}$$

where

$$\mu(k+1) = (1-\beta)^{k+1}\|\mathbf{Y}(0)\|_F + \frac{2\alpha C}{\beta}\cdot \tag{B.27}$$

2. We now apply the delayed version of the *Grönwall-Bellman* inequality for a finite sum in Eq. (B.26). Indeed, let $w(k)$ be

$$w(k) = \sum_{t=0}^{k} (1-\beta)^{-t} \|\mathbf{Y}(t-\tau)\|_F.$$

Thus by convention we have $w(-1) = 0$ and $w(k)$ is a non-decreasing non-negative function of time. Moreover, by Eq. (B.26) we have

$$\|\mathbf{Y}(k)\|_F \leq \mu(k) + \sigma_2\beta\,(1-\beta)^{k-1}\, w(k-1).$$

Consider

$$\begin{aligned}
&w(k+1) - w(k)\\
&\quad= \sum_{t=0}^{k+1} (1-\beta)^{-t} \|\mathbf{Y}(t-\tau)\|_F - \sum_{t=0}^{k} (1-\beta)^{-t} \|\mathbf{Y}(t-\tau)\|_F\\
&\quad= (1-\beta)^{-k-1} \|\mathbf{Y}(k+1-\tau)\|,
\end{aligned}$$

which implies that

$$\begin{aligned}
w(k+1) &= (1-\beta)^{-k-1} \|\mathbf{Y}(k+1-\tau)\| + w(k)\\
&\leq (1-\beta)^{-k-1} \mu(k+1-\tau)\\
&\qquad + \sigma_2\beta\,(1-\beta)^{-k-1}(1-\beta)^{k-\tau}\, w(k-\tau) + w(k)\\
&= (1-\beta)^{-k-1} \mu(k+1-\tau)\\
&\qquad + \sigma_2\beta\,(1-\beta)^{-\tau-1}\, w(k-\tau) + w(k)\\
&\leq (1-\beta)^{-k-1} \mu(k+1-\tau)\\
&\qquad + \left(1 + \frac{\sigma_2\beta}{(1-\beta)^{\tau+1}}\right) w(k)\\
&= \sum_{t=0}^{k+1} \left(1 + \frac{\sigma_2\beta}{(1-\beta)^{\tau+1}}\right)^{k+1-t} (1-\beta)^{-t} \mu(t-\tau), \qquad \text{(B.28)}
\end{aligned}$$

where $w(-1) = 0$. Substituting Eq. (B.28) into Eq. (B.26) we have

$$\|\mathbf{Y}(k+1)\|_F \leq \mu(k+1) + \sigma_2\beta \sum_{t=0}^{k} \left(1 - \beta + \frac{\sigma_2\beta}{(1-\beta)^{\tau}}\right)^{k-t} \mu(t-\tau). \qquad \text{(B.29)}$$

Let

$$\eta = 1 - \beta + \frac{\sigma_2 \beta}{(1-\beta)^{\tau}}.$$

Since $\beta \in \left(0, 1 - e^{-\ln(1/\sigma_2)/\tau}\right)$ we have $\eta \in (0,1)$. First, using $\eta$ into the second term on the right-hand side of Eq. (B.29) gives

$$\sum_{t=0}^{k} \eta^{k-t} (1-\beta)^{t-\tau} \leq \frac{(1-\beta)^{-\tau}}{\eta + \beta - 1} \eta^{k+1} = \frac{1}{\sigma_2 \beta} \eta^{k+1}. \tag{B.30}$$

Second, we have

$$\sum_{t=0}^{k} \eta^{k-t} = \eta^{k} \frac{1 - \eta^{-k-1}}{1 - \eta^{-1}} \leq \frac{1}{1-\eta}\cdot \tag{B.31}$$

Thus, using Eqs. (B.27), (B.30), and (B.31) into Eq. (B.29) we obtain

$$\begin{aligned} &\sum_{t=0}^{k} \left(1 - \beta + \frac{\sigma_2 \beta}{(1-\beta)^{\tau}}\right)^{k-t} \mu(t-\tau) \\ &\quad = \sum_{t=0}^{k} \eta^{k-t} \left( (1-\beta)^{k-\tau} \|\mathbf{Y}(0)\|_F + \frac{2\alpha C}{\beta} \right) \\ &\quad \leq \frac{\|\mathbf{Y}(0)\|_F}{\sigma_2 \beta} \eta^{k+1} + \frac{2C\alpha}{\beta(1-\eta)}, \end{aligned}$$

which by Eqs. (B.26) and (B.27) implies Eq. (B.20).

□

Let $\mathbf{x}^* \in \mathcal{X}^*$ be a solution of problem (2.1). Consider the following lemma.

**Lemma 20.** *Suppose that Assumptions 1 and 3 hold. Let $\{\mathbf{x}_i(k)\}$, for all $i \in \mathcal{V}$, be updated by Algorithm 6 and $\mathbf{d}(k) = \bar{\mathbf{x}}(k) - \mathbf{x}^*$. Suppose that $\beta \in (0,1)$. Then for all $\ell \in \mathcal{V}$*

$$\begin{aligned} \|\mathbf{d}(k+1)\|^2 \leq \|\mathbf{d}(k)\|^2 &+ \frac{2C^2\alpha^2}{n(1-\beta)} \\ &+ \beta \left( \|\mathbf{d}(k-\tau)\|^2 - \|\mathbf{d}(k)\|^2 \right) \\ &+ \frac{6C\alpha}{n} \|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F - \frac{2\alpha}{n} \left( f(\mathbf{x}_\ell(k)) - f^* \right). \end{aligned} \tag{B.32}$$

*Proof.* By Eqs. (B.15) and (B.18) we have

$$
\begin{aligned}
\|\mathbf{d}(k+1)\|^2 &= \|\bar{\mathbf{x}}(k+1) - \mathbf{x}^*\|^2 \\
&= \left\| (1-\beta)\bar{\mathbf{x}}(k) + \beta\bar{\mathbf{x}}(k-\tau) - \mathbf{x}^* - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) - \bar{\zeta}(\bar{\mathbf{z}}(k)) \right\|^2 \\
&= \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + 2\beta(\bar{\mathbf{x}}(k) - \mathbf{x}^*)^T(\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)) \\
&\qquad - 2\left(\bar{\mathbf{x}}(k) - \mathbf{x}^*\right)^T \left( \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right) \\
&\qquad + \left\| \beta(\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)) - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) - \bar{\zeta}(\bar{\mathbf{z}}(k)) \right\|^2 \\
&= \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \beta\|\bar{\mathbf{x}}(k-\tau) - \mathbf{x}^*\|^2 \\
&\qquad - \beta\left( \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \|\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)\|^2 \right) \\
&\qquad - 2\left(\bar{\mathbf{x}}(k) - \mathbf{x}^*\right)^T \left( \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right) \\
&\qquad + \left\| \beta(\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)) - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) - \bar{\zeta}(\bar{\mathbf{z}}(k)) \right\|^2 . \qquad \text{(B.33)}
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality to the last term on the right-hand side of Eq. (B.33) gives

$$
\begin{aligned}
\|\mathbf{d}(k+1)\|^2 \le{}& \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \beta\|\bar{\mathbf{x}}(k-\tau) - \mathbf{x}^*\|^2 \\
&- \beta\left( \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \|\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)\|^2 \right) \\
&- 2\left(\bar{\mathbf{x}}(k) - \mathbf{x}^*\right)^T \left( \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right) \\
&+ \frac{\beta^2}{2}\|\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)\|^2 + 2\left\| \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right\|^2 \\
={}& \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \beta\left( \|\bar{\mathbf{x}}(k-\tau) - \mathbf{x}^*\|^2 - \|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 \right) \\
&- \beta\|\bar{\mathbf{x}}(k-\tau) - \bar{\mathbf{x}}(k)\|^2 + 2\left\| \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right\|^2 \\
&- 2\left(\bar{\mathbf{x}}(k) - \mathbf{x}^*\right)^T \left( \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k)) \right) \\
\le{}& \|\mathbf{d}(k)\|^2 + \beta(\|\mathbf{d}(k-\tau)\|^2 - \|\mathbf{d}(k)|\|^2) + H_1 + H_2, \qquad \text{(B.34)}
\end{aligned}
$$

where $H_1$ and $H_2$ are defined as

$$\begin{aligned} H_1 &= 2\left\|\frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k))\right\|^2 \\ H_2 &= -2\left(\bar{\mathbf{x}}(k) - \mathbf{x}^*\right)^T\left(\frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k))\right). \end{aligned}$$

We first analyze $H_1$. Indeed, we have

$$\begin{aligned} \left\|\frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k)) + \bar{\zeta}(\bar{\mathbf{z}}(k))\right\| &\leq \left\|\frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k))\right\| + \left\|\bar{\zeta}(\bar{\mathbf{z}}(k))\right\| \\ &\leq \frac{2\alpha C}{n}, \end{aligned}$$

which implies that

$$H_1 \leq \frac{8\alpha^2C^2}{n^2}\cdot \tag{B.35}$$

Second, we analyze $H_2$. In particular, consider the first term of $H_2$

$$\begin{aligned} &\frac{-2\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k))^T(\bar{\mathbf{x}}(k) - \mathbf{x}^*) \\ &= -\frac{2\alpha}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}_i(k))\right)^T\left(\bar{\mathbf{x}}(k) - \mathbf{x}_i(k) + \mathbf{x}_i(k)\right)^T(\mathbf{x}_i(k) - \mathbf{x}^*) \\ &\leq \frac{2\alpha}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i(k))\|\,\|\bar{\mathbf{x}}(k) - \mathbf{x}_i(k)\| - \frac{2\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k))^T(\mathbf{x}_i(k) - \mathbf{x}^*) \\ &\leq \frac{2C\alpha}{n}\|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F - \frac{2\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k))^T(\mathbf{x}_i(k) - \mathbf{x}^*)\cdot \end{aligned} \tag{B.36}$$

For some fixed $\ell \in \mathcal{V}$, we have

$$\begin{aligned} -\frac{2\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k))^T(\mathbf{x}_i(k) - \mathbf{x}^*) &\leq -\frac{2\alpha}{n}\left(F(\mathbf{X}(k)) - f^*\right) \\ &= -\frac{2\alpha}{n}\left(F(\mathbf{X}(k)) - F(\mathbf{1}\bar{\mathbf{x}}(k)^T)\right) \\ &\quad - \frac{2\alpha}{n}\left(F(\mathbf{1}\bar{\mathbf{x}}(k)^T) - f(\mathbf{x}_\ell(k))\right) - \frac{2\alpha}{n}\left(f(\mathbf{x}_\ell(k)) - f^*\right) \\ &\overset{(2.10)}{\leq} \frac{4C\alpha}{n}\|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F - \frac{2\alpha}{n}\left(f(\mathbf{x}_\ell(k)) - f^*\right). \end{aligned} \tag{B.37}$$

Substituting Eq. (B.37) into Eq. (B.36) gives

$$\frac{-2\alpha}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k))^T(\bar{\mathbf{x}}(k)-\mathbf{x}^*)$$
$$\leq \frac{6C\alpha}{n}\|\mathbf{X}(k)-\mathbf{1}\bar{\mathbf{x}}(k)^T\|_F - \frac{2\alpha}{n}\left(f(\mathbf{x}_\ell(k))-f^*\right). \qquad \text{(B.38)}$$

We now consider the second term of $H_2$. Specifically, denote by $\mathbf{p}_i(k)$

$$\mathbf{p}_i(k) = \mathbf{z}_i(k) - \beta\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(k-\tau) = (1-\beta)\mathbf{x}_i(k) - \alpha\nabla f_i(\mathbf{x}_i(k)).$$

Note that since $\mathbf{A}$ is doubly stochastic,

$$\mathbf{q}(k) = \beta\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(k-\tau) + (1-\beta)\mathbf{x}^* \in \mathcal{X}.$$

Then by Lemma 16(a) we have

$$-(\mathbf{p}_i(k)-(1-\beta)\mathbf{x}^*)^T\zeta_i(\mathbf{z}_i(k)) \leq -\left\|\mathbf{z}_i(k)-\mathcal{P}_{\mathcal{X}}[\mathbf{z}_i(k)]\right\|^2.$$

In addition, using the double stochasticity of $\mathbf{A}$ we obtain

$$\begin{aligned}
&\sum_{i=1}^{n}\|(1-\beta)\bar{\mathbf{x}}(k)-\zeta_i(k)\|\|\zeta_i(\mathbf{z}_i(k))\| \\
&\quad= \sum_{i=1}^{n}\|(1-\beta)(\bar{\mathbf{x}}(k)-\mathbf{x}_i(k))+\alpha\nabla f_i(\mathbf{x}_i(k))\|\|\zeta_i(\mathbf{z}_i(k))\| \\
&\quad\leq (1-\beta)\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}(k)-\sum_{j=1}^{n}a_{ij}\mathbf{x}_j(k)\right\|\|\zeta_i(\mathbf{z}_i(k))\| \\
&\qquad+\sum_{i=1}^{n}\|\alpha\nabla f_i(\mathbf{x}_i(k))\|\|\zeta_i(\mathbf{z}_i(k))\| \\
&\quad\leq (1-\beta)C\alpha\|\mathbf{X}(k)-\mathbf{1}\bar{\mathbf{x}}(k)^T\|_F + C^2\alpha^2\cdot
\end{aligned}$$

Thus, using the last two relations into the second term of $H_2$ gives

$$-2\left(\bar{\mathbf{x}}(k-\mathbf{x}^*\right)^T\bar{\zeta}(\bar{\mathbf{z}}(k)) \leq \frac{2C\alpha}{n}\|\mathbf{X}(k)-\mathbf{1}\bar{\mathbf{x}}(k)^T\|_F + \frac{2C^2\alpha^2}{n(1-\beta)}. \qquad \text{(B.39)}$$

Using Eqs. (B.35), (B.36) and (B.39) into Eq. (B.34) we obtain Eq. (B.32).

□

Using Lemmas 19 and 20, we now show the convergence rate of Algorithm 6. In particular, we show that the objective function $f$ of problem (2.1) converges to the neighborhood of $f^*$ with a rate $\mathcal{O}(1/k)$. Our analysis is based on considering the discrete-time candidate of the Krasovskii Lyapunov function $V$ in Chapter 3.

$$V(\bar{\mathbf{x}}(k)) = (1-\beta)\|\bar{\mathbf{x}}(k) - \mathbf{x}^*\|^2 + \beta \sum_{t=k-\tau}^{k} \|\bar{\mathbf{x}}(t) - \mathbf{x}^*\|^2. \tag{B.40}$$

**Theorem 18.** *Suppose that Assumptions 1 and 3 hold. Let the sequence $\{\mathbf{x}_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Algorithm 6 where*

$$\beta \in \left(0, 1 - e^{-\ln(1/\sigma_2)/\tau}\right).$$

*Moreover, suppose that each node $i$ maintains a variable $\mathbf{v}_i$ and updates as*

$$\mathbf{v}_i(k+1) = \frac{1}{k+1}\sum_{t=0}^{k} \mathbf{x}_i(t), \quad \forall k \geq 0. \tag{B.41}$$

*Then using $\eta$ in Lemma 19, we have for all $i \in \mathcal{V}$,*

$$f(\mathbf{v}_i(k)) - f^* \leq \frac{1}{k+1}\left(\frac{nV(\bar{\mathbf{x}}(0))}{2\alpha} + \frac{6C\|\mathbf{Y}(0)\|_F}{1-\eta}\right) + \frac{C^2\alpha}{1-\beta} + \frac{12C^2\alpha}{1-\eta}. \tag{B.42}$$

*Proof.* Let $\mathbf{x}^* \in \mathcal{X}^*$ and $\mathbf{d}(k) = \bar{\mathbf{x}}(k) - \mathbf{x}^*$. Adding both sides of Eq. (B.32) by $\beta\sum_{t=k+1-\tau}^{k} \|\bar{\mathbf{x}}(t) - \mathbf{x}^*\|$ and using the definition of $V$ in Eq. (B.40) implies

$$\begin{aligned}
&V(\bar{\mathbf{x}}(k+1))\\
&\quad \leq \|\mathbf{d}(k)\|^2 + \beta\sum_{t=k+1-\tau}^{k}\|\mathbf{d}(t)\| + \beta\left(\|\mathbf{d}(k-\tau)\|^2 - \|\mathbf{d}(k)\|^2\right)\\
&\qquad + \frac{2C^2\alpha^2}{n(1-\beta)} + \frac{6C\alpha}{n}\|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F - \frac{2\alpha}{n}\left(f(\mathbf{x}_\ell(k)) - f^*\right)\\
&\quad = V(\bar{\mathbf{x}}(k)) + \frac{2C^2\alpha^2}{n(1-\beta)} + \frac{6C\alpha}{n}\|\mathbf{X}(k) - \mathbf{1}\bar{\mathbf{x}}(k)^T\|_F\\
&\qquad - \frac{2\alpha}{n}\left(f(\mathbf{x}_\ell(k)) - f^*\right).
\end{aligned}$$

Iteratively updating the previous relation over $k$ we have

$$V(\bar{\mathbf{x}}(k+1)) \leq V(\bar{\mathbf{x}}(0)) + \frac{6C\alpha}{n}\sum_{t=0}^{k}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F + \frac{2C^2\alpha^2}{n(1-\beta)}(k+1) - \frac{2\alpha}{n}\sum_{t=0}^{k}\left(f(\mathbf{x}_\ell(t)) - f^*\right)\cdot \tag{B.43}$$

Using Eq. (B.20) and $\eta \in (0,1)$ gives

$$\sum_{t=0}^{k}\|\mathbf{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)^T\|_F \leq \frac{2\|\mathbf{Y}(0)\|_F}{1-\eta} + \frac{4C\alpha}{1-\eta}(k+1).$$

Substituting the previous relation into Eq. (B.43) we have

$$V(\bar{\mathbf{x}}(k+1)) \leq V(\bar{\mathbf{x}}(0)) + \frac{2C^2\alpha^2}{n(1-\beta)}(k+1) + \frac{24C^2\alpha^2}{n(1-\eta)}(k+1) + \frac{12C\alpha\|\mathbf{Y}(0)\|_F}{n(1-\eta)} - \frac{2\alpha}{n}\sum_{t=0}^{k}\left(f(\mathbf{x}_\ell(t)) - f^*\right).$$

Reorganizing above and dropping the non-negative term $V(\bar{\mathbf{x}}(k+1)$ imply

$$\sum_{t=0}^{k}\left(f(\mathbf{x}_\ell(t)) - f^*\right) \leq \frac{nV(\bar{\mathbf{x}}(0))}{2\alpha} + \frac{C^2\alpha}{1-\beta}(k+1) + \frac{12C^2\alpha}{1-\eta}(k+1) + \frac{6C\|\mathbf{Y}(0)\|_F}{1-\eta}.$$

Thus, dividing both sides of the preceding equation by $(k+1)$ and by the convexity of $f$ gives Eq. (B.42). □

# Appendix C

# Proofs of Lemmas 7 and 8 in Chapter 4

## C.1 Proof of Lemma 7

Recall from Eq. (4.15) that

$$\theta = \sqrt{1 - \alpha \frac{2L\mu}{n(\mu + L)}} \in (\sigma_2, 1).$$

Using Eq. (4.13) gives

$$\begin{aligned}
\mathbb{E}\Big[\big|\bar{x}(k+1) - x^*\big|\Big] &\leq \theta \mathbb{E}\Big[\big|\bar{x}(k) - x^*\big|\Big] + \frac{C\alpha}{n} + \frac{L\alpha}{n} E\Big[\big\|\mathbf{x}^\dagger(k)\big\|\Big] \\
&\overset{(4.7)}{\leq} \theta \mathbb{E}\Big[\big|\bar{x}(k) - x^*\big|\Big] + \frac{C\alpha}{n} \\
&\quad + \frac{L\alpha}{n}\left(\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]\sigma_2^k + \alpha \sum_{t=0}^{k-1} \sigma_2^{k-1-t} \mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]\right),
\end{aligned}$$

which when iterating over $k$ gives

$$\begin{aligned}
&\mathbb{E}\Big[\big|\bar{x}(k+1) - x^*\big|\Big] \\
&\quad\leq \theta^{k+1} \mathbb{E}\Big[\big|\bar{x}(0) - x^*\big|\Big] + \frac{C\alpha}{n} \sum_{t=0}^{k} \theta^t + \frac{L\alpha \mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]}{n} \sum_{t=0}^{k} \theta^{k-t} \sigma_2^t \\
&\qquad + \frac{L\alpha^2}{n} \sum_{t=0}^{k} \theta^{k-t} \sum_{\ell=0}^{t-1} \sigma_2^{t-1-\ell} \mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big] \\
&\quad\leq \theta^{k+1} \mathbb{E}\Big[\big|\bar{x}(0) - x^*\big|\Big] + \frac{C\alpha}{n(1-\theta)} + \frac{L\alpha \mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]}{n(\theta - \sigma_2)} \theta^{k+1} \\
&\qquad + \frac{L\alpha^2}{n} \sum_{t=0}^{k} \theta^{k-t} \sum_{\ell=0}^{t-1} \sigma_2^{t-1-\ell} \mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]. \qquad \text{(C.1)}
\end{aligned}$$

We provide a bound for the last term on the right-hand side of Eq. (C.1)

$$\begin{aligned}
&\sum_{t=0}^{k}\theta^{k-t}\sum_{\ell=0}^{t-1}\sigma_2^{t-1-\ell}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]\\
&\quad=\theta^k\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]\sigma_2^{-\ell-1}\sum_{t=\ell+1}^{k}\Big(\frac{\sigma_2}{\theta}\Big)^t\\
&\quad\leq\theta^k\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]\sigma_2^{-\ell-1}\frac{\big(\frac{\sigma_2}{\theta}\big)^{\ell+1}}{1-\frac{\sigma_2}{\theta}}\leq\frac{1}{\theta-\sigma_2}\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]\theta^{k-\ell}.
\end{aligned}$$

Using above and the notation in Eq. (4.15) into Eq. (C.2) gives

$$\begin{aligned}
&\mathbb{E}\Big[\big|\bar{x}(k+1)-x^*\big|\Big]\\
&\quad\leq\left(\mathbb{E}\Big[\big|\bar{x}(0)-x^*\big|\Big]+\frac{\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]}{n(1+\tau)(\gamma-\sigma_2)}\right)\theta^{k+1}\\
&\qquad+\frac{C\alpha}{n(1-\theta)}+\frac{L\alpha^2}{n(\theta-\sigma_2)}\sum_{\ell=0}^{k-1}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(\ell)\big\|\Big]\theta^{k-\ell}\\
&\quad=\beta_1\theta^{k+1}+\frac{C\alpha}{n(1-\theta)}+\beta_2\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big].
\end{aligned}\qquad\text{(C.2)}$$

Thus by Eq. (C.2) we obtain Eq. (4.16), i.e.,

$$\begin{aligned}
&E\Big[\big\|\bar{x}(k+1)\mathbf{1}-x^*\mathbf{1}\big\|\Big]+E\Big[\big\|\bar{x}(k)\mathbf{1}-x^*\mathbf{1}\big\|\Big]\\
&\quad\leq\beta_1\theta^{k+1}+\frac{C\alpha}{n(1-\sigma_2)}+\beta_2\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]\\
&\qquad+\beta_1\theta^{k}+\frac{C\alpha}{n(1-\sigma_2)}+\beta_2\sum_{t=0}^{k-2}\theta^{k-1-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]\\
&\quad\leq\beta_1\gamma^{k+1}+\frac{C\alpha}{n(1-\sigma_2)}+\beta_2\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]\\
&\qquad+\beta_1\gamma^{k}+\frac{C\alpha}{n(1-\sigma_2)}+\beta_2\theta^{-1}\sum_{t=0}^{k-2}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]\\
&\quad\leq2\beta_1\theta^{k}+\frac{2C\alpha}{n(1-\sigma_2)}+\frac{2\beta_2}{\theta}\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big].
\end{aligned}$$

## C.2 Proof of Lemma 8

By Eq. (4.2) and the triangle inequality we have

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{g}(\mathbf{x}(k+1)) - \mathbf{g}(\mathbf{x}(k))\big\|\Big] \\
&\quad \le \mathbb{E}\Big[\big\|\nabla F(\mathbf{x}(k+1)) - \nabla F(\mathbf{x}(k))\big\|\Big] + \mathbb{E}\Big[\big\|\boldsymbol{\xi}(\mathbf{x}(k+1)) - \boldsymbol{\xi}(\mathbf{x}(k))\big\|\Big] \\
&\quad \underset{(2.8)}{\overset{(4.2)}{\le}} L\mathbb{E}\Big[\big\|\mathbf{x}(k+1) - \mathbf{x}(k)\big\|\Big] + 2C \\
&\quad \le L\mathbb{E}\Big[\big\|\mathbf{x}(k+1) - \bar{x}(k+1)\mathbf{1}\big\|\Big] + L\mathbb{E}\Big[\big\|\bar{x}(k+1)\mathbf{1} - x^*\mathbf{1}\big\|\Big] \\
&\qquad\quad + L\mathbb{E}\Big[\big\|\bar{x}(k)\mathbf{1} - \mathbf{x}(k)\big\|\Big] + L\mathbb{E}\Big[\big\|x^*\mathbf{1} - \bar{x}(k)\mathbf{1}\big\|\Big] + 2C. \qquad \text{(C.3)}
\end{aligned}
$$

Using Eq. (4.7) gives

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{x}(k+1) - \bar{x}(k+1)\mathbf{1}\big\|\Big] + \mathbb{E}\Big[\big\|\bar{x}(k)\mathbf{1} - \mathbf{x}(k)\big\|\Big] \\
&\quad \le \sigma_2^{k+1}\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big] + \alpha\sum_{t=0}^{k}\sigma_2^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big] \\
&\qquad\quad + \sigma_2^{k}\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big] + \alpha\sum_{t=0}^{k-1}\sigma_2^{k-1-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big] \\
&\quad \le 2\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]\sigma_2^k + \alpha\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(k)\big\|\Big] + \frac{2\alpha}{\sigma_2}\sum_{t=0}^{k-1}\sigma_2^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big]. \qquad \text{(C.4)}
\end{aligned}
$$

Substituting Eqs. (C.4) and (4.16) into Eq. (C.4) we obtain Eq. (4.17), i.e.,

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{g}(\mathbf{x}(k+1)) - \mathbf{g}(\mathbf{x}(k))\big\|\Big] \\
&\quad \le 2C + 2L\mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]\sigma_2^k + L\alpha\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(k)\big\|\Big] + \frac{2L\alpha}{\sigma_2}\sum_{t=0}^{k-1}\sigma_2^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big] \\
&\qquad\quad + 2L\beta_1\theta^k + \frac{2LC\alpha}{n(1-\sigma_2)} + \frac{2L\beta_2}{\theta}\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big] \\
&\quad \le 2C + \frac{2LC\alpha}{n(1-\sigma_2)} + 2L\left(\beta_1 + \mathbb{E}\Big[\big\|\mathbf{x}^\dagger(0)\big\|\Big]\right)\theta^k + L\alpha\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(k)\big\|\Big] \\
&\qquad\quad + \frac{2L(\beta_2+\alpha)}{\sigma_2}\sum_{t=0}^{k-1}\theta^{k-t}\mathbb{E}\Big[\big\|\mathbf{y}^\dagger(t)\big\|\Big],
\end{aligned}
$$

where $\beta_3$ and $\beta_4$ are defined in Eq. (4.15).

# Appendix D

# Proofs of Lemmas 9 and 10 in Chapter 6

## D.1 Proof of Lemma 9

*Proof.* Since $\mathbf{x}^* \in \mathcal{X}$ then $\mathbf{x}^* \in \mathcal{X}_i$. By Eq. (6.9) we have

$$\begin{aligned}\left\|\mathbf{x}_i(k+1)-\mathbf{x}^*\right\|^2 &= \left\|\mathbf{v}_i(k)-\mathbf{x}^*-\Big(\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\Big)\right\|^2\\ &= \left\|\mathbf{v}_i(k)-\mathbf{x}^*\right\|^2+\left\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\right\|^2\\ &\quad -2\Big(\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\Big)^T\left(\mathbf{v}_i(k)-\mathbf{x}^*\right)\\ &\leq \left\|\mathbf{v}_i(k)-\mathbf{x}^*\right\|^2-\left\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\right\|^2, \qquad \text{(D.1)}\end{aligned}$$

where in the last inequality we use the projection inequality to have

$$\begin{aligned}&-2\Big(\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\Big)^T\left(\mathbf{v}_i(k)-\mathbf{x}^*\right)\\ &\quad = -2\left\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\right\|^2\\ &\qquad -2\Big(\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\Big)^T\Big(\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]-\mathbf{x}^*\Big)\\ &\quad \leq -2\left\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\right\|^2,\end{aligned}$$

Taking the expectation and averaging on both sides of Eq. (D.1) gives

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[\left\|\mathbf{x}_i(k+1)-\mathbf{x}^*\right\|^2\Big]\\ &\quad\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[\left\|\mathbf{v}_i(k)-\mathbf{x}^*\right\|^2\Big]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[\left\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\right\|^2\Big]. \qquad \text{(D.2)}\end{aligned}$$

We now use the regularity Assumption 10 to bound the last term on the right-hand side of Eq. (D.2). First, we note that $\left\|\mathbf{a}\right\|^2 \geq 2(\mathbf{a}-\mathbf{b})^T\mathbf{b}+\left\|\mathbf{b}\right\|^2$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Let $\mathbf{a} = \mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]$ and $\mathbf{b} = \bar{\mathbf{x}}(k) - \mathcal{P}_{\mathcal{X}}\big[\bar{\mathbf{x}}(k)\big]$ implying $\mathbf{e}(k) = \mathbf{b}$. This gives

$$\begin{aligned}
&\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2 \\
&\geq 2\Big(\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big] - \bar{\mathbf{x}}(k) + \mathcal{P}_{\mathcal{X}}\big[\bar{\mathbf{x}}(k)\big]\Big)^T \mathbf{e}(k) + \big\|\mathbf{e}(k)\big\|^2 \\
&\geq -2\Big(\big\|\mathbf{v}_i(k) - \bar{\mathbf{x}}(k)\big\| + \big\|\mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big] - \mathcal{P}_{\mathcal{X}}\big[\bar{\mathbf{x}}(k)\big]\big\|\Big)\big\|\mathbf{e}(k)\big\| + \big\|\mathbf{e}(k)\big\|^2,
\end{aligned}$$

where using the non-expansiveness property of the projection gives

$$\begin{aligned}
\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2 &\geq -4\big\|\mathbf{v}_i(k) - \bar{\mathbf{x}}(k)\big\|\big\|\mathbf{e}(k)\big\| + \big\|\mathbf{e}(k)\big\|^2 \\
&\overset{(6.8)}{=} -4\big\|\alpha(k) g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))\big\|\big\|\mathbf{e}(k)\big\| + \big\|\mathbf{e}(k)\big\|^2 \\
&\geq -8\alpha^2(k)\big\|g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))\big\|^2 - \frac{1}{2}\big\|\mathbf{e}(k)\big\|^2 + \big\|\mathbf{e}(k)\big\|^2 \\
&= -8\alpha^2(k)\big\|g_i(\bar{\mathbf{x}}(k), \boldsymbol{\omega}_i(k))\big\|^2 + \frac{1}{2}\big\|\mathbf{e}(k)\big\|^2, \qquad \text{(D.3)}
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality. Taking the expectation of Eq. (D.3), using Eq. (6.7), and summing over $i = 1, \ldots, n$ on both sides yields

$$\sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2\Big] \geq -8\alpha^2(k)\sum_{i=1}^{n} C_i^2 + \frac{n}{2}\mathbb{E}\Big[\big\|\mathbf{e}(k)\big\|^2\Big], \qquad \text{(D.4)}$$

By the regularity Assumption 10 we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2 \\
&\qquad \leq \frac{1}{n}\sum_{i=1}^{n} D \max_{i=1,\ldots,n} \mathbb{E}\Big[\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\big\|^2 \,|\, \mathcal{F}(k)\Big] \\
&\qquad \leq D\sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\big\|^2 \,|\, \mathcal{F}(k)\Big],
\end{aligned}$$

which by taking the expectation on both sides yields

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2\Big] \leq D\sum_{i=1}^{n} \mathbb{E}\Big[\big\|\mathbf{v}_i(k) - \mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\big\|^2\Big].$$

Using Eq. (D.4) into the preceding relation implies

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}_{i\zeta_i(k)}}\big[\mathbf{v}_i(k)\big]\big\|^2\Big]\\
&\qquad\geq \frac{1}{Dn^2}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathcal{P}_{\mathcal{X}}\big[\mathbf{v}_i(k)\big]\big\|^2\Big]\\
&\qquad\overset{(\text{D.4})}{\geq} \frac{-8\alpha^2(k)}{Dn^2}\sum_{i=1}^{n}C_i^2+\frac{1}{2Dn}\mathbb{E}\Big[\big\|\mathbf{e}(k)\big\|^2\Big]\\
&\qquad\geq -\frac{8C^2\alpha^2(k)}{Dn^2}+\frac{1}{2Dn}\mathbb{E}\Big[\big\|\mathbf{e}(k)\big\|^2\Big]. \qquad (\text{D.5})
\end{aligned}$$

Taking the expectation of Eq. (D.5) and substituting into Eq. (D.2), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{x}_i(k+1)-\mathbf{x}^*\big\|^2\Big]\\
&\qquad\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathbf{x}^*\big\|^2\Big]+\frac{8C^2\alpha^2(k)}{Dn^2}-\frac{1}{2Dn}\mathbb{E}\Big[\big\|\mathbf{e}(k)\big\|^2\Big],
\end{aligned}$$

which using the Jensen inequality on the 2-norm function gives Eq. (6.11). □

## D.2 Proof of Lemma 10

*Proof.* Since $\mathbf{x}^*\in\mathcal{X}$, by Eq. (6.8) we first have

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathbf{x}^*\big\|^2\Big]=\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*-\alpha(k)g_i(\bar{\mathbf{x}}(k),\boldsymbol{\omega}_i(k))\big\|^2\Big]\\
&\quad=\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big]+\mathbb{E}\Big[\big\|\alpha(k)g_i(\bar{\mathbf{x}}(k),\boldsymbol{\omega}_i(k))\big\|^2\Big]\\
&\qquad-2\alpha(k)\mathbb{E}\Big[(\bar{\mathbf{x}}(k)-\mathbf{x}^*)^Tg_i(\bar{\mathbf{x}}(k),\boldsymbol{\omega}_i(k))\Big]\\
&\quad\overset{(6.7)}{\leq}\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big]+C_i^2\alpha^2(k)-2\alpha(k)\mathbb{E}\Big[\big(\bar{\mathbf{x}}(k)-\mathbf{x}^*\big)^T\nabla f_i(\bar{\mathbf{x}}(k))\Big], \quad (\text{D.6})
\end{aligned}$$

where using the strong convexity of $f_i$, the last term is bounded by

$$\begin{aligned}
&-2\alpha(k)\mathbb{E}\Big[\big(\bar{\mathbf{x}}(k)-\mathbf{x}^*\big)^T\nabla f_i(\bar{\mathbf{x}}(k))\Big]\\
&\qquad\leq -2\alpha(k)\mathbb{E}\Big[f_i(\bar{\mathbf{x}}(k))-f(\mathbf{x}^*)\Big]-\mu_i\alpha(k)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big].
\end{aligned}$$

Let $\mathbf{y} \in \mathcal{X}$. The preceding equation can be further written as

$$\begin{aligned}
&-2\alpha(k)\mathbb{E}\Big[\big(\bar{\mathbf{x}}(k)-\mathbf{x}^*\big)^T\nabla f_i(\bar{\mathbf{x}}(k))\Big]\\
&\quad= -2\alpha(k)\mathbb{E}\Big[f_i(\mathbf{y})-f_i(\mathbf{x}^*)\Big] - \mu_i\alpha(k)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big]\\
&\qquad\quad -2\alpha(k)\mathbb{E}\Big[f_i(\bar{\mathbf{x}}(k))-f_i(\mathbf{y}).\Big]\\
&\quad\overset{(6.7)}{\leq} -2\alpha(k)\mathbb{E}\Big[f_i(\mathbf{y})-f_i(\mathbf{x}^*)\Big] - \mu_i\alpha(k)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big]\\
&\qquad\quad +2C_i\alpha(k)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{y}\big\|\Big]\\
&\quad\leq -2\alpha(k)\mathbb{E}\Big[f_i(\mathbf{y})-f_i(\mathbf{x}^*)\Big] - \mu_i\alpha(k)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big]\\
&\qquad\quad +\frac{1}{4Dn}\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{y}\big\|^2\Big] + 4DnC_i^2\alpha^2(k),
\end{aligned}\tag{D.7}$$

where the last inequality is due to the Cauchy-Schwarz inequality. Substituting Eq. (D.7) into Eq. (D.6) we obtain

$$\begin{aligned}
\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathbf{x}^*\big\|^2\Big] &\leq (1-\mu_i\alpha(k))\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big] + (4Dn+1)C_i^2\alpha^2(k)\\
&\quad -2\alpha(k)\mathbb{E}\Big[f_i(\mathbf{y})-f_i(\mathbf{x}^*)\Big] + \frac{\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{y}\big\|^2\Big]}{4Dn}\cdot
\end{aligned}\tag{D.8}$$

1) If $0 < \mu_i$, for all $i \in \mathcal{V}$, then by averaging over $i$ on both sides of Eq. (D.8) and using Eq. (6.10), we obtain Eq. (6.12), i.e.,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big\|\mathbf{v}_i(k)-\mathbf{x}^*\big\|^2\Big]\\
&\quad\leq \left(1-\frac{\mu\alpha(k)}{n}\right)\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{x}^*\big\|^2\Big] + (4Dn+1)C^2\alpha^2(k)\\
&\qquad\quad -\frac{2\alpha(k)}{n}\mathbb{E}\Big[F(\mathbf{y})-F(\mathbf{x}^*)\Big] + \frac{1}{4Dn}\mathbb{E}\Big[\big\|\bar{\mathbf{x}}(k)-\mathbf{y}\big\|^2\Big].
\end{aligned}$$

2) Similarly, if $\mu_i = 0$, for all $i \in \mathcal{V}$, then we obtain Eq. (6.13). □

# Appendix E

# Proofs of Main Results in Chapter 7

We provide here the proofs of main results in Sections 7.2 and 7.3. We start with some preliminaries and notation.

## E.1 Preliminaries and Notation

In the sequel, we denote by $\mathcal{X}$ the Cartesian products of $\mathcal{X}_i$, i.e.,

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n.$$

The feasible set $\mathcal{S}$ of problem $\mathsf{P}$ is given as

$$\mathcal{S} = \left\{ \mathbf{x} \in \mathcal{X} \;\middle|\; \sum_{i \in \mathcal{V}} x_i = b \right\}.$$

Let $f$ denote the sum of $f_i$, i.e.,

$$f(\mathbf{x}) = \sum_{i \in \mathcal{V}} f_i(x_i).$$

Since the constraint set $\mathcal{S}$ is compact and the function $f$ is continuous, there exists a vector $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*) \in \mathcal{S}$ which achieves the minimum of $\mathsf{P}$. However, this solution is not unique. We denote the set of solutions of $\mathsf{P}$ as $\mathcal{S}^*$. Let $\lambda^*$ be an optimizer of $\mathsf{DP}$ and let $\mathbf{x}^*$ be the corresponding optimizer of $\mathsf{P}$, such that, $(\mathbf{x}^*, \lambda^*)$ is a saddle point of $\mathcal{L}$ in Eq. (7.5), i.e.,

$$\mathcal{L}(\mathbf{x}^*, \lambda) \leq \mathcal{L}(\mathbf{x}^*, \lambda^*) \leq \mathcal{L}(\mathbf{x}, \lambda^*), \qquad \forall \mathbf{x} \in \mathcal{X},\ \lambda \in \mathbb{R}. \tag{E.1}$$

We present here the so-called distributed perturbed consensus algorithm studied in [26], a noisy version of Eq. (2.14). As will be seen, the distributed subgradient method in Eq. (2.16) is a special case of this algorithm. The distributed perturbed consensus method on a given sequence of time-varying undirected graphs $G(k) = (\mathcal{V}, \mathcal{E}(k))$, is given as

$$\lambda_i(k+1) = \sum_{j \in \mathcal{N}_i(k)} a_{ij}(k)\lambda_i(k) + \epsilon_i(k), \tag{E.2}$$

where $\epsilon_i(k)$ is some perturbation or disturbance at node $i$ and $a_{ij}(k)$ is the weight which node $i$ assigns for the message received from node $j$ at time $k$. By allowing $\epsilon_i(k)$ to take different forms, or using differing assumptions, we can study different classes of algorithms for different problems. For example, this method reduces to distributed subgradient methods for dual problem (7.6) when $\epsilon_i(k) = -\alpha(k)\partial q_i(\lambda_i(k))$, for all $i \in \mathcal{V}$ and $k \geq 0$.

We state here some important results which we will utilize in our development later. We first state a result on almost supermartingale convergence studied in [115] (see also in [116, Lemma 11, Chapter 2]), which may refer to as Robbins-Siegmund Lemma. We then present an important lemma of distributed perturbed consensus methods, an extension of Lemma 2.

**Lemma 21** ( [115]). *Let $\{y(k)\}$, $\{z(k)\}$, $\{w(k)\}$, and $\{\beta(k)\}$ be non-negative sequences of random variables. Suppose that these sequences satisfy*

$$\mathbb{E}\Big[y(k+1)|\mathcal{F}_k\Big] \leq (1+\beta(k))y(k) - z(k) + w(k) \tag{E.3}$$

$$\sum_{k=0}^{\infty} \beta(k) < \infty \;\; a.s, \qquad \sum_{k=0}^{\infty} w(k) < \infty \;\; a.s, \tag{E.4}$$

*where $\mathcal{F}_k = \{y(0), y(1), \ldots, y(k)\}$, the history of $y$ up to time $k$. Then the sequence $\{y(k)\}$ converges a.s., and $\sum_{k=0}^{\infty} z(k) < \infty$ a.s.*

**Lemma 22** ( [26]). *Suppose that Assumptions 2 and 4 hold. Let the sequence $\{\lambda_i(k)\}$, for all $i \in \mathcal{V}$, be generated by Eq.* (E.2) *with an arbitrary initial condition $\lambda_i(0) \in \mathbb{R}$, for all $i \in \mathcal{V}$. Then we have*

1. *For all $i \in \mathcal{V}$ and $k \geq 0$*

$$\big|\lambda_i(k) - \bar{\lambda}(k)\big| \leq \delta^k \big\|\boldsymbol{\lambda}(0)\big\| + \sum_{t=1}^{k} \delta^{k-t} \big\|\boldsymbol{\epsilon}(t)\big\|, \tag{E.5}$$

*where $\delta \leq \min\{(1-\frac{1}{4n^3})^{1/B}, \max_{k\geq 0} \sigma_2(\mathbf{A}(k))\}$.*

2. *Further if* $\lim_{k\to\infty} \epsilon_i(k) = 0$, *for all* $i \in \mathcal{V}$, *then we have*

$$\lim_{k\to\infty} \left|\lambda_i(k) - \bar{\lambda}(k)\right| = 0, \quad \forall i \in \mathcal{V}. \tag{E.6}$$

3. *If we are further given a non-increasing positive scalar sequence* $\{\alpha(k)\}$ *such that* $\sum_{k=0}^{\infty} \alpha(k)|\epsilon_i(k)| < \infty$, *for all* $i \in \mathcal{V}$, *then we obtain*

$$\sum_{k=0}^{\infty} \alpha(k)\left|\lambda_i(k) - \bar{\lambda}(k)\right| < \infty, \qquad \forall i \in \mathcal{V}. \tag{E.7}$$

## E.2 Proofs of Results in Section 7.2

We now present the proof of part (b) in Theorem 14; recall that part (a) is a consequence of Theorem 5. Let $C = \sum_{i\in\mathcal{V}} C_i$ where $C_i$ is given in Eq. (7.10).

*Proof of part (b) Theorem 14.* Let $(\mathbf{x}^*, \lambda^*)$ is a saddle point of $\mathcal{L}$, i.e., $(\mathbf{x}^*, \lambda^*)$ satisfies Eq. (E.1). Recall from Eq. (7.8) that

$$\sum_{i\in\mathcal{V}} \mathcal{L}_i(x_i, v_i) = \sum_{i\in\mathcal{V}} f_i(x_i) + v_i(x_i - b_i).$$

To show our main result, we will show the following relation,

$$\begin{aligned} 0 &\leq \sum_{i\in\mathcal{V}} \mathcal{L}_i(x_i^*, v_i(k+1)) - \mathcal{L}_i(x_i(k+1), v_i(k+1)) \\ &\leq C\left\|\mathbf{v}(k+1) - \lambda^*\mathbf{1}\right\| + \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i). \end{aligned} \tag{E.8}$$

We note that by part (a) $\lim_{k\to\infty} \lambda_i(k) = \lambda^*$ for all $i \in \mathcal{V}$. This implies that $\lim_{k\to\infty} v_i(k) = \lambda^*$ since

$$v_i(k) = \sum_{j\in\mathcal{N}_i} a_{ij}(k)\lambda_j(k-1)$$

and $\mathbf{A}(k)$ is a doubly stochastic matrix. In addition, since $\mathbf{x}^* \in \mathcal{S}$ we have

$$\lim_{k\to\infty} \left\|\mathbf{v}(k+1) - \lambda^*\mathbf{1}\right\| = 0 \qquad \text{and} \qquad \lim_{k\to\infty} \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i) = 0.$$

Thus, we obtain

$$\begin{aligned}0 &\leq \lim_{k\to\infty}\sum_{i\in\mathcal{V}}\mathcal{L}_i(x_i^*,v_i(k+1))-\mathcal{L}_i(x_i(k+1),v_i(k+1))\\&=\lim_{k\to\infty}\sum_{i\in\mathcal{V}}\mathcal{L}_i(x_i^*,\lambda^*)-\mathcal{L}_i(x_i(k+1),\lambda^*)=0.\end{aligned}$$

By Eqs. (7.5) and (7.8) the preceding equation implies that

$$\begin{aligned}0 &\leq \mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*)-\lim_{k\to\infty}\mathcal{L}(\mathbf{x}(k+1),\boldsymbol{\lambda}(k))\\&=f(\mathbf{x}^*)-\lim_{k\to\infty}\mathcal{L}(\mathbf{x}(k+1),\boldsymbol{\lambda}(k))=0,\end{aligned}\tag{E.9}$$

where we use the fact that $\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*)=f(\mathbf{x}^*)$ and

$$\lim_{k\to\infty}\boldsymbol{\lambda}(k)=\lim_{k\to\infty}\mathbf{v}(k)=\boldsymbol{\lambda}^*.$$

We now proceed to show Eq. (E.8). Since $x_i(k)$ satisfies Eq. (7.3) and by the definition of $\mathcal{L}_i$ we have for any $k\geq 0$

$$0\leq\mathcal{L}_i(x_i^*,v_i(k+1))-\mathcal{L}_i(x_i(k+1),v_i(k+1)),\quad\forall i\in\mathcal{V},$$

which when summing over $i\in\mathcal{V}$ implies that

$$\begin{aligned}0 &\leq\sum_{i\in\mathcal{V}}\mathcal{L}_i(x_i^*,v_i(k+1))-\mathcal{L}_i(x_i(k+1),v_i(k+1))\\&=\sum_{i\in\mathcal{V}}f_i(x_i^*)+v_i(k+1)(x_i^*-b_i)\\&\quad-\sum_{i\in\mathcal{V}}f_i(x_i(k+1))+v_i(k+1)(x_i(k+1)-b_i).\end{aligned}\tag{E.10}$$

By Assumption 11 we have the strong duality holds, i.e.,

$$\sum_{i\in\mathcal{V}}f_i(x_i^*)=-\sum_{i\in\mathcal{V}}q_i(\lambda^*).$$

Moreover by Eqs. (7.6) and (7.3) in Algorithm 4 we have

$$q_i(v_i(k+1))=-f_i(x_i(k+1))-v_i(k+1)(x_i(k+1)-b_i).$$

Substituting the previous two preceding relations into Eq. (E.10) we obtain

$$\begin{aligned}
0 &\leq \sum_{i\in\mathcal{V}} \mathcal{L}_i(x_i^*, v_i(k+1)) - \mathcal{L}_i(x_i(k+1), v_i(k+1))\\
&= \sum_{i\in\mathcal{V}} \Big(q_i(v_i(k+1)) - q_i(\lambda^*) + v_i(k+1)(x_i^* - b_i)\Big)\\
&\overset{(7.10)}{\leq} \sum_{i\in\mathcal{V}} \Big( - \partial q_i(v_i(k+1))(\lambda^* - v_i(k+1)) + v_i(k+1)(x_i^* - b_i)\Big)\\
&\leq \sum_{i\in\mathcal{V}} \Big(C_i\big|\lambda^* - v_i(k+1)\big| + v_i(k+1)(x_i^* - b_i)\Big)\\
&\leq C\big\|\lambda^*\mathbf{1} - \mathbf{v}(k+1)\big\| + \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i),
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality. This concludes our proof. □

## E.3 Proofs of Results in Section 7.3

In this section, we provide the proofs for main results on the convergence of the distributed randomized Lagrangian method presented in Section 7.3. The distributed noisy sub-gradient method has also been studied in [26] where the authors consider the case of strongly convex objective functions. We analyze here the convergence of such methods when the objective function is convex. To do this we need to introduce more notation. Let $\mathcal{L}_i^s : \mathcal{X}_i \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the local stochastic Lagrangian function at node $i$ defined as,

$$\mathcal{L}_i^s(x_i, v_i, \ell_i) = f_i(x_i) + v_i(x_i - \ell_i). \tag{E.11}$$

We define $\mathcal{F}_k$ to be all the information generated by the randomized distributed primal-dual method up to time $k$, i.e., all the $x_i(k)$, $v_i(k)$, $\lambda_i(k)$, $g_i(v_i(k))$, and so forth for $k \geq 0$. We start with the analysis of Theorem 15.

*Proof of Theorem 15.* Let $D = \sum_{i\in\mathcal{V}} D_i$ where $D_i$ is given in Eq. (7.17). Recall that the dual function in DP is given as

$$q(\lambda) = \sum_{i\in\mathcal{V}} q_i(\lambda).$$

1. *Proof of part (a)*: Let $\lambda^*$ be a dual solution of dual problem (7.6). By Eqs. (7.12) and (7.14) in Algorithm 5, and Eq. (7.15) we have

$$\begin{aligned}
&\sum_{i\in\mathcal{V}}\Big(\lambda_i(k+1)-\lambda^*\Big)^2\\
&\qquad=\sum_{i\in\mathcal{V}}\Big(v_i(k+1)-\alpha(k)g_i(v_i(k+1))-\lambda^*\Big)^2\\
&\qquad=\sum_{i\in\mathcal{V}}\Big(v_i(k+1)-\lambda^*\Big)^2+\sum_{i\in\mathcal{V}}\alpha^2(k)\Big[g_i(v_i(k+1))\Big]^2\\
&\qquad\qquad-2\alpha(k)\sum_{i\in\mathcal{V}}g_i(v_i(k+1))\Big(v_i(k+1)-\lambda^*\Big). \qquad\text{(E.12)}
\end{aligned}$$

Recall from Eqs. (7.7) and (7.15) that

$$\partial q_i(v_i(k))=x_i(k)-b_i \qquad \text{and} \qquad g_i(v_i(k))=x_i(k)-\ell_i(k-1).$$

Thus, by Assumption 12 and Eq. (7.11) we have

$$\mathbb{E}\Big[g_i(v_i(k+1))\,|\,\mathcal{F}_k\Big]=\partial q_i(v_i(k+1)).$$

Taking the conditional expectation of above with respect to $\mathcal{F}_k$ and using Eq. (7.17) we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(k+1)-\lambda^*\mathbf{1}\big\|^2\,|\,\mathcal{F}_k\Big]\\
&\qquad=\big\|\mathbf{v}(k+1)-\lambda^*\mathbf{1}\big\|^2+\alpha^2(k)\mathbb{E}\left[\sum_{i\in\mathcal{V}}\Big(g_i(v_i(k+1))\Big)^2\,|\,\mathcal{F}_k\right]\\
&\qquad\qquad-2\alpha(k)\sum_{i\in\mathcal{V}}\partial q_i(v_i(k+1))\Big(v_i(k+1)-\lambda^*\Big)\\
&\qquad\overset{(7.17)}{\leq}\big\|\mathbf{v}(k+1)-\lambda^*\mathbf{1}\big\|^2+\alpha^2(k)\sum_{i\in\mathcal{V}}D_i^2\\
&\qquad\qquad-2\alpha(k)\sum_{i\in\mathcal{V}}\partial q_i(v_i(k+1))\Big(v_i(k+1)-\lambda^*\Big)\\
&\qquad\leq\big\|\mathbf{v}(k+1)-\lambda^*\mathbf{1}\big\|^2+D\alpha^2(k)\\
&\qquad\qquad+2\alpha(k)\sum_{i\in\mathcal{V}}\Big(q_i(\lambda^*)-q_i(v_i(k+1))\Big), \qquad\text{(E.13)}
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality and the convexity of $q_i$, for all $i \in \mathcal{V}$. Consider the last term on the right-hand side of Eq. (E.13)

$$\begin{aligned}
&\sum_{i\in\mathcal{V}} \Big(q_i(\lambda^*) - q_i(v_i(k+1))\Big) \\
&\qquad = \sum_{i\in\mathcal{V}} \Big(q_i(\lambda^*) - q_i(\bar{\lambda}(k)) + q_i(\bar{\lambda}(k)) - q_i(v_i(k+1))\Big) \\
&\qquad \le \sum_{i\in\mathcal{V}} \Big(q_i(\lambda^*) - q_i\big(\bar{\lambda}(k)\big)\Big) + \sum_{i\in\mathcal{V}} \Big(D_i\big|\bar{\lambda}(k) - v_i(k+1)\big|\Big) \\
&\qquad \le q^* - q\big(\bar{\lambda}(k)\big) + D\|\bar{\lambda}(k)\mathbf{1} - \mathbf{v}(k+1)\|.
\end{aligned}$$

Substituting the preceding relation into (E.13) yields

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(k+1) - \lambda^*\mathbf{1}\big\|^2 \,|\, \mathcal{F}_k\Big] \\
&\qquad \le \big\|\mathbf{v}(k+1) - \lambda^*\mathbf{1}\big\|^2 + D\alpha^2(k) \\
&\qquad\qquad + 2\alpha(k)\Big(q^* - q\big(\bar{\lambda}(k)\big) + D\|\bar{\lambda}(k)\mathbf{1} - \mathbf{v}(k+1)\|\Big) \\
&\qquad \le \big\|\boldsymbol{\lambda}(k) - \lambda^*\mathbf{1}\big\|^2 + D\alpha^2(k) \\
&\qquad\qquad + 2\alpha(k)\Big(q^* - q\big(\bar{\lambda}(k)\big) + D\|\bar{\lambda}(k)\mathbf{1} - \boldsymbol{\lambda}(k)\|\Big),
\end{aligned} \tag{E.14}$$

where the last inequality is due to

$$\begin{aligned}
\big\|\mathbf{v}(k+1) - \bar{\lambda}(k)\mathbf{1}\|_1 &\le \|\boldsymbol{\lambda}(k) - \bar{\lambda}(k)\mathbf{1}\| \\
\big\|\mathbf{v}(k+1) - \lambda^*\mathbf{1}\big\|^2 &\le \|\boldsymbol{\lambda}(k) - \lambda^*\mathbf{1}\big\|^2.
\end{aligned}$$

Recall that Eq. (7.14) in Algorithm 5 is a special case of the perturbed consensus protocol in Eq. (E.2) where

$$\epsilon_i(k) = -\alpha(k)g_i(v_i(k+1)) = -\alpha(k)(\partial q_i(v_i(k+1)) - \eta_i(k)).$$

Since $\alpha(k)$ satisfies Eq. (7.16) and by Eq. (7.17) we obtain

$$\begin{aligned}
\sum_{k=0}^{\infty} \alpha(k)\|\boldsymbol{\epsilon}(k)\| &= \sum_{k=0}^{\infty} \alpha^2(k)\|\nabla q(\mathbf{v}(k+1)) + \boldsymbol{\eta}(k)\| \\
&\le nD\sum_{k=0}^{\infty} \alpha^2(k) < \infty \text{ a.s.},
\end{aligned}$$

which satisfies Eq. (E.5). Thus we obtain

$$\sum_{k=0}^{\infty} \alpha(k) \big| \lambda_i(k) - \bar{\lambda}(k) \big| < \infty \text{ a.s.,} \quad \forall i \in \mathcal{V}.$$

This implies

$$nD^2 \sum_{k=0}^{\infty} \alpha^2(k) + 2C \sum_{k=0}^{\infty} \alpha(k) \| \boldsymbol{\lambda}(k) - \bar{\lambda}(k)\mathbf{1} \| < \infty \text{ a.s.}$$

Thus, since Eq. (E.14) satisfies all conditions in Lemma 21 we obtain

$$\Big\{ \| \boldsymbol{\lambda}(k) - \lambda^* \mathbf{1} \| \Big\} \text{ converges a.s. for each } \lambda^* \tag{E.15}$$

$$\sum_{k=0}^{\infty} \alpha(k) \Big( q\big(\bar{\lambda}(k)\big) - q^* \Big) < \infty \text{ a.s.,} \tag{E.16}$$

which since $\sum_{k=0}^{\infty} \alpha(k) = \infty$ gives

$$\liminf_{k \to \infty} \; q\big(\bar{\lambda}(k)\big) = q^* \text{ a.s.}$$

Eq. (E.15) implies that the sequence $\{\bar{\lambda}(k)\}$ is bounded, so there exists a bounded subsequence $\{\bar{\lambda}(k_\ell)\}$ of $\{\bar{\lambda}(k)\}$ such that

$$\lim_{k_\ell \to \infty} q\big(\bar{\lambda}(k_\ell)\big) = \liminf_{k \to \infty} \; q\big(\bar{\lambda}(k)\big) = q^* \text{ a.s.} \tag{E.17}$$

This bounded subsequence $\{\bar{\lambda}(k_\ell)\}$ has a convergent subsequence. By Eq. (E.17) and the continuity of $f$ this subsequence converges to a point in $\mathcal{S}^*$. Call this point $\tilde{\lambda}$, a solution of Eq. (7.1). By letting $\lambda^* = \tilde{\lambda}$ in Eq. (E.15) we obtain that

$$\lim_{k \to \infty} \bar{\lambda}(k) = \tilde{\lambda} \text{ a.s.}$$

Finally, since $\alpha(k)$ satisfies Eq. (7.16), $\lim_{k \to \infty} \alpha(k) = 0$. Thus, by Eq. (E.6) we have

$$\lim_{k \to \infty} |\lambda_i(k) - \bar{\lambda}(k)| = 0 \text{ a.s.,} \qquad \forall i \in \mathcal{V},$$

which further implies that

$$\lim_{k \to \infty} \lambda_i(k) = \tilde{\lambda} \text{ a.s.,} \qquad \forall i \in \mathcal{V}.$$

2. *Proof of part (b)*

Recall that $(\mathbf{x}^*, \lambda^*)$ is a saddle point of the Lagrangian in Eq. (7.5). Since $x_i(k)$ satisfies Eq. 7.13 and by Eq. (E.11) we have for all $i \in \mathcal{V}$ and $k \geq 0$

$$0 \leq \mathcal{L}_i^s(x_i^*, v_i(k+1), \ell_i(k)) - \mathcal{L}_i^s(x_i(k+1), v_i(k+1), \ell_i(k)),$$

which when summing over $i \in \mathcal{V}$ implies

$$\begin{aligned} 0 \leq & \sum_{i\in\mathcal{V}} \mathcal{L}_i^s(x_i^*, v_i(k+1), \ell_i(k)) - \mathcal{L}_i^s(x_i(k+1), v_i(k+1), \ell_i(k)) \\ & = \sum_{i\in\mathcal{V}} f_i(x_i^*) + v_i(k+1)(x_i^* - \ell_i(k)) \\ & \quad - \sum_{i\in\mathcal{V}} f_i(x_i(k+1)) + v_i(k+1)(x_i(k+1) - \ell_i(k)). \end{aligned} \tag{E.18}$$

By Assumption 11 the strong duality holds, i.e.,

$$\sum_{i\in\mathcal{V}} f_i(x_i^*) = -\sum_{i\in\mathcal{V}} q_i(\lambda^*). \tag{E.19}$$

Recall from Eq. (7.6) that

$$q_i(v_i(k)) = -f_i(x_i(k)) - v_i(k)(x_i(k) - b_i). \tag{E.20}$$

Taking the conditional expectation of Eq. (E.18) with respect to $\mathcal{F}_k$, and using Eqs. (E.19) and (E.19) give

$$\begin{aligned} 0 \leq & \sum_{i\in\mathcal{V}} \mathbb{E}\Big[\mathcal{L}_i^s(x_i^*, v_i(k+1), \ell_i(k)) \,|\, \mathcal{F}_k\Big] \\ & \quad - \sum_{i\in\mathcal{V}} \mathbb{E}\Big[\mathcal{L}_i^s(x_i(k+1), v_i(k+1), \ell_i(k)) \,|\, \mathcal{F}_k\Big] \\ & = \sum_{i\in\mathcal{V}} \mathbb{E}\Big[q_i(v_i(k+1)) - q_i(\lambda^*) \,|\, \mathcal{F}_k\Big] \\ & \quad + \sum_{i\in\mathcal{V}} \mathbb{E}\Big[v_i(k+1)(x_i^* - b_i) \,|\, \mathcal{F}_k\Big] \\ & \leq \sum_{i\in\mathcal{V}} \mathbb{E}\Big[-\partial q_i(v_i(k+1))(\lambda^* - v_i(k+1)) \,|\, \mathcal{F}_k\Big] \\ & \quad + \sum_{i\in\mathcal{V}} \mathbb{E}\Big[v_i(k+1)(x_i^* - b_i) \,|\, \mathcal{F}_k\Big], \end{aligned}$$

which by using the Cauchy-Schwarz inequality and Eq. (7.17) gives

$$\begin{aligned}0 &\leq \sum_{i\in\mathcal{V}}\mathbb{E}\Big[\mathcal{L}_i^s(x_i^*, v_i(k+1), \ell_i(k))\,|\,\mathcal{F}_k\Big]\\ &\qquad - \sum_{i\in\mathcal{V}}\mathbb{E}\Big[\mathcal{L}_i^s(x_i(k+1), v_i(k+1), \ell_i(k))\,|\,\mathcal{F}_k\Big]\\ &\leq \sum_{i\in\mathcal{V}} C_i\big|\lambda^* - v_i(k+1)\big| + \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i)\\ &\leq C\|\mathbf{v}(k+1) - \lambda^*\mathbf{1}\| + \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i)\\ &\leq C\|\boldsymbol{\lambda}(k) - \lambda^*\mathbf{1}\| + \sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i).\end{aligned}$$

Taking the limit as $k\to\infty$ of the preceding relation gives

$$\begin{aligned}0 &\leq \lim_{k\to\infty}\sum_{i\in\mathcal{V}}\mathbb{E}\Big[\mathcal{L}_i^s(x_i^*, v_i(k+1), \ell_i(k))\,|\,\mathcal{F}_k\Big]\\ &\qquad - \lim_{k\to\infty}\sum_{i\in\mathcal{V}}\mathbb{E}\Big[\mathcal{L}_i^s(x_i(k+1), v_i(k+1), \ell_i(k))\,|\,\mathcal{F}_k\Big]\\ &\leq C\lim_{k\to\infty}\|\boldsymbol{\lambda}(k) - \lambda^*\mathbf{1}\| + \lim_{k\to\infty}\sum_{i\in\mathcal{V}} v_i(k+1)(x_i^* - b_i) \;=\; 0 \;\text{ a.s.,}\end{aligned} \tag{E.21}$$

where we use the conditions

$$\lim_{k\to\infty}\lambda_i(k) = \lambda^* \qquad \text{and} \qquad \sum_{i\in\mathcal{V}} x_i^* - b_i = 0.$$

By Assumption 12 and recall the definition of $\mathcal{L}_i$ in Eq. (7.8) we have

$$\begin{aligned}\mathbb{E}\Big[\mathcal{L}_i^s(x_i^*, v_i(k), \ell_i(k))\Big] &= \mathbb{E}\Big[\mathcal{L}_i(x_i^*, v_i(k))\Big]\\ \mathbb{E}\Big[\mathcal{L}_i^s(x_i(k), v_i(k), \ell_i(k))\Big] &= \mathbb{E}\Big[\mathcal{L}_i(x_i(k), v_i(k))\Big].\end{aligned}$$

Thus taking the expectation of Eq. (E.21) and using the convergence of $\lambda_i(k)$, for all $i\in\mathcal{V}$, give

$$\lim_{k\to\infty}\mathbb{E}\left[\sum_{i\in\mathcal{V}}\mathcal{L}_i(x_i(k+1), \lambda_i(k))\right] = f^*.$$

□

Finally, we present the proof of Theorem 16.

*Proof of Theorem 16.* Recall that Eq. (7.14) in Algorithm 5 is a special case of a perturbed averaging protocol in Eq. (E.2), where

$$\epsilon_i(k) = -\alpha(k) g_i(v_i(k+1)) = -\alpha(k)\Big(\partial q_i(v_i(k+1)) - \eta_i(k)\Big).$$

By Eqs. (E.5) and (7.17), for all $i \in \mathcal{V}$ and $K \geq 1$, we have

$$\sum_{k=1}^{K} \alpha(k) \big|\lambda_i(k) - \bar{\lambda}(k)\big| \leq \|\boldsymbol{\lambda}(0)\| \sum_{k=1}^{K} \delta^k \alpha(k) + D \sum_{k=1}^{K} \alpha(k) \sum_{t=1}^{k} \delta^{k-t} \alpha(t) \quad a.s. \tag{E.22}$$

Since $\alpha(k) = 1/\sqrt{k} \leq 1$ and $\delta < 1$, Eq. (E.22) is equivalent to

$$\begin{aligned}
\sum_{k=1}^{K} \alpha(k) |\lambda_i(k) - \bar{\lambda}(k)| &\leq \frac{\delta}{1-\delta} \|\boldsymbol{\lambda}(0)\| + D_i \sum_{k=1}^{K} \sum_{t=1}^{k} \frac{\delta^{k-t}}{t} \\
&= \frac{\delta}{1-\delta} \|\boldsymbol{\lambda}(0)\| + D_i \sum_{t=1}^{K} \frac{1}{t} \sum_{\ell=0}^{K-t} \delta^{\ell} \\
&\leq \frac{\delta}{1-\delta} \|\boldsymbol{\lambda}(0)\| + \frac{D_i}{1-\delta} \sum_{t=1}^{K} \frac{1}{t} \\
&\leq \frac{\delta}{1-\delta} \|\boldsymbol{\lambda}(0)\| + \frac{D_i(1+\ln(K))}{1-\delta},
\end{aligned} \tag{E.23}$$

where the last inequality is due to

$$\sum_{t=1}^{K} \frac{1}{t} = 1 + \sum_{t=2}^{K} \frac{1}{t} \leq 1 + \int_1^K \frac{du}{u} = 1 + \ln(K). \tag{E.24}$$

Since $\alpha(0) = 1$ we have

$$\alpha(0) \|\boldsymbol{\lambda}(0) - \bar{\lambda}(0)\mathbf{1}\| \leq \|\boldsymbol{\lambda}(0)\|. \tag{E.25}$$

Adding Eq. (E.25) to both sides of Eq. (E.22) and using $D = \sum_{i \in \mathcal{V}} D_i$ give

$$\sum_{k=0}^{K} \alpha(k) \big\|\boldsymbol{\lambda}(k) - \bar{\lambda}(k)\mathbf{1}\big\| \leq \frac{1}{1-\delta} \|\boldsymbol{\lambda}(0)\| + \frac{D(1+\ln(K))}{1-\delta} \quad a.s. \tag{E.26}$$

Recall from Eq. (E.14) that

$$\mathbb{E}\Big[\|\boldsymbol{\lambda}(k+1)-\lambda^*\mathbf{1}\big\|^2\,|\,\mathcal{F}_k\Big] \leq \big\|\boldsymbol{\lambda}(k)-\lambda^*\mathbf{1}\big\|^2 - 2\alpha(k)(q(\bar{\lambda}(k))-q^*) \\ + D^2\alpha^2(k) + 2D\alpha(k)\|\boldsymbol{\lambda}(k)-\bar{\lambda}(k)\mathbf{1}\|. \quad \text{(E.27)}$$

Summing both sides of Eq. (E.27) over $k = 0, \ldots, K$ for some $K \geq 0$, and using Eqs. (E.23) and (E.24), give

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(K+1)-\lambda^*\mathbf{1}\big\|^2\,|\,\mathcal{F}_k\Big] \\
&\quad\leq \big\|\boldsymbol{\lambda}(0)-\lambda^*\mathbf{1}\big\|^2 - 2\sum_{k=0}^{K}\alpha(k)(q(\bar{\lambda}(k))-q^*) \\
&\qquad + D^2\sum_{k=0}^{K}\alpha^2(k) + 2D\sum_{k=0}^{K}\alpha(k)\big\|\boldsymbol{\lambda}(k)-\bar{\lambda}(k)\mathbf{1}\big\|, \\
&\quad\underset{\text{(E.24)}}{\overset{\text{(E.23)}}{\leq}} \big\|\boldsymbol{\lambda}(0)-\lambda^*\mathbf{1}\big\|^2 - 2\sum_{k=0}^{K}\alpha(k)(q(\bar{\lambda}(k))-q^*) \\
&\qquad + \frac{2D\big\|\boldsymbol{\lambda}(0)\big\|}{1-\delta} + \frac{2D^2(2+\ln(K))}{1-\delta},
\end{aligned} \quad \text{(E.28)}$$

which when taking the expectation of both sides implies

$$\mathbb{E}\Big[\|\boldsymbol{\lambda}(K+1)-\lambda^*\mathbf{1}\big\|^2\Big] \leq \mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)-\lambda^*\mathbf{1}\big\|^2\Big] - 2\sum_{k=0}^{K}\alpha(k)\mathbb{E}\Big[q(\bar{\lambda}(k))-q^*\Big] \\ + \frac{2D\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)\big\|\Big]}{1-\delta} + \frac{2D^2(2+ln(K))}{1-\delta}\cdot \quad \text{(E.29)}$$

Dividing both sides of the preceding relation by $2\sum_{k=0}^{K}\alpha(k+1)$, using $\delta \in (0,1)$, and rearranging the terms give

$$\begin{aligned}
&\frac{\sum_{k=0}^{K}\alpha(k)\mathbb{E}\Big[q(\bar{\lambda}(k))\Big]}{\sum_{k=0}^{K}\alpha(k)} - q^* \\
&\quad\leq \frac{\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)-\boldsymbol{\lambda}^*\big\|^2\Big]}{2\sum_{k=0}^{K}\alpha(k)} + \frac{D\delta\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)\big\|\Big]}{(1-\delta)\sum_{k=0}^{K}\alpha(k)} + \frac{D^2(2+ln(K))}{(1-\delta)\sum_{k=0}^{K}\alpha(k)} \\
&\quad\leq \frac{\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)-\boldsymbol{\lambda}^*\big\|^2\Big]}{2\sum_{k=0}^{K}\alpha(k)} + \frac{D\mathbb{E}\Big[\big\|\boldsymbol{\lambda}(0)\big\|\Big]}{(1-\delta)\sum_{k=0}^{K}\alpha(k)} + \frac{D^2(2+ln(K))}{(1-\delta)\sum_{k=0}^{K}\alpha(k)}\cdot
\end{aligned} \quad \text{(E.30)}$$

Since $\alpha(k) = 1/\sqrt{k}$ and $\alpha(1) = 1$ we have

$$\sum_{k=0}^{K} \alpha(k) = \sum_{k=0}^{K} \frac{1}{\sqrt{k}} \geq \int_{0}^{K+1} \frac{du}{\sqrt{u}} = 2\sqrt{K+1}. \tag{E.31}$$

Using Eq. (E.31) and the Jensen inequality into Eq. (E.30) gives

$$\begin{aligned}
&\mathbb{E}\left[q\left(\frac{\sum_{k=0}^{K}\alpha(k)\bar{\boldsymbol{\lambda}}(k)}{\sum_{k=0}^{K}\alpha(k)}\right)\right] - q^* \\
&\quad \leq \frac{\mathbb{E}\Big[\|\boldsymbol{\lambda}(0) - \lambda^*\mathbf{1}\|^2\Big]}{2\sqrt{K+1}} + \frac{D\delta\mathbb{E}\Big[\|\boldsymbol{\lambda}(0)\|\Big]}{2(1-\delta)\sqrt{K+1}} + \frac{D^2(2+\ln(K))}{2(1-\delta)\sqrt{K+1}} \cdot
\end{aligned} \tag{E.32}$$

Fixed some $i \in \mathcal{V}$. Using Eqs. (E.32) and (7.18) gives

$$\begin{aligned}
&\mathbb{E}\left[q(y_i(K+1)) - q\left(\frac{\sum_{k=0}^{K}\alpha(k)\bar{\lambda}(k)}{\sum_{k=0}^{K}\alpha(k)}\right) \mid \mathcal{F}_k\right] \\
&\quad = \mathbb{E}\left[q\left(\frac{\sum_{k=0}^{K}\alpha(k)\lambda_i(k)}{\sum_{k=0}^{K}\alpha(k)}\right) - q\left(\frac{\sum_{k=0}^{K}\alpha(k)\bar{\lambda}(k)}{\sum_{k=0}^{K}\alpha(k)}\right)\right] \\
&\quad \leq D\mathbb{E}\left[\left|\frac{\sum_{k=0}^{K}\alpha(k)\lambda_i(k)}{\sum_{k=0}^{K}\alpha(k)} - \frac{\sum_{k=0}^{K}\alpha(k)\bar{\lambda}(k)}{\sum_{k=0}^{K}\alpha(k)}\right|\right] \\
&\quad \leq \frac{D}{\sum_{k=0}^{K}\alpha(k)} \sum_{k=0}^{K} \alpha(k)|\lambda_i(k) - \bar{\lambda}(k)| \\
&\quad \leq \frac{D}{2\sqrt{K+1}} \sum_{k=0}^{K} \alpha(k)|\lambda_i(k) - \bar{\lambda}(k)| \\
&\quad \overset{\text{(E.23)}}{\leq} \frac{D\|\boldsymbol{\lambda}(0)\|}{2(1-\delta)\sqrt{K+1}} + \frac{D^2(1+\ln(K))}{2(1-\delta)\sqrt{K+1}}.
\end{aligned} \tag{E.33}$$

Thus, taking the expectation of Eq. (E.33) and then adding the result to Eq. (E.32) give Eq. (7.19). This completes our proof of Theorem 16. □

# References

[1] T. Hastie, T. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. New York: Springe-Verlag, 2009.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. Cambridge University Press, 2014.

[3] G. Meteos, J. Bazerque, and G. Giannakis, "Distributed sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, pp. 5262–5276, 2010.

[4] R. Srikant, *The Mathematics of Internet Congestion Control.* Birkhauser, 2004.

[5] M. White and R. Pike, "ISO New England and New York ISO inter-regional interchange scheduling: analysis and options," Jan. 2011.

[6] B. H. Kim and R. Baldick, "Coarse-grained distributed optimal power flow," *IEEE Transactions on Power Systems*, vol. 12, no. 2, pp. 932–939, May 1997.

[7] A. J. Conejo and J. A. Aguado, "Multi-area coordinated decentralized dc optimal power flow," *IEEE Transactions on Power Systems*, vol. 13, no. 4, pp. 1272–1278, Nov 1998.

[8] R. Baldick, B. H. Kim, C. Chase, and Y. Luo, "A fast distributed implementation of optimal power flow," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 858–864, Aug 1999.

[9] X. Wang, Y. H. Song, and Q. Lu, "Lagrangian decomposition approach to active power congestion management across interconnected regions," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 148, no. 5, pp. 497–503, Sep 2001.

[10] X. Lai, L. Xie, Q. Xia, H. Zhong, and C. Kang, "Decentralized multi-area economic dispatch via dynamic multiplier-based Lagrangian relaxation," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3225–3233, Nov 2015.

[11] T. T. Doan, S. Bose, and C. L. Beck, "Distributed Lagrangian method for tie-line scheduling in power grids under uncertainty," *SIGMETRICS Perform. Eval. Rev.*, vol. 45, no. 2, Oct. 2017.

[12] J. Cortes, S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Transactions on Automatic Control*, vol. 20, no. 2, pp. 243 – 255, 2004.

[13] P. Sharma, S. M. Salapaka, and C. L. Beck, "Entropy-based framework for dynamic coverage and clustering problems," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 135–150, Jan 2012.

[14] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[15] J. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, M.I.T., Nov. 1984.

[16] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods.* Prentice-Hall, 1989.

[17] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[18] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[19] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[20] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, no. 99, 2017.

[21] A. Nedíc, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[22] S. Ram, A. Nedić, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2009.

[23] B. Touri and B. Gharesifard, "Continuous-time distributed convex optimization on time-varying directed networks," in *IEEE 54th Annual Conference on Decision and Control (CDC)*, Japan, Dec 2015.

[24] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.

[25] K. Tsianos and M. Rabbat, "Distributed strongly convex optimization," in *Proc. of Allerton Conference on Communication, Control, and Computing*, 2012.

[26] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936 – 3947, 2016.

[27] A. Olshevsky, "Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control," arXiv preprint: https://arxiv.org/pdf/1411.4186v6.pdf, 2016.

[28] A. Nedíc, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," Available at: https://arxiv.org/abs/1709.08765.

[29] A. Nedić, D. Bertsekas, and V. Borkar, "Distributed asynchronous incremental subgradient methods," ser. Studies in Computational Mathematics. Elsevier, 2001, vol. 8, pp. 381 – 407.

[30] F. Niu, B. Recht, C. Ré, and S. Wright, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, ser. NIPS'11, 2011, pp. 693–701.

[31] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 873–881.

[32] S. Sra, A. Yu, M. Li, and A. Smola, "Adadelay: delay adaptive distributed stochastic optimization," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 51, Cadiz, Spain, 09–11 May 2016, pp. 957–965.

[33] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng, "Large scale distributed deep networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12, 2012.

[34] O. Shamir and N. Srebro, "Distributed stochastic optimization and learning," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014, pp. 850–857.

[35] S. Zhang, A. Choromanska, and Y. LeCun, "Deep learning with elastic averaging SGD," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15, 2015.

[36] M. A. Zinkevich, M. Weimer, A. Smola, and L. Li, "Parallelized stochastic gradient descent," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'10, 2010, pp. 2595–2603.

[37] C. Godsil and G. Royle, *Algebraic Graph Theory*, ser. Graduate Texts in Mathematics. New York: Springe-Verlag, 2001, vol. 207.

[38] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge, U.K.: Cambridge Univ. Press, 1985.

[39] M. Fiedler, "Algebraic connectivity of graphs," *Czech. Math. J.*, vol. 23, no. 98, pp. 298–305, 1973.

[40] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization.* Cambridge, MA: Athena Scientific, 2004.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization.* New York, NY: Cambridge University Press, 2004.

[42] R. T. Rockafellar, *Convex Analysis*, 1970.

[43] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, pp. 107–194, 2012.

[44] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, Sept 2004.

[45] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effect," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[46] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course.* Norwell, MA: Kluwer Academic Publishers, 2004.

[47] Y. E. D.P. Palomar, *Convex Optimization in Signal Processing and Communications*, 1st ed. Cambridge University Press, Dec. 2009.

[48] K. Tsianos, S. Lawlor, and M. Rabbat, "Distributed dual averaging for convex optimization under communication delays," in *Proc. of American Control Conference (ACC)*, 2012.

[49] V. Blondel, J. Hendrickx, A. Olshevsky, and J. Tsitsiklis, “Convergence in multiagent coordination, consensus, and flocking,” in *Proc. of the Joint 44th Conference on Decision and Control And European Control Conference*, 2005, pp. 2996–3000.

[50] A. Nedić and A. Ozdaglar, “Convergence rate for consensus with delays,” *Journal of Global Optimization*, vol. 47, no. 3, p. 437456, 2010.

[51] U. Münz, A. Papachristodoulou, and F. Allgöwer, “Consensus in multiagent systems with coupling delays and switching topology,” *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2976 – 2982, 2011.

[52] K. Tsianos and M. Rabbat, “The impact of communication delays on distributed consensus algorithms,” arXiv preprint: https://arxiv.org/pdf/1207.5839.pdf, 2012.

[53] T. Charalambous, Y. Yuan, T. Yang, W. Pan, C. N. Hadjicostis, and M. Johansson, “Distributed finite-time average consensus in digraphs in the presence of time delays,” *IEEE Transactions on Control of Network Systems*, vol. 2, no. 4, pp. 370–381, Dec 2015.

[54] T. T. Doan, C. L. Beck, and R. Srikant, “On the convergence rate of distributed gradient methods for finite-sum optimization under communication delays,” *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, pp. 37:1–37:27, Dec. 2017.

[55] T. T. Doan, C. L. Beck, and R. Srikant, “Convergence rate of distributed subgradient methods under communication delays,” in *Proc. of American Control Conference (ACC)*, 2018.

[56] H. K. Khalil, *Nonlinear System*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[57] J. Hale and S. Lunel, *Introduction to Functional Diffential Equations*. Springer-Verlag, 1993, vol. 99.

[58] J. Duchi, A. Agarwal, and M. Wainwright, “Dual averaging for distributed optimization: convergence analysis and network scaling,” *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[59] T. T. Doan, “Aggregating stochastic gradients in distributed optimization,” in *Proc. of American Control Conference (ACC)*, 2018.

[60] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, Nov 2015.

[61] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization.* Wiley-Interscience series in discrete mathematics, Wiley, 1983.

[62] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operation Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[63] B.-T. Aharon, M. Tamar, and A. Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," *SIAM Journal on Optimization*, vol. 12, pp. 79–108, 2001.

[64] M. Raginsky and J. Bouvrie, "Continuous-time stochastic mirror descent on a network: variance reduction, consensus, convergence," in *Proc. IEEE 51st Conference on Decision and Control*, USA, Dec 2012, pp. 6793–6800.

[65] W. Krichene, A. M. Bayen, and P. L. Bartlett, "Accelerated mirror descent in continuous and discrete time," in *Proc. Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Canada, Dec 2015.

[66] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic mirror descent," in *Proc. 49th Annual Allerton Conference*, USA, Sept 2011, pp. 701–706.

[67] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal of Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[68] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *Proc. IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Mexico, Dec 2015, pp. 517–520.

[69] M. Nokleby and W. U. Bajwa, "Stochastic optimization from distributed, streaming data in rate-limited networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 170, no. 3, pp. 1–13, 2017. [Online]. Available: https://arxiv.org/abs/1704.07888

[70] J. Li, G. Li, Z. Wu, and C. Wu, "Stochastic mirror descent method for distributed multi-agent optimization," *Optimization Letters*, vol. 24, no. 1, pp. 1–19, 2014.

[71] Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. Glynn, "Mirror descent learning in continuous games," in *Proc. IEEE 56th Conference on Decision and Control*, Scotland, Dec 2017, pp. 5776–5783.

[72] E. C. Hall and R. M. Willett, "Online convex optimization in dynamic environments," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 647–662, 2015.

[73] G. S. Ledva, L. Balzano, and J. L. Mathieu, "Inferring the behavior of distributed energy resources with online learning," in *Proc. 53th Annual Allerton Conference*, USA, Oct 2015, pp. 6321–6326.

[74] A. Nedic, A. Olshevsky, and C. A. Uribe, "Distributed learning with infinitely many hypotheses," in *Proc. 2016 IEEE 55th Conference on Decision and Control*, USA, Dec 2016, pp. 6321–6326.

[75] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. PP, no. 99, pp. 1–12, 2017.

[76] Z. Zhou, P. Mertikopoulos, N. Bambos, P. Glynn, and C. Tomlin, "Countering feedback delays in multi-agent learning," in *Proc. The 31st International Conference on Neural Information Processing Systems*, USA, Dec 2017, pp. 5776–5783.

[77] T. T. Doan and C. L. Beck, "Distributed Lagrangian methods for network resource allocation," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, 2017, pp. 650–655.

[78] W. Krichene, S. Krichen, and A. Bayen, "Convergence of mirror descent dynamics in the routing game," in *Proc. 2015 European Control Conference*, Austria, Jul 2015, pp. 569–574.

[79] T. T. Doan, S. Bose, D. H. Nguyen, and C. L. Beck, "Convergence of the iterates in mirror descent methods," Submitted to Control Systems Letters, available at: https://www.dropbox.com/s/solia6o00mqsu7i/mirrordescent_LCSS.pdf?dl=0, 2018.

[80] H. Bauschke and J. Borwein, "Joint and separate convexity of Bregman distance," *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, pp. 23–36, 2001.

[81] O. J. Karst, "Linear curve fitting using least deviations," *Journal of the American Statistical Association*, vol. 53, no. 281, pp. 118–132, 1958.

[82] D. Precup and R. S. Sutton, "Exponentiated gradient methods for reinforcement learning," in *Proceedings of International Conference on Machine Learning*, 1997, pp. 272–277.

[83] M. Wang, Y. Chen, J. Liu, and Y. Gu, "Random multi-constraint projection: stochastic gradient methods for convex optimization with many constraints," arXiv preprint: https://arxiv.org/pdf/1511.03760.pdf, 2015.

[84] A. Nedić, "Random algorithms for convex minimization problems," *Mathematical Programming*, vol. 129, no. 2, pp. 225–253, Oct 2011.

[85] M. Wang and D. P. Bertsekas, "Incremental constraint projection methods for variational inequalities," *Math. Program.*, vol. 150, no. 2, pp. 321–363, May 2015.

[86] S. Lee and A. Nedić, "Distributed random projection algorithm for convex optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 221–229, 2013.

[87] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, vol. 38, no. 3, pp. 367–426, 1996.

[88] H. H. Bauschke, "Projection algorithms and monotone operators," Ph.D. dissertation, 1996.

[89] L. Bottou, F. Curtis, and J. Nocedal, "Optimization methods for large-ccale machine learning," arXiv preprint: https://arxiv.org/pdf/1606.04838.pdf, 2016.

[90] L. Gubin, B. Polyak, and E. Raik, "The method of projections for finding the common point of convex sets," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1–24, 1976.

[91] S. Agmon, "The relaxation method for linear inequalities," *Canadian Journal of Mathematics*, vol. 6, no. 3, pp. 382–392, 1954.

[92] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[93] C. Zhao, U. Topcu, N. Li, and S. Low, "Design and stability of load-side primary frequency control in power systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1177–1189, April 2014.

[94] F. Dörfler, J. Simpson-Porco, and F. Bullo, "Breaking the hierarchy: distributed control and economic optimality in microgrids," *IEEE Transactions on Automatic Control*, vol. 3, no. 3, pp. 241 – 253, Sept. 2016.

[95] A. Domínguez-García, S. Cady, and C. Hadjicostis, "Decentralized optimal dispatch of distributed energy resources," in *51st IEEE Conference on Decision and Control*, 2012, pp. 3688–3693.

[96] S. Yang, S. Tan, and J.-X. Xu, "Consensus based approach for economic dispatch problem in a smart grid," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4416–4426, Oct. 2013.

[97] G. Binetti, A. Davoudi, F. Lewis, D. Naso, and B. Turchiano, "Distributed consensus-based economic dispatch with transmission losses," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1711–1720, June 2014.

[98] S. Kar and G. Hug, "Distributed robust economic dispatch in power systems: a consensus + innovations approach," in *2012 IEEE Power and Energy Society General Meeting*, 2012, pp. 1–8.

[99] H. Xing, Y. Mou, M. Fu, and Z. Lin, "Distributed bisection method for economic power dispatch in smart grid," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3024–3035, Aug. 2015.

[100] A. Cherukuri and J. Cortes, "Distributed generator coordination for initialization and anytime optimization in economic dispatch," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 3, pp. 226–237, Sept 2015.

[101] H. Lakshmanan and D. de Farias, "Decentralized resource allocation in dynamic networks of agents," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 911–940, 2008.

[102] A. Nedíc, A. Olshevsky, and W. Shi, "Improved convergence rates for distributed resource allocation," Available at: https://arxiv.org/abs/1706.05441.

[103] T. T. Doan and C. L. Beck, "Distributed Resource Allocation Over Dynamic Networks with Uncertainty," Also available at: https://arxiv.org/abs/1708.03543.

[104] T. Yang, J. Lu, D. Wu, J. Wu, G. Shi, Z. Meng, and K. H. Johansson, "A distributed algorithm for economic dispatch over time-varying directed networks with delays," *IEEE Transactions on Industrial Electronics*, vol. PP, no. 99, 2016.

[105] J. Zhu, *Optimization of Power System Operation*, 2nd ed. Wiley-IEEE Press, 2008.

[106] D. Bertsekas, *Convex Optimization Theory*. Cambridge, MA: Athena Scientific, 2009.

[107] D. Bertsekas, *Nonlinear Programming: 2nd Edition*. Cambridge, MA: Athena Scientific, 1999.

[108] "IEEE 14 bus system [online]," http://icseg.iti.illinois.edu/ieee-14-bus-system/.

[109] "IEEE 118 bus system [online]," http://motor.ece.iit.edu/data/JEAS_IEEE118.doc.

[110] I. Necoara, “Random coordinate descent algorithms for multi-agent convex optimization over networks,” *IEEE Transactions on Automatic Control*, vol. 58, no. 8, pp. 2001–2012, 2013.

[111] I. Necoara, “Distributed and parallel random coordinate descent methods for huge convex programming over networks,” in *54th IEEE Conference on Decision and Control*, 2015, pp. 425–430.

[112] A. Beck, “The 2-coordinate descent method for solving double-sided simplex constrained minimization problems,” *Journal of Optimization Theory and Applications*, vol. 162, no. 3, pp. 892–919, 2014.

[113] Y. C. Ho, L. D. Servi, and R. Suri, “A class of center-free resource allocation algorithms,” *Large Scale Systems*, vol. 1, pp. 51–62, 1980.

[114] L. Xiao and S. Boyd, “Optimal scaling of a gradient method for distributed resource allocation,” *Journal of Optimization Theory and Application*, vol. 129, no. 3, pp. 469–488, 2006.

[115] H. Robbins and D. Siegmund, “A convergence theorem for nonnegative almost supermartingales and some applications,” *Optimization Methods in Statistics, Academic Press, New York*, pp. 233–257, 1971.

[116] B. Polyak, *Introduction to Optimization.* Publication division, New York, 1987: Optimization software, Inc., 1987.