Distributed Optimization Algorithms in Large-Scale Directed Networks

by

Chenguang Xi

A dissertation submitted

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Department of Electrical and Computer Engineering

Tufts University

Advisor: Professor Usman Ahmed Khan

February 2017

Dedicated to my love, Yuan

Abstract

In the interconnected world of today, distributed computation and optimization over large-scale multi-agents networks are ubiquitous. The applications can be found in various fields ranging from machine learning, signal processing, to computational finance. The increasing interest in distributed optimization is further motivated by the emergence of big data application. In one aspect, large datasets push centralized computation to the limit and the need for distributed algorithms arises quite naturally. On a similar note, transmitting the data collected in a distributed manner to a center is either too costly or violates privacy. In this thesis, we aim to solve the distributed optimization problem over multi-agent networks, where each agent has local private information and objective functions. The goal is to have agents collaborate with each other to optimize the sum of these local objective functions. Existing algorithms mostly deal with the corresponding problems under the assumption that the underlying multi-agent network is strongly-connected and undirected, i.e., if agent i sends information to agent j, then agent j also sends information to agent i. The contribution of this work lies in the relaxation of such assumptions on the network topology. In particular, we assume the communication between the agents is described by a *directed* graph. We mainly propose four algorithms, Directed-Distributed Subgradient Descent (D-DSD),

Directed-Distributed Projection Subgradient (D-DPS), DEXTRA, and ADD-OPT. D-DSD and D-DPS focus on the case when the objective functions are convex, but not necessarily differentiable. Both of the proposed algorithms achieve convergence rates of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations. D-DPS is the generalization of D-DSD from unconstrained cases to constrained cases. When the objective functions are relaxed to be smooth, i.e., they are convex as well as differentiable, we propose DEXTRA and ADD-OPT that accelerate the convergence to a linear rate $O(\tau^k)$ for $0 < \tau < 1$. Moreover, ADD-OPT supports a wider and more realistic range of step-sizes than DEXTRA. All four algorithms achieves the best known rate of convergence for this class of problems under the appropriate assumptions.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my adviser, Professor Usman Khan, for his guidance, support, encouragement and leadership through each stage of my Ph.D career. It has been a great privilege to work with him and to learn from him over the past four years. From him, I learn to be a qualified researcher as well as a tolerant person. He has set a good example that will have a lifetime influence on my attitude and behavior towards work and life. Without him, I would never be where I am today.

Many thanks to my thesis committee members, Prof. Shuchin Aeron from Electrical and Computer Engineering Department, Tufts University, Prof. Jason Rife from Mechanical Engineering Department, Tufts University, and Prof. Jose Bento from Computer Science Department, Boston College. I appreciate all their useful suggestions and insightful comments to enhance the quality of this work.

I am indebted to my wonderful colleges, Sam Safavi, Mohammadreza Doostmohammadian, Fakhteh Saadatniaki, for much joyful time together as well as many useful discussions, suggestions and encouragements. Meanwhile, I want to thank Liu Chao, Qiong Wu, Jincheng Pang, Tianyi Luo, Zhi Li, Tinghao Liang, Liangwang Chi, Shuo Zhao, for being such great friends at Tufts.

I owe my deepest gratitude to my family. I am indebted to my parents,

Bolong Xi and Yizhen Shen, 's unreserved love and care. My wife Yuan Li has shown unswerving support to me. Her encouragement helps me to overcome the most challenging time of my life. I am very lucky to spend my life with her.

Chapter 1

Introduction

1.1 Motivation

In the interconnected world of today, distributed computation, [1], and optimization, [2], over large-scale multi-agent networks are ubiquitous. The increasing interest in distributed algorithms is further motivated by the emergence of big data applications, [3]. Large datasets push centralized computation to the limit. It is often impossible for a single processor to implement any real-time algorithm due to the large data volumes. For example, in machine learning, large scales of training examples may prevent a problem from being solved effectively on a single machine. In contrast, it is much more effective to use multiple processing processors, especially when the information is naturally distributed. Thus, the need for distributed algorithms arises quite

naturally.

Besides the efficiency in computation, distributed algorithm outperforms centralized methods in other aspects. It is usually the case that data are collected in a distributed manner. Thus, to transmit the huge volume of raw data to a center is costly. A preferable solution is to process the data locally, and exchange the processed information between local processors. On the other hand, local agents often need to reserve their private information. The existence of any centralized processor may violate the privacy.

In view of these considerations, there has recently been a growing interest in extending conventional (centralized) methods, [4, 5], to distributed methods for solving optimization problems where information is distributed over a network of agents. Usually each agent in the network owns local information that is private. The agents cooperatively solve a global optimization problem through local computation and information exchange over the network. Specifically, we consider the problem of minimizing a sum of objectives, $\sum_{i=1}^{n} f_i(\mathbf{x})$, where $f_i : \mathbb{R}^p \to \mathbb{R}$ is a private objective function that belongs to the ith agent in the network. This general form has applications in distributed large-scale machine learning, [6–10], distributed averaging, [11, 12], model predictive control, [13, 14], cognitive networks, [15, 16], wireless communication, [17], coordination, [18–20], distributed source localization, [21, 22], distributed sparse optimization, [23, 24], decentralized low-rank matrix completion, [25] and fac-

torization, [26], resource scheduling, [27], message routing, [28], and interference [29], etc.

Existing distributed methods for solving optimization problems where information is distributed over multi-agent networks mostly deal with the corresponding problems under the assumption that the network is undirected and connected, i.e., if agent i can send information to agent j, then agent j can also send information to agent i. However, in practice, it is desirable to merely rely on one directional communication between agents. For instance, in a bi-directional case where each node blocks until it receives a response, deadlocks can occur when the network has cycles. In other cases, agents may broadcast at different power levels, implying communication capability in one direction, but not the other. Moreover, a one directional communication is more robust to noise interference than bi-directional communication. It is obvious that the distributed optimization problem over directed networks has wider applications than that over undirected graphs since the network topology is more flexible. For example, agents may be able to reduce the communication overhead when they have a large number of neighbors. Besides, if there exist some slow communication links, it is good for algorithms to eliminate the link such that the convergence can be accelerated. This results a directed graph. Therefore, we focus on the case in this thesis when the communication between the agents is described by a *directed* graph.

1.2 Preliminaries and Notations

In this section, we review some basics in convex optimization as well as graph theory, which support the rest of contents in this thesis.

Properties of Functions

We first list some standard definitions and properties subject to functions of our interest. The details can be found in standard literature, e.g., Ref. [4].

- **Convex functions:** A function $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is convex if, for any points $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^p$, and $\theta \in [0, 1]$, it satisfies that

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

- **Smooth and nonsmooth functions:** A function $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable if its derivative exists and is continuous. It is smooth if it has derivatives of all orders. However, in the context of convex optimization, nonsmooth functions, usually refer to functions that do not even have a first-order derivative. For example, $f(x) = x^2$ is convex and smooth, and $f(x) = |x|$ is convex but not smooth.

- **Gradient and subgradient:** When a function $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is smooth, its first-order derivative $Df(\mathbf{x}) \in \mathbb{R}^{1\times p}$ exists. The transpose of this derivative is called function $f(\mathbf{x})$'s gradient, denoted by $\nabla f(\mathbf{x}) \in$

$\mathbb{R}^p$. Its components are the partial derivatives of $f(\mathbf{x})$: $[\nabla f(\mathbf{x})]_i = \frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_i}$, $\forall i \in [1, n]$. If $f(\mathbf{x})$ is convex, the gradient $\nabla f(\mathbf{x})$ at point $\mathbf{x}$ satisfies

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}),$$

for all $\mathbf{y} \in \mathbb{R}^p$.

If a function $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is convex but not necessarily smooth, we extend the definition of gradient to subgradient. A vector $\mathbf{g}(\mathbf{x}) \in \mathbb{R}^p$ is called the subgradient of $f(\mathbf{x})$ at point $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}),$$

for all $\mathbf{y} \in \mathbb{R}^p$. The set of all subgradients of $f(\mathbf{x})$ at $\mathbf{x}$ is called the subdifferential and is denoted as $\partial f(\mathbf{x})$.

For a smooth convex function, the subgradient coincides with its gradient.

- **Lipschitz-continuous:** A function $\nabla : \mathbb{R}^p \to \mathbb{R}$ is called Lipschitz-continuous if there exists a constant $L > 0$, such that

$$\|\nabla(\mathbf{x}) - \nabla(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^p$.

- **Strong-convexity:** A smooth convex function $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ is further said to be strongly convex if there exists some positive m such that for

any point $\mathbf{x}$, $\mathbf{y}$ it satisfies

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq m\|\mathbf{x} - \mathbf{y}\|^2,$$

where $\|\cdot\|$ is the euclidean norm.

Graph Theory

We next present some basic definitions and relations regarding graphs and matrices. The details can be found in [30].

- **Undirected graph and directed graph:** An undirected graph is a graph where all the edges are bidirectional. In contrast, a graph where there exists at least one edge point in a direction is called a directed graph.

- **Strongly connected graph:** A graph is said to be strongly connected if every node is reachable from every other node.

- **Stochastic Matrix:** A stochastic matrix is a matrix used to describe the transitions of a Markov chain. Each of its entries is a nonnegative real number representing a probability. A row-stochastic matrix is a real square stochastic matrix, with each row summing to 1. A column-stochastic matrix is a real square stochastic matrix, with each column

summing to 1. A doubly-stochastic matrix is a square matrix of nonnegative real numbers with each row and column summing to 1.

For a doubly-stochastic matrix, it always satisfies that both the left and right eigenvector corresponding to eigenvalue 1 are an all-one-vector, i.e., for any $W \in \mathbb{R}^{n\times n}$ being a doubly-stochastic matrix, it satisfies

$$W\mathbf{1}_n = \mathbf{1}_n, \qquad \mathbf{1}_n^\top W = \mathbf{1}_n^\top,$$

where we denote $\mathbf{1}_n \in \mathbb{R}^n$ the n-dimensional all-one vector. In contrast, we have that for any row-stochastic matrix, W, it satisfies that

$$W\mathbf{1}_n = \mathbf{1}_n, \qquad \boldsymbol{\pi}_n^\top W = \boldsymbol{\pi}_n^\top,$$

where $\boldsymbol{\pi}_n \in \mathbb{R}^n$ is not necessary equals to $\mathbf{1}_n$. Similarly, we have for any column-stochastic matrix, W,

$$W\boldsymbol{\pi}_n = \boldsymbol{\pi}_n, \qquad \mathbf{1}_n^\top W = \mathbf{1}_n^\top.$$

Convergence rate for iterative methods

- **Linear convergence rate:** Suppose that the sequence $\{\mathbf{x}_k\}$ over k converges to the limit $\mathbf{x}^*$. We say that this sequence converges linearly to $\mathbf{x}^*$, if there exists a number $\tau \in (0, 1)$ such that

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \tau.$$

The number τ is called the rate of convergence. Moreover, if the sequence $\{\mathbf{x}_k\}$ satisfies for all k that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \tau^k,$$

for $\tau \in (0, 1)$, we call that the sequence $\{\mathbf{x}_k\}$ converges at an R-linear rate. The R-linear rate extends the definition of linear convergence rate in that the overall convergence rate remains linear while the convergence "speed" at every iteration may vary.

- **Sublinear convergence rate:** Suppose that the sequence $\{\mathbf{x}_k\}$ over k converges to the limit $\mathbf{x}^*$. We say that this sequence converges sublinearly to $\mathbf{x}^*$, if there exists a number $\tau_k \to 1$ for $k \to \infty$ such that

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \tau_k.$$

Typical sublinear rates that will appear in this thesis include $O(\frac{\ln k}{\sqrt{k}})$, $O(\frac{\ln k}{k})$.

Notations

Besides some notations that already appeared above, we use the following notations in the rest of this thesis. We use lowercase bold letters to denote vectors and uppercase italic letters to denote matrices. We denote by $[\mathbf{x}]_i$, the ith component of a vector, $\mathbf{x}$. For a matrix, A, we denote by $[A]_i$, the ith

row of A, and by $[A]_{ij}$, the (i,j)th element of A. The matrix, I_n, represents the $n \times n$ identity, and $\mathbf{1}_n$ and $\mathbf{0}_n$ are the n-dimensional vector of all 1's and 0's. The inner product of two vectors, $\mathbf{x}$ and $\mathbf{y}$, is $\langle \mathbf{x}, \mathbf{y} \rangle$. The Euclidean norm of any vector, $\mathbf{x}$, is denoted by $\|\mathbf{x}\|$. We define the square of A-matrix norm, $\|\mathbf{x}\|_A^2$, of any vector, $\mathbf{x}$, as

$$\|\mathbf{x}\|_A^2 \triangleq \langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, A^\top \mathbf{x} \rangle = \langle \mathbf{x}, \frac{A + A^\top}{2} \mathbf{x} \rangle,$$

where A is not necessarily symmetric. Note that the A-matrix norm is non-negative only when $A + A^\top$ is Positive Semi-Definite (PSD). If a symmetric matrix, A, is PSD, we write $A \succeq 0$, while $A \succ 0$ means A is Positive Definite (PD). The largest and smallest eigenvalues of a matrix A are denoted as $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$. The smallest *nonzero* eigenvalue of a matrix A is denoted as $\widetilde{\lambda}_{\min}(A)$. For any $f(\mathbf{x})$, $\nabla f(\mathbf{x})$ denotes the (sub)gradient of f at $\mathbf{x}$. Finally, we use $\mathcal{P}_{\mathcal{X}}[\mathbf{x}]$ for the projection of a vector $\mathbf{x}$ on the set $\mathcal{X}$, i.e., $\mathcal{P}_{\mathcal{X}}[\mathbf{x}] = \arg\min_{\mathbf{v} \in \mathcal{X}} \|\mathbf{v} - \mathbf{x}\|^2$.

1.3 Related Literature

In this section, we discuss existing research on the topic of distributed optimization that divide the computation load among multiple agents. Related work may be divided into three categories: Incremental method, distributed methods over undirected graphs, and distributed methods over di-

rected graphs. The earliest work of this problem can be found in [2, 31], where each agent has access to the same global objective function. Note that in our setting in this thesis, we assume that agents own private local objective functions that are unknown to others agents in the network. In the following, we briefly describe these approaches.

Incremental Methods

Incremental methods, [32–36], may be regarded the first big contribution to distributed optimization where the computation load is divided over the multi-agent network. In each iteration of these type of algorithms, the iterator passes around the network and updates itself with each agent's private local objective. Although the computation load is divided, incremental method is different from distributed methods that we study in this thesis due to the following reason. In Incremental methods, there is always exactly one agent updating at each iteration while in distributed method all agents are updated simultaneously in every single iteration. Each agent in distributed method maintains an estimate of the global optimal. Compared to distributed method, incremental methods do not fully utilize the distributed properties. Incremental methods rely on cyclic order of passing the iterator through the network. In contrast, distributed methods consider more general network topology where

agents communicate to multiple neighbors at the same time.

Distributed Methods over undirected graphs

Distributed methods for solving optimization problems where information is distributed over multi-agent networks are well-established. However, the majority of work deals with the corresponding problems under the assumption that the multi-agent network is undirected and connected, i.e., if agent i can send information to agent j, then agent j can also send information to agent i. We describe these algorithms here. The well-known Distributed Gradient Descent (DGD), [37], is the very first distributed method to solve the corresponding problem, whose convergence rate and fault tolerance are well analyzed in related literature, [38–44]. At each iteration, a gradient-related step is calculated, followed by averaging over the neighbors in the network. [45] applies a similar idea to develop a distributed algorithm in dual domain. The main advantage of DGD is computational simplicity. However, the convergence rate is slow due to the diminishing step-size, which is required to ensure exact convergence. The convergence rate of DGD with a diminishing step-size is shown to be $O(\frac{\ln k}{\sqrt{k}})$, [37]. Under a constant step-size, the algorithm accelerates to $O(\frac{1}{k})$ at the cost of inexact convergence to a neighborhood of the optimal solution, [46]. When the objective functions are further strongly-

convex, DGD has a faster convergence rate of $O(\frac{\ln k}{k})$, and DGD with a constant step-size converges linearly to a neighborhood of the optimal solution. To accelerate the slow convergence rate, some alternate approaches include the Nesterov-based methods, e.g., Distributed Nesterov Gradient (DNG) with a convergence rate of $O(\frac{\ln k}{k})$, and Distributed Nesterov gradient with Consensus iterations (DNC), [47]. The algorithm, DNC, can be interpreted to have an inner loop, where information is exchanged, within every outer loop where the optimization-step is performed. The time complexity of the outer loop is $O(\frac{1}{k^2})$ whereas the inner loop performs a substantial $O(\ln k)$ information exchanges within the kth outer loop. Therefore, the equivalent convergence rate of DNC is $O(\frac{\ln k}{k^2})$. Both DNG and DNC assume the gradient to be bounded and Lipschitz continuous at the same time. Another related approach is EXTRA, see [48] for details. It uses a constant step-size and the gradients of the last two iterates. The method converges at $O(\frac{1}{k})$ for general convex functions and converges linearly under the strong-convexity assumption.

Other related algorithms include the distributed implementation of ADMM, based on augmented Lagrangian, where at each iteration the primal and dual variables are solved to minimize a Lagrangian-related function, [49–51]. Comparing to the gradient-based methods with diminishing step-sizes, this type of method converges exactly to the optimal solution with a faster rate of $O(\frac{1}{k})$ owing to the constant step-size; and further has a linear convergence when

the objective functions are strongly-convex. However, the disadvantage is a high computation burden because each agent needs to optimize a subproblem at each iteration. To resolve this issue, Decentralized Linearized ADMM (DLM), [52], is proposed, which can be considered as a first-order approximation of decentralized ADMM. DLM converges at a linear rate if the local objective functions are strongly-convex.

Considering the methods mentioned above, Refs. [37–45] solve the corresponding distributed problem when the objective functions are not necessarily differentiable and smooth. The assumption needed is that the (sub)-gradients of objective functions are bounded. Refs. [46–52] solve the corresponding distributed problem when the objective functions are required to be differentiable and smooth, i.e., the gradient of objective functions are required to satisfy be Lipschistz continuous. Compared to the bounded gradient assumption, Lipschistz continuous assumption on the gradient is relatively more restrictive. As a result, methods under Lipschistz continuous assumption achieve faster convergence rate. All the methods mentioned above require only the first-order information of functions. To accelerate the convergence speed, some methods, [53–60], exploit the second-order information of the corresponding gradients.

All these distributed algorithms, [37–60], assume the multi-agent network to be an undirected and connected graph. In contrast, the literature concerning directed graphs is relatively limited. The challenge lies in the imbalance

caused by the asymmetric information exchange among the agents. Recall that this asymmetry is because agent i sending information to agent j does not necessarily imply that agent j can send information to agent i.

Distributed Methods over directed graphs

In the context of directed graphs, Gradient-Push (GP), [61–65], combines gradient descent and push-sum consensus. The push-sum algorithm, [66, 67], is first proposed in consensus problems[1] to achieve average-consensus given a column-stochastic matrix. The idea is based on computing the stationary distribution (the left eigenvector of the weight matrix corresponding to eigenvalue 1) for the Markov chain characterized by the multi-agent network and canceling the imbalance by dividing with the left-eigenvector. The algorithms in [61–65] follow a similar spirit of push-sum consensus and propose nonlinear (because of division) methods.

The work in [74] combines gradient descent and weight-balancing to design a distributed gradient-based method over directed graphs. The notion of weights that balance a directed network was proposed in [75], where the column-stochastic weighting matrix that describes a directed graph simultaneously converges to a doubly-stochastic matrix, which corresponds to a undirected graph. In this thesis, we refer to the method in [74] as the Weighting

[1] See, [68–73], for additional information on average consensus problems.

Balancing-Distributed Gradient Descent (WB-DGD).

Neither GP and WB-DGD requires the objective functions to have Lipschitz-continuous gradient nor being strongly-convex. This is to say that GP and WB-DGD solve the corresponding distributed problems when the objective functions are not necessarily smooth. The convergence rate of both algorithms is $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary convex functions.

When the objective functions are smooth, Refs. [76–78] modify the distributed implementation of ADMM by changing the weights that are used for communication between agents. This improvement makes it possible to balance the weights from in-neighbors and out-neighbors, which make the convergence over directed graphs possible. For general convex functions, the convergence rate is $O(\frac{1}{k})$. When the objective functions are strongly convex, the convergence rate is linear. However, this category of methods suffers a high computation burden since at each iteration, a sub-optimize problem needs solving.

1.4 Contributions

In this thesis, we relax the assumption on the underlying network topology to follow a directed graph. The main contribution of this work is that we propose four algorithms, Directed-Distributed Subgradient Descent (D-DSD),

Directed-Distributed Projection Subgradient (D-DPS), DEXTRA, and ADD-OPT that converge over the directed graph. We now summarize each algorithm in the following, the content of which can also be found in our work, [79–86].

- **Chapter III: Directed-Distributed Subgradient Descent (D-DSD):** D-DSD is a subgradient-based method that combines surplus-consensus techniques and DGD [37] to minimize the sum of local objective functions when the network topology among agents is described by a directed graph. It converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions are convex, but not necessarily differentiable. Therefore, we stick to the notation subgradient instead of gradient, implying that gradient may not exist. We will apply the standard bounded subgradient assumption in nonsmooth optimization as what can be found in GP [61], WB-DGD [74], or other related work, [37, 38]. We provide the convergence analysis and show that D-DSD converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

- **Chapter IV: Directed-Distributed Projection Subgradient (D-DPS):** D-DPS solves the distributed optimization problem over directed graphs subject to additional convex constraints. D-DPS can be viewed as a generalization of D-DSD when the constrained set changes from $\mathbb{R}^p$, implying no constraint, to a convex constrained set, $\mathcal{X} \subseteq \mathbb{R}^p$. Similar

to D-DSD, D-DPS converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions are convex, but not necessarily differentiable. The convergence analysis shows that D-DPS converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

- **Chapter V: DEXTRA:** Though D-DSD and D-DPS successfully solve the distributed optimization problem over directed networks, the convergence rates, however, are sub-linear, which is relatively slow. Compared to them, DEXTRA harnesses the smoothness to obtain a much faster convergence rate. In other words, DEXTRA converges to the optimal solution in smooth convex optimization, i.e., the local objective functions are convex and differentiable. We show that, with the appropriate step-size, DEXTRA converges at a linear rate $O(\tau^k)$ for $0 < \tau < 1$, given that the objective functions are restricted strongly-convex.

- **Chapter VI: ADD-OPT:** ADD-OPT is an improvement over DEXTRA that solves the distributed smooth optimization problem over directed graphs. Same as DEXTRA, it achieves the best known rate of convergence for this class of problems, $O(\mu^k)$ for $0 < \mu < 1$ given that the objective functions are strongly-convex, where k is the number of iterations. However, ADD-OPT supports a wider and more realistic range of step-sizes. In particular, the greatest lower bound of DEXTRA's step-

size is strictly greater than zero while that of ADD-OPT's equals exactly to zero.

In table 1.1, we summarize algorithms solving distributed optimization problems, over both undirected and directed networks. In either case, the corresponding algorithms are categorized as either primal domain methods or dual domain methods. It can be found that related literature considering the directed case is limited. We label our contributions in red text. The proposed algorithms in this thesis partially complete the distributed optimization algorithm framework.

			Algorithms	References
undirected	nonsmooth	primal	DSD, DPS, etc	[37–44]
		dual	DDA,D-ADMM	[45, 49]
	smooth	primal	DGD, DNG, DNC, EXTRA, NN, etc	[46–48, 58–60]
		dual	D-ADMM, DLM, DQM	[49–57]
directed	nonsmooth	primal	GP, WB-DGD, D-DSD, D-DPS	
	smooth	primal	DEXTRA, ADD-OPT	

Table 1.1: Distributed optimization algorithms summary.

1.5 Outline

The remainder of this thesis is organized as follows. Chapter II formulates the problem, recaps some related distributed optimization methods for solving the

problem, either over undirected graphs or directed graphs. These methods include Distributed Gradient Descent (DGD) [37] for general convex functions over undirected graphs, EXTRA [48] for smooth and strongly-convex functions over undirected graphs, and Gradient-Push (GP) [61] for general convex functions over directed graphs. Chapter III proposes the Directed-Distributed Subgradient Descent (D-DSD), which solves the distributed optimization problem over directed graphs for general convex functions. D-DSD can be viewed as an generalization of DGD. In Chapter IV, we generalize D-DSD to Directed-Distributed Projection Subgradient (D-DPS), which solves the problem with an additional convex constraint. In Chapter V, we propose a fast algorithm, termed DEXTRA, which combines EXTRA and push-sum consensus. An extended version of DEXTRA, termed ADD-OPT, is proposed in Chapter VI. Chapter VII concludes this thesis and discuss the possibility of some future works.

1.6 Summary

In this chapter, we motivate the work, and summarize the contribution made in this thesis: the distributed algorithm that converges over directed graph is important and necessary, yet not well established. The proposed algorithms in this thesis complete the distributed optimization algorithm framework.

Chapter 2

Model and Previous Work

In this chapter, we formulate the problem, and recap some related distributed optimization methods for solving the problem, either over undirected graphs or directed graphs. These methods include Distributed Gradient Descent (DGD) [37] for general convex functions over undirected graphs, EXTRA [48] for smooth and strongly-convex functions over undirected graphs, and Gradient-Push (GP) [61] for general convex functions over directed graphs.

2.1 Problem Formulation

Consider a strongly-connected network of n agents communicating over a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents, and $\mathcal{E}$ is the collection of ordered pairs, $(i, j), i, j \in \mathcal{V}$, such that agent j can send information to agent i.

Define $\mathcal{N}_i^{\text{in}}$ to be the collection of in-neighbors, i.e., the set of agents that can send information to agent i. Similarly, $\mathcal{N}_i^{\text{out}}$ is the set of out-neighbors of agent i. We allow both $\mathcal{N}_i^{\text{in}}$ and $\mathcal{N}_i^{\text{out}}$ to include the node i itself. Note that in a directed graph when $(i,j) \in \mathcal{E}$, it is not necessary that $(j,i) \in \mathcal{E}$. Consequently, $\mathcal{N}_i^{\text{in}} \neq \mathcal{N}_i^{\text{out}}$, in general. We focus on solving a convex optimization problem that is distributed over the above multi-agent network. In particular, the network of agents cooperatively solve the following optimization problem:

$$\text{P1}: \quad \min\ f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}),$$

where each local objective function, $f_i : \mathbb{R}^p \to \mathbb{R}$, is convex, and known only by agent i. Our goal is to develop a distributed iterative algorithm such that each agent converges to the global solution of Problem P1 when the communications between agents are described by a directed graph, $\mathcal{G}$.

2.2 Distributed Gradient Descent

Consider the Distributed Gradient Descent (DGD) [37] to solve P1 over undirected graphs. At each iteration, a gradient related step is calculated, followed by averaging with neighbors in the network. More specifically, at kth iteration, agent i updates its estimate, $\mathbf{x}_i^{k+1} \in \mathbb{R}^p$, as follows:

$$\mathbf{x}_i^{k+1} = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j^k - \alpha_k \nabla \mathbf{f}_i^k, \tag{2.1}$$

where w_{ij} is a non-negative weight such that $W = \{w_{ij}\}$ is doubly-stochastic. The scalar, α_k, is a diminishing but non-negative step-size, satisfying the persistence conditions, [39, 87]:

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \tag{2.2}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{2.3}$$

The vector $\nabla \mathbf{f}_i^k \in \mathbb{R}^p$ is a sub-gradient of f_i at $\mathbf{x}_i^k$. According to [37], the iteration in Eq. (2.1) drives all agents to reach consensus, i.e., $\lim_{k\to\infty} \mathbf{x}_i^k = \lim_{k\to\infty} \mathbf{x}_j^k$, $\forall i, j$, and the consensus (accumulation) point approaches to the optimal solution, i.e., $\lim_{k\to\infty} \mathbf{x}_i^k = \mathbf{x}^*$, where $\mathbf{x}^*$ is the optimal solution of P1. The convergence rate of DGD is $O(\frac{\ln k}{\sqrt{k}})$ for general convex objective functions, under the assumption that the local private convex objective function, $f_i(\mathbf{x})$, is bounded for all $\mathbf{x}$. DGD is valid for nonsmooth convex functions, i.e., the objective functions are not necessarily differentiable.

We now derive an informal but intuitive proof showing how DGD pushes agents to achieve consensus and reach the optimal solution. We first write the matrix form of Eq. (2.1). Denote $\mathbf{x}^k$, $\nabla \mathbf{f}(\mathbf{x}^k) \in \mathbb{R}^{n\times p}$, and $W \in \mathbb{R}^{n\times n}$ as follow,

$$\mathbf{x}^k = \begin{bmatrix} \left(\mathbf{x}_1^k\right)^\top \\ \vdots \\ \left(\mathbf{x}_n^k\right)^\top \end{bmatrix}, \quad \nabla \mathbf{f}(\mathbf{x}^k) = \begin{bmatrix} \nabla \mathbf{f}_1^\top(\mathbf{x}_1^k) \\ \vdots \\ \nabla \mathbf{f}_n^\top(\mathbf{x}_n^k) \end{bmatrix}, \quad W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}.$$

Thus, Eq. (2.1) can be written in a matrix form as

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha_k \nabla \mathbf{f}(\mathbf{x}^k). \tag{2.4}$$

In Eq. (2.4), we have that the weighting matrix, W, is doubly-stochastic. For the sake of argument, let us assume that the sequences, $\{\mathbf{x}^k\}$, generated by Eq. (2.4), converge to its own limit, $\mathbf{x}^\infty$, (not necessarily true). Note that $\lim_{k\to\infty} \alpha_k = 0$. From Eq. (2.4), we have that

$$\begin{aligned} \mathbf{x}^\infty &= W\mathbf{x}^\infty - \alpha_\infty \nabla \mathbf{f}(\mathbf{x}^\infty) \\ &= W\mathbf{x}^\infty, \end{aligned} \tag{2.5}$$

which implies that $\mathbf{x}^\infty \in \text{span}\{\mathbf{1}_n\}$ considering that $\mathbf{1}_n = W\mathbf{1}_n$ is satisfied for any doubly stochastic matrix. Therefore, the consensus property is established. We now consider the optimality property. It follows from Eq. (2.4) that

$$\begin{aligned} \overline{\mathbf{x}}^{k+1} \triangleq& \frac{1}{n}\mathbf{1}_n^\top \mathbf{x}^{k+1} \\ =& \frac{1}{n}\mathbf{1}_n^\top W\mathbf{x}^k - \frac{1}{n}\alpha_k \mathbf{1}_n^\top \nabla \mathbf{f}(\mathbf{x}^k) \\ =& \overline{\mathbf{x}}^k - \frac{1}{n}\alpha_k \nabla \overline{\mathbf{f}}^k, \end{aligned} \tag{2.6}$$

where we denote $\nabla \overline{\mathbf{f}}^k$ as $\mathbf{1}_n^\top \nabla \mathbf{f}(\mathbf{x}^k)$. By considering the consensus property, it follows that the preceding relation can be regarded as an inexact centralized gradient descent method (using $\nabla \overline{\mathbf{f}}^k = \mathbf{1}_n^\top \nabla \mathbf{f}(\mathbf{x}^k)$ instead of $\nabla f^\top(\overline{\mathbf{x}}^k)$) with step-size $\frac{\alpha_k}{n}$ to minimize the global objective function $f(\mathbf{x})$ of Problem P1. Therefore, the optimality property is achieved.

2.3 EXTRA

EXTRA is a fast exact first-order algorithm that solve Problem P1 when the communication network is undirected. At the kth iteration, agent i performs the following update:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_j^k - \sum_{j\in\mathcal{N}_i} \widetilde{w}_{ij}\mathbf{x}_j^{k-1} - \alpha\left[\nabla f_i(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^{k-1})\right], \qquad (2.7)$$

where the weights, w_{ij}, form a weighting matrix, $W = \{w_{ij}\}$, that is symmetric and doubly-stochastic. The collection $\widetilde{W} = \{\widetilde{w}_{ij}\}$ satisfies $\widetilde{W} = \theta I_n + (1-\theta)W$, with some $\theta \in (0, \frac{1}{2}]$. The update in Eq. (2.7) converges to the optimal solution at each agent i with a convergence rate of $O(\frac{1}{k})$ and converges linearly when the objective functions are strongly-convex. To better represent EXTRA, we write Eq. (2.7) in a matrix form. Let $\mathbf{x}^k$, $\nabla\mathbf{f}(\mathbf{x}^k) \in \mathbb{R}^{n\times p}$ be the collections of all agent states and gradients at time k,

$$\mathbf{x}^k = \begin{bmatrix} \left(\mathbf{x}_1^k\right)^\top \\ \vdots \\ \left(\mathbf{x}_n^k\right)^\top \end{bmatrix}, \quad \nabla\mathbf{f}(\mathbf{x}^k) = \begin{bmatrix} \nabla f_1^\top(\mathbf{x}_1^k) \\ \vdots \\ \nabla f_n^\top(\mathbf{x}_n^k) \end{bmatrix},$$

and W, $\widetilde{W} \in \mathbb{R}^{n\times n}$ be the weighting matrices collecting weights, w_{ij}, $\widetilde{w}_{ij}$, respectively. Then, Eq. (2.7) can be represented in a matrix form as:

$$\mathbf{x}^{k+1} = (I_n + W)\mathbf{x}^k - \widetilde{W}\mathbf{x}^{k-1} - \alpha\left[\nabla\mathbf{f}(\mathbf{x}^k) - \nabla\mathbf{f}(\mathbf{x}^{k-1})\right]. \qquad (2.8)$$

Note that the difference between EXTRA and DGD lies in two aspects. Firstly, it uses two weighting matrices instead of just one weighting matrix in DGD. Secondly, the step-size used in EXTRA is constant while that in DGD is diminishing. This is the real reason why EXTRA has a much faster convergence rate compared to DGD. As the index number, k, going large, a diminishing step-size, α_k, is small, which slows the performance.

We now derive an informal but intuitive proof showing that how EXTRA pushes agents to achieve consensus and reach the optimal solution. For the sake of argument, let us assume that the sequences, $\{\mathbf{x}^k\}$, generated by Eq. (2.8), converge to its own limit, $\mathbf{x}^\infty$, (not necessarily true). We first show the consensus property. Let k goes to infinity, we obtain form Eq. (2.8) that

$$\mathbf{x}^\infty = (I_n + W)\mathbf{x}^\infty - \widetilde{W}\mathbf{x}^\infty - \alpha\left[\nabla\mathbf{f}(\mathbf{x}^\infty) - \nabla\mathbf{f}(\mathbf{x}^\infty)\right], \tag{2.9}$$

which implies that $(W - \widetilde{W})\mathbf{x}^\infty = \mathbf{0}_{n\times p}$, or $\mathbf{x}^\infty \in \text{span}\{\mathbf{1}_n\}$, where the consensus is reached. Summing up Eq. (2.8) over k from 0 to ∞, we obtain that

$$\mathbf{x}^\infty = W\mathbf{x}^\infty - \alpha\nabla\mathbf{f}(\mathbf{x}^\infty) - \sum_{r=0}^{\infty}\left(\widetilde{W} - W\right)\mathbf{x}^r. \tag{2.10}$$

Consider that $\mathbf{x}^\infty = W\mathbf{x}^\infty$ from the consensus property and the preceding relation, it follows that

$$\alpha\nabla\mathbf{f}(\mathbf{x}^\infty) = \sum_{r=0}^{\infty}\left(\widetilde{W} - W\right)\mathbf{x}^r. \tag{2.11}$$

Therefore, we obtain that

$$\alpha \mathbf{1}_n^\top \nabla \mathbf{f}(\mathbf{x}^\infty) = -\mathbf{1}_n^\top \left(\widetilde{W} - W\right) \sum_{r=0}^{\infty} \mathbf{x}^r = \mathbf{0}_p^\top, \tag{2.12}$$

which is the optimality condition of Problem P1 considering that $\mathbf{x}^\infty \in \text{span}\{\mathbf{1}_n\}$.

2.4 Gradient-Push

Unlike DGD and EXTRA which are two methods solving Problem P1 over undirected graphs, GP is the first distributed method that solves Problem P1 over directed networks. We describe the implementation of GP as follow. Each agent, $j \in \mathcal{V}$, maintains two vector variables: $\mathbf{x}_j^k$, $\mathbf{z}_j^k \in \mathbb{R}^p$, as well as a scalar variable, $y_j^k \in \mathbb{R}$, where k is the discrete-time index. At the kth iteration, agent j weights its states, $a_{ij}\mathbf{x}_j^k$ and $a_{ij}y_j^k$, and sends these to each of its out-neighbors, $i \in \mathcal{N}_j^{\text{out}}$, where the weights, a_{ij},'s are such that:

$$a_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{i=1}^{n} a_{ij} = 1, \forall j, \tag{2.13}$$

where $\theta \in (0, \frac{1}{2}]$. With agent i receiving the information from its in-neighbors, $j \in \mathcal{N}_i^{\text{in}}$, it calculates the state, $\mathbf{z}_i^k$, by dividing $\mathbf{x}_i^k$ over y_i^k, and updates $\mathbf{x}_i^{k+1}$ and y_i^{k+1} as follows:

$$\mathbf{z}_i^k = \frac{\mathbf{x}_i^k}{y_i^k}, \tag{2.14a}$$

$$\mathbf{x}_i^{k+1} = \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij}\mathbf{x}_j^k\right) - \alpha_k \nabla f_i(\mathbf{z}_i^k), \tag{2.14b}$$

$$y_i^{k+1} = \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij} y_j^k \right). \tag{2.14c}$$

In the above, $\nabla f_i(\mathbf{z}_i^k)$ is the gradient of the function $f_i(\mathbf{z})$ at $\mathbf{z} = \mathbf{z}_i^k$. The scalar, α_k, is a diminishing but non-negative step-size, satisfying the persistence conditions, [39, 87]:

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \tag{2.15}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{2.16}$$

The method is initiated with an arbitrary vector, $\mathbf{x}_i^0$, and with $y_i^0 = 1$ for any agent i. We note that the implementation of Eq. (2.14) needs each agent to have the knowledge of its out-neighbors (such that it can design the weights according to Eq. (5.1). In a more restricted setting, e.g., a broadcast application where it may not be possible to know the out-neighbors, we may use $a_{ij} = |\mathcal{N}_j^{\text{out}}|^{-1}$; thus, the implementation only requires each agent to know its out-degree. We write Eq. (2.14) in a matrix form. Let $\mathbf{x}^k$, $\mathbf{z}^k$, $\nabla \mathbf{f}(\mathbf{z}^k) \in \mathbb{R}^{n \times p}$, and $\mathbf{y}^k \in \mathbb{R}^n$ be the collections of all agent states and gradients at time k,

$$\mathbf{x}^k = \begin{bmatrix} \left(\mathbf{x}_1^k\right)^\top \\ \vdots \\ \left(\mathbf{x}_n^k\right)^\top \end{bmatrix}, \mathbf{y}^k = \begin{bmatrix} \left(y_1^k\right)^\top \\ \vdots \\ \left(y_n^k\right)^\top \end{bmatrix}, \mathbf{z}^k = \begin{bmatrix} \left(\mathbf{z}_1^k\right)^\top \\ \vdots \\ \left(\mathbf{z}_n^k\right)^\top \end{bmatrix}, \nabla \mathbf{f}(\mathbf{z}^k) = \begin{bmatrix} \nabla f_1^\top(\mathbf{z}_1^k) \\ \vdots \\ \nabla f_n^\top(\mathbf{z}_n^k) \end{bmatrix}.$$

Let $A = \{a_{ij}\}$ be the weighting matrix collecting all the weights. It is straightforward that A is a column stochastic matrix. Define a diagonal matrix, $D^k \in \mathbb{R}^{n\times n}$, for each k, such that the ith element of D^k is y_i^k, i.e.,

$$D^k = \text{diag}\left(\mathbf{y}^k\right) = \text{diag}\left(A^k \cdot \mathbf{1}_n\right). \tag{2.17}$$

Given that the graph, $\mathcal{G}$, is strongly-connected and the corresponding weighting matrix, A, is non-negative, it follows that D^k is invertible for any k. Then, we can write Eq. (2.14) in the matrix form equivalently as follows: Based on the above notations, we have that

$$\mathbf{z}^k = (D^k)^{-1}\mathbf{x}^k, \tag{2.18a}$$

$$\mathbf{x}^{k+1} = A\mathbf{x}^k - \alpha_k \nabla \mathbf{f}(\mathbf{z}^k), \tag{2.18b}$$

$$\mathbf{y}^{k+1} = A\mathbf{y}^k. \tag{2.18c}$$

Compare Eqs. (2.18)(b) and (2.4), we realize that the difference between GP and DGD is that the weighting matrix changes from a doubly-stochastic matrix, W, to a column-stochastic matrix, A. To overcome the difficulties, GP arguments two additional variables $\mathbf{y}^k$ and $\mathbf{z}^k$. The consensus property can still be achieved by dividing $\mathbf{x}^k$ over $\mathbf{y}^k$, Eq. (2.18)(a), which cancels the imbalance causing by the asymmetric information exchange in directed graphs.

2.5 Summary

In this chapter, we formulated the problem, and recapped some related distributed optimization methods for solving the problem, either over undirected graphs or directed graphs. Distributed Gradient Descent (DGD) [37] converges at $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary convex functions over undirected graphs in nonsmooth optimization. EXTRA [48] converges at $O(\frac{1}{k})$ for smooth convex functions over undirected graphs and a linear convergence rate can be achieved when the objective functions are strongly-convex. Gradient-Push (GP) [61] is the first work considering directed graph topology. It is valid in nonsmooth optimization with the same convergence rate as DGD.

Chapter 3

D-DSD for Nonsmooth Convex Optimization

In this chapter, we introduce an algorithm, termed Directed-Distributed Subgradient Descent (D-DSD), to solve the distributed optimization problem, P1, over directed networks. D-DSD converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions in Problem P1 are convex, but not necessarily differentiable. We first discuss the performance of DGD, [37], over directed graphs. In particular, we show the reason why DGD fails to converge to the optimal solution over directed graphs. Motivated by the analysis, we propose D-DSD, which can be viewed as an extension of DGD to the case of directed networks. We provide the convergence analysis

and show that D-DSD converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations. Numerical experiments illustrate the findings.

3.1 Motivation

When we restrict the assumption of network topology from undirected networks to directed networks, we no longer achieve a doubly-stochastic weighting matrix. Instead, we can only construct a row-stochastic weighting matrix as well as a column-stochastic weighting matrix. For a doubly-stochastic matrix, it always satisfies that both the left and right eigenvector corresponding to eigenvalue 1 are an all-one-vector, i.e., for any W being a doubly-stochastic matrix, it satisfies

$$W\mathbf{1}_n = \mathbf{1}_n, \qquad \mathbf{1}_n^\top W = \mathbf{1}_n^\top. \tag{3.1}$$

In contrast, we have that for any row-stochastic matrix, W, it satisfies that

$$W\mathbf{1}_n = \mathbf{1}_n, \qquad \boldsymbol{\pi}_n^\top W = \boldsymbol{\pi}_n^\top, \tag{3.2}$$

where $\boldsymbol{\pi}_n \in \mathbb{R}^n$ is not necessary equals to $\mathbf{1}_n$. Similarly, we have for any column-stochastic matrix, W,

$$W\boldsymbol{\pi}_n = \boldsymbol{\pi}_n, \qquad \mathbf{1}_n^\top W = \mathbf{1}_n^\top. \tag{3.3}$$

Recall the update of DGD, Eq. (2.1),

$$\mathbf{x}_i^{k+1} = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j^k - \alpha_k \nabla f_i^k, \tag{3.4}$$

and its matrix form, Eq. (2.4)

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha_k \nabla \mathbf{f}(\mathbf{x}^k). \tag{3.5}$$

For the sake of argument, consider W to be row-stochastic but not column-stochastic. Clearly, $\mathbf{1}_n$ is a right eigenvector of W, and let $\boldsymbol{\pi}_n = \{\pi_i\}$ be its left eigenvector corresponding to eigenvalue 1. Summing over i in Eq. (3.4), we get

$$\begin{aligned}
\widehat{\mathbf{x}}^{k+1} &\triangleq \sum_{i=1}^{n} \pi_i \mathbf{x}_i^{k+1} \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \pi_i w_{ij} \right) \mathbf{x}_j^k - \alpha_k \sum_{i=1}^{n} \pi_i \nabla f_i(\mathbf{x}_i^k) \\
&= \widehat{\mathbf{x}}^k - \alpha_k \sum_{i=1}^{n} \pi_i \nabla f_i(\mathbf{x}_i^k),
\end{aligned} \tag{3.6}$$

where $\pi_j = \sum_{i=1}^{n} \pi_i w_{ij}, \forall i, j$. If we assume that the agents reach an agreement, then Eq. (3.6) can be viewed as an inexact (central) gradient descent (with $\sum_{i=1}^{n} \pi_i \nabla f_i(\mathbf{x}_i^k)$ instead of $\sum_{i=1}^{n} \pi_i \nabla f_i(\widehat{\mathbf{x}}^k)$) minimizing a new objective, $\widehat{f}(\mathbf{x}) \triangleq \sum_{i=1}^{n} \pi_i f_i(\mathbf{x})$. As a result, the agents reach consensus and converge to the minimizer of $\widehat{f}(\mathbf{x})$.

Now consider the weight matrix, W, to be column-stochastic but not row-stochastic. Let $\overline{\mathbf{x}}^k$ be the average of agents estimates at time k, then Eq. (3.4)

leads to

$$\begin{aligned}\overline{\mathbf{x}}^{k+1} &\triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{k+1}\\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}w_{ij}\right)\mathbf{x}_j^k - \frac{\alpha_k}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^k)\\ &= \overline{\mathbf{x}}^k - \left(\frac{\alpha_k}{n}\right)\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^k). \end{aligned} \tag{3.7}$$

Eq. (3.7) reveals that the average, $\overline{\mathbf{x}}^k$, of agents estimates follows an inexact (central) gradient descent ($\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^k)$ instead of $\sum_{i=1}^{n}\nabla f_i(\overline{\mathbf{x}}^k)$) with stepsize α^k/n, thus reaching the minimizer of $f(\mathbf{x})$. Despite the fact that the average, $\overline{\mathbf{x}}^k$, reaches the optima, $\mathbf{x}^*$, of $f(\mathbf{x})$, the optima is not achievable for each agent because consensus can not be reached with a matrix that is not necessary row-stochastic.

Eqs. (3.6) and (3.7) explain the importance of doubly-stochastic matrices in consensus-based optimization. The row-stochasticity guarantees all of the agents to reach a consensus, while column-stochasticity ensures each local gradient to contribute equally to the global objective. In other words, we note that reaching a consensus requires the right eigenvector (corresponding to eigenvalue 1) to lie in span$\{\mathbf{1}_n\}$, and minimizing the global objective requires the corresponding left eigenvector to lie in span$\{\mathbf{1}_n\}$. Both the left and right eigenvectors of a doubly-stochastic matrix are $\mathbf{1}_n$, which, in general, is not possible in directed graphs. Therefore, we aim to construct a new

weight matrix, $W \in \mathbb{R}^{2n\times 2n}$, whose left and right eigenvectors (corresponding to eigenvalue 1) are in the form: $[\mathbf{1}_n^\top, \mathbf{v}^\top]$ and $[\mathbf{1}_n^\top, \mathbf{u}^\top]^\top$, for some vector $\mathbf{v}$ and $\mathbf{u}$.

3.2 Algorithm and Assumptions

We introduce *Directed-Distributed Subgradient Descent* (D-DSD) that overcomes the above issues by augmenting an additional variable at each agent and thus constructing a new weight matrix, $M \in \mathbb{R}^{2n\times 2n}$, whose left and right eigenvectors (corresponding to eigenvalue 1) are in the form: $[\mathbf{1}_n^\top, \mathbf{v}^\top]$ and $[\mathbf{1}_n^\top, \mathbf{u}^\top]^\top$, for some vector $\mathbf{v}$ and $\mathbf{u}$. Formally, we describe D-DSD as follows.

At kth iteration, each agent, $j \in \mathcal{V}$, maintains two vectors: $\mathbf{x}_j^k$ and $\mathbf{y}_j^k$, both in $\mathbb{R}^p$. Agent j sends its state estimate, $\mathbf{x}_j^k$, as well as a weighted auxiliary variable, $b_{ij}\mathbf{y}_j^k$, to each out-neighbor, $i \in \mathcal{N}_j^{\text{out}}$, where b_{ij}'s are such that:

$$b_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{i=1}^{n} b_{ij} = 1, \qquad \forall j.$$

Agent i updates the variables, $\mathbf{x}_i^{k+1}$ and $\mathbf{y}_i^{k+1}$, with the information received from its in-neighbors, $j \in \mathcal{N}_i^{\text{in}}$, as follows:

$$\mathbf{x}_i^{k+1} = \sum_{j=1}^{n} a_{ij}\mathbf{x}_j^k + \epsilon\mathbf{y}_i^k - \alpha_k \nabla f_i(\mathbf{x}_i^k), \tag{3.8a}$$

$$\mathbf{y}_i^{k+1} = \mathbf{x}_i^k - \sum_{j=1}^{n} a_{ij}\mathbf{x}_j^k + \sum_{j=1}^{n} b_{ij}\mathbf{y}_j^k - \epsilon\mathbf{y}_i^k, \tag{3.8b}$$

where:

$$a_{ij} = \begin{cases} > 0, & j \in \mathcal{N}_i^{\text{in}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{j=1}^{n} a_{ij} = 1, \qquad \forall i.$$

The diminishing step-size, $\alpha_k \geq 0$, satisfies the persistence conditions, [39, 87]: $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. The scalar, ϵ, is a small positive number, which plays a key role in the convergence of the algorithm[1]. For an illustration of the message passing between agents in the implementation of Eq. (3.8), see

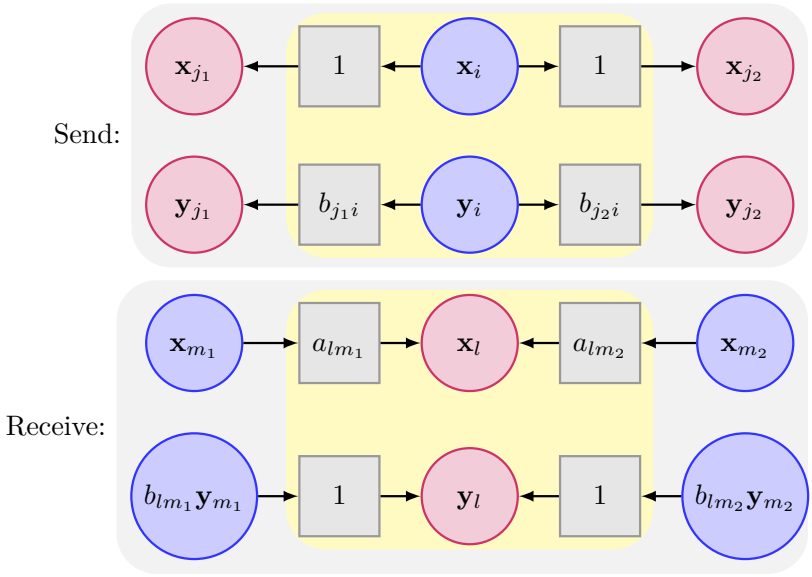


Figure 3.1: Illustration of the message passing between agents by Eq. (3.8).

[1]Note that in the implementation of Eq. (3.8), each agent needs the knowledge of its out-neighbors. In a more restricted setting, e.g., a broadcast application where it may not be possible to know the out-neighbors, we may use $b_{ij} = |\mathcal{N}_j^{\text{out}}|^{-1}$; thus, the implementation only requires knowing the out-degrees, see, e.g., [61, 62] for similar assumptions.

Fig. 3.1 on how agent i sends information to its out-neighbors and agent l receives information from its in-neighbors. In Fig. 3.1, the weights $b_{j_1 i}$ and $b_{j_2 i}$ are designed by the sender, agent i, and satisfy $b_{ii} + b_{j_1 i} + b_{j_2 i} = 1$. The weights a_{lm_1} and a_{lm_2} are designed by the receiver, agent l, and satisfy $b_{ll} + b_{lm_1} + b_{lm_2} = 1$. To analyze the algorithm, we denote $\mathbf{z}_i^k \in \mathbb{R}^p$, $\mathbf{g}_i^k \in \mathbb{R}^p$, and $M \in \mathbb{R}^{2n \times 2n}$ as follows:

$$\mathbf{z}_i^k = \begin{cases} \mathbf{x}_i^k, & i \in \{1, ..., n\}, \\ \mathbf{y}_{i-n}^k, & i \in \{n+1, ..., 2n\}, \end{cases}$$

$$\mathbf{g}_i^k = \begin{cases} \nabla f_i(\mathbf{x}_i^k), & i \in \{1, ..., n\}, \\ 0_p, & i \in \{n+1, ..., 2n\}, \end{cases}$$

$$M = \begin{bmatrix} A & \epsilon I \\ I - A & B - \epsilon I \end{bmatrix}, \tag{3.9}$$

where $A = \{a_{ij}\}$ is row-stochastic, $B = \{b_{ij}\}$ is column-stochastic. Consequently, Eq. (3.8) can be represented compactly as follows: for any $i \in \{1, ..., 2n\}$, at $k+1$th iteration,

$$\mathbf{z}_i^{k+1} = \sum_{j=1}^{2n} [M]_{ij} \mathbf{z}_j^k - \alpha_k \mathbf{g}_i^k. \tag{3.10}$$

We refer to the iterative relation in Eq. (3.10) as the Directed-Distributed Subgradient Descent (D-DSD) method, since it has the same form as DGD except the dimension doubles due to a new weight matrix $M \in \mathbb{R}^{2n \times 2n}$ as defined in Eq. (3.9). It is worth mentioning that even though Eq. (3.10) looks

similar to DGD, [37], the convergence analysis of D-DSD does not exactly follow that of DGD. This is because the weight matrix, M, has negative entries. Besides, M is not a doubly-stochastic matrix, i.e., the row sum is not 1. Hence, the tools in the analysis of DGD are not applicable, e.g., $\|\sum_j [M]_{ij}\mathbf{z}_j - \mathbf{x}^*\| \leq \sum_j [M]_{ij}\|\mathbf{z}_j - \mathbf{x}^*\|$ does not necessarily hold because $[M]_{ij}$ are not non-negative.

We now describe the assumptions to ensure the convergence of D-DSD to the optimal solution of Problem P1 over directed networks.

Assumption A1. *In order to ensure the convergence of D-DSD to the optimal solution of Problem P1 over directed networks, we make the following assumptions:*

(a) The agent graph, $\mathcal{G}$, is strongly-connected.

(b) The optimizer of Problem P1 and the corresponding optimal value exist and are unique. Formally, we have

$$f(\mathbf{x}^*) = f^* = \min f(\mathbf{x}).$$

(c) Each local function, $f_i : \mathbb{R}^p \to \mathbb{R}$, is convex, $\forall i \in \mathcal{V}$.

(d) For each local objective function, it is not necessarily differentiable. The sub-gradient, $\nabla f_i(\mathbf{x})$, is bounded:

$$\|\nabla f_i(\mathbf{x})\| \leq D,$$

for all $\mathbf{x} \in \mathbb{R}^p, i \in \mathcal{V}$.

The Assumptions A1 are standard in nonsmooth distributed optimization, see related literature, [38], and references therein. Note that in Assumption A1(d), we claim that the objective functions are not necessarily differentiable. Therefore, we use term sub-gradient instead of gradient in this chapter since the gradients of functions do not necessary exist. We now present the convergence analysis of D-DSD.

3.3 Convergence Analysis

The convergence analysis of D-DSD can be divided into two parts. In the first part, we discuss the *consensus property* of D-DSD, i.e., we capture the decrease in $\left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\|$ for $i \in \{1, ..., n\}$, as D-DSD progresses, where we define $\overline{\mathbf{z}}^k$ as the accumulation point:

$$\overline{\mathbf{z}}^k \triangleq \frac{1}{n}\sum_{j=1}^{2n} \mathbf{z}_i^k = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_i^k + \frac{1}{n}\sum_{j=1}^{n} \mathbf{y}_i^k. \tag{3.11}$$

The decrease in $\left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\|$ reveals that all agents approach a common accumulation point. We then show the *optimality property* in the second part, i.e., the decrease in the difference between the function evaluated at the accumulation point and the optimal solution, $f(\overline{\mathbf{z}}^k) - f(\mathbf{x}^*)$. We combine the two parts to establish the convergence.

Consensus Property

To show the consensus property, we study the convergence behavior of the weight matrices, M^k, in Eq. (3.9) as k goes to infinity. We use an existing results on such matrices M, based on which we show the convergence behavior as well as the convergence rate. We borrow the following from [88].

Lemma 1. *(Cai et al. [88]) Assume the graph is strongly-connected. M is the weighting matrix defined in Eq.* (3.9)*, and the constant ϵ in M satisfies $\epsilon \in (0, \Upsilon)$, where $\Upsilon := \frac{1}{(20+8n)^n}(1 - |\lambda_3|)^n$, where λ_3 is the third largest eigenvalue of M in Eq.* (3.9) *by setting $\epsilon = 0$. Then the weighting matrix, M, defined in Eq.* (3.9)*, has a simple eigenvalue* 1 *and all other eigenvalues have magnitude smaller than one.*

Based on Lemma 1, we now provide the convergence behavior as well as the convergence rate of the weight matrix, M.

Lemma 2. *Assume that the network is strongly-connected, and M is the weight matrix that defined in Eq.* (3.9)*. Then,*

(a) The sequence of $\{M^k\}$, as k goes to infinity, converges to the following limit:

$$\lim_{k\to\infty} M^k = \begin{bmatrix} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} & \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\ 0 & 0 \end{bmatrix};$$

(b) For all $i, j \in \mathcal{V}$, the entries $\left[M^k\right]_{ij}$ converge to their limits as $k \to \infty$ at a geometric rate, i.e., there exist bounded constants, $\Gamma \in \mathbb{R}$, and $0 < \gamma < 1$, such that

$$\left\| M^k - \begin{bmatrix} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} & \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\ 0 & 0 \end{bmatrix} \right\|_\infty \leq \Gamma \gamma^k.$$

Proof. Note that the sum of each column of M equals one, so 1 is an eigenvalue of M with a corresponding left (row) eigenvector $[\mathbf{1}_n^\top \ \mathbf{1}_n^\top]$. We further have $M[\mathbf{1}_n^\top \ \mathbf{0}_n^\top]^\top = [\mathbf{1}_n^\top \ \mathbf{0}_n^\top]^\top$, so $[\mathbf{1}_n^\top \ \mathbf{0}_n^\top]^\top$ is a right (column) eigenvector corresponding to the eigenvalue 1. According to Lemma 1, 1 is a simple eigenvalue of M and all other eigenvalues have magnitude smaller than one. We represent M^k in the Jordan canonical form for some P_i and Q_i

$$M^k = \frac{1}{n}[\mathbf{1}_n^\top \ \mathbf{0}_n^\top]^\top [\mathbf{1}_n^\top \ \mathbf{1}_n^\top] + \sum_{i=2}^{n} P_i J_i^k Q_i, \tag{3.12}$$

where the diagonal entries in J_i are smaller than one in magnitude for all i. The statement (a) follows by noting that $\lim_{k\to\infty} J_i^k = 0$, for all i.

From Eq. (3.12), and with the fact that all eigenvalues of M except 1 have magnitude smaller than one, there exist some bounded constants, Γ and $\gamma \in (0, 1)$, such that

$$\begin{aligned} \left\| M^k - \begin{bmatrix} \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} & \frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \\ 0 & 0 \end{bmatrix} \right\| &= \left\| \sum_{i=2}^{n} P_i J_i^k Q_i \right\|, \\ &\leq \sum_{i=2}^{n} \|P_i\| \, \|Q_i\| \, \left\|J_i^k\right\| \end{aligned}$$

$$\leq \Gamma\gamma^k,$$

from which we get the desired result. □

Using the result from Lemma 1, Lemma 2 shows the convergence behavior of the power of the weight matrix, and further show that its convergence is bounded by a geometric rate. Lemma 2 plays a key role in proving the consensus properties of D-DSD. Based on Lemma 2, we bound the difference between agent estimates in the following lemma. More specifically, we show that the agent estimates, $\mathbf{x}_i^k$, approaches the accumulation point, $\overline{\mathbf{z}}^k$, and the auxiliary variable, $\mathbf{y}_i^k$, goes to $\mathbf{0}_n$, where $\overline{\mathbf{z}}^k$ is defined in Eq. (3.11).

Lemma 3. *Let the Assumptions A1 hold. Let* $\left\{\mathbf{z}_i^k\right\}$ *be the sequence over* k *generated by the D-DSD algorithm, Eq.* (3.10)*. Then, there exist some bounded constants,* Γ *and* $0 < \gamma < 1$*, such that:*

(a) for $1 \leq i \leq n$*, and* $k \geq 1$*,*

$$\left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\| \leq \Gamma\gamma^k \sum_{j=1}^{2n} \left\|\mathbf{z}_j^0\right\| + n\Gamma D \sum_{r=1}^{k-1} \gamma^{k-r}\alpha_{r-1} + 2D\alpha_{k-1};$$

(b) for $n+1 \leq i \leq 2n$*, and* $k \geq 1$*,*

$$\left\|\mathbf{z}_i^k\right\| \leq \Gamma\gamma^k \sum_{j=1}^{2n} \left\|\mathbf{z}_j^0\right\| + n\Gamma D \sum_{r=1}^{k-1} \gamma^{k-r}\alpha_{r-1}.$$

Proof. For any $k \geq 1$, we write Eq. (3.10) recursively

$$\mathbf{z}_i^k = \sum_{j=1}^{2n} [M^k]_{ij}\mathbf{z}_j^0 - \sum_{r=1}^{k-1}\sum_{j=1}^{2n} [M^{k-r}]_{ij}\alpha_{r-1}\mathbf{g}_j^{r-1} - \alpha_{k-1}\mathbf{g}_i^{k-1}. \tag{3.13}$$

Since every column of M sums up to one, we have for any r $\sum_{i=1}^{2n}[M^r]_{ij} = 1$. Considering the recursive relation of $\mathbf{z}_i^k$ in Eq. (3.13), we obtain that $\overline{\mathbf{z}}^k$ can be represented as

$$\overline{\mathbf{z}}^k = \sum_{j=1}^{2n}\frac{1}{n}\mathbf{z}_j^0 - \sum_{r=1}^{k-1}\sum_{j=1}^{2n}\frac{1}{n}\alpha_{r-1}\mathbf{g}_j^{r-1} - \frac{1}{n}\sum_{j=1}^{2n}\alpha_{k-1}\mathbf{g}_j^{k-1}. \tag{3.14}$$

Subtracting Eq. (3.14) from (3.13) and taking the norm, we obtain that for $1 \leq i \leq n$,

$$\begin{aligned}\left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\| \leq &\sum_{j=1}^{2n}\left\|[M^k]_{ij} - \frac{1}{n}\right\|\left\|\mathbf{z}_j^0\right\| + \sum_{r=1}^{k-1}\sum_{j=1}^{n}\left\|[M^{k-r}]_{ij} - \frac{1}{n}\right\|\alpha_{r-1}\left\|\nabla f_j(\mathbf{x}_j^{r-1})\right\| \\ &+ \alpha_{k-1}\left\|\nabla f_i(\mathbf{x}_i^{k-1})\right\| + \frac{1}{n}\sum_{j=1}^{n}\alpha_{k-1}\left\|\nabla f_j(\mathbf{x}_j^{k-1})\right\|. \end{aligned}\tag{3.15}$$

The proof of part (a) follows by applying the result of Lemma 2 to Eq. (3.15) and noticing that the gradient is bounded by a constant D. Similarly, by taking the norm of Eq. (3.13), we obtain that for $n+1 \leq i \leq 2n$,

$$\left\|\mathbf{z}_i^k\right\| \leq \sum_{j=1}^{2n}\left\|[M^k]_{ij}\right\|\left\|\mathbf{z}_j^0\right\| + \sum_{r=1}^{k-1}\sum_{j=1}^{n}\left\|[M^{k-r}]_{ij}\right\|\alpha_{r-1}\left\|\nabla f_j(\mathbf{x}_j^{r-1})\right\|.$$

The proof of part (b) follows by applying the result of Lemma 2 to the preceding relation and considering the boundedness of gradient in Assumption A1(d). □

Using the above lemma, we now draw our first conclusion on the consensus property at the agents. Proposition 1 reveals that all agents asymptotically reach consensus.

Proposition 1. *Let the Assumptions A1 hold. Let* $\left\{\mathbf{z}_i^k\right\}$ *be the sequence over* k *generated by the D-DSD algorithm, Eq.* (3.10)*. Then,* $\mathbf{z}_i^k$ *satisfies*

(a) for $1 \leq i \leq n$,

$$\sum_{k=1}^{\infty} \alpha_k \left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\| < \infty;$$

(b) for $n+1 \leq i \leq 2n$,

$$\sum_{k=1}^{\infty} \alpha_k \left\|\mathbf{z}_i^k\right\| < \infty.$$

Proof. Based on the result of Lemma 3(a), we obtain, for $1 \leq i \leq n$,

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k \left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\| \leq & \Gamma \left(\sum_{j=1}^{2n} \left\|\mathbf{z}_j^0\right\|\right) \sum_{k=1}^{K} \alpha_k \gamma^k + n\Gamma D \sum_{k=1}^{K} \sum_{r=1}^{k-1} \gamma^{(k-r)} \alpha_k \alpha_{r-1} \\ & + 2D \sum_{k=0}^{K-1} \alpha_k^2. \end{aligned} \tag{3.16}$$

With the basic inequality $ab \leq \frac{1}{2}(a^2 + b^2)$, $a, b \in \mathbb{R}$, we have:

$$2\sum_{k=1}^{K} \alpha_k \gamma^k \leq \sum_{k=1}^{K} \left[\alpha_k^2 + \gamma^{2k}\right] \leq \sum_{k=1}^{K} \alpha_k^2 + \frac{1}{1-\gamma^2};$$

and

$$\begin{aligned}\sum_{k=1}^{K} \sum_{r=1}^{k-1} \gamma^{(k-r)} \alpha_k \alpha_{r-1} &\leq \frac{1}{2} \sum_{k=1}^{K} \alpha_k^2 \sum_{r=1}^{k-1} \gamma^{(k-r)} + \frac{1}{2} \sum_{r=1}^{K-1} (\alpha_{r-1})^2 \sum_{k=r+1}^{K} \gamma^{(k-r)} \\ &\leq \frac{1}{1-\gamma} \sum_{k=1}^{K} \alpha_k^2. \end{aligned}$$

The proof of part (a) follows by applying the preceding relations to Eq. (3.16) along with $\sum_{k=0}^{K} \alpha_k^2 < \infty$ as $K \to \infty$. Following the same spirit in the proof of part (b), we can reach the conclusion of part (b). □

Since $\sum_{k=1}^{\infty} \alpha_k = \infty$, Proposition 1 shows that all agents reach consensus at the accumulation point, $\overline{\mathbf{z}}^k$, asymptotically, i.e., for all $1 \leq i \leq n$, $1 \leq j \leq n$,

$$\lim_{k\to\infty} \mathbf{z}_i^k = \lim_{k\to\infty} \overline{\mathbf{z}}^k = \lim_{k\to\infty} \mathbf{z}_j^k, \tag{3.17}$$

and for $n+1 \leq i \leq 2n$, the states, $\mathbf{z}_i^k$, asymptotically, converge to zero, i.e., for $n+1 \leq i \leq 2n$,

$$\lim_{k\to\infty} \mathbf{z}_i^k = 0. \tag{3.18}$$

We next show how the accumulation point, $\overline{\mathbf{z}}^k$, approaches the optima, $\mathbf{x}^*$, as D-DSD progresses.

Optimality Property

The following lemma gives an upper bound on the difference between the objective evaluated at the accumulation point, $f(\overline{\mathbf{z}}^k)$, and the optimal objective value, f^*.

Lemma 4. *Let the Assumptions A1 hold. Let* $\left\{\mathbf{z}_i^k\right\}$ *be the sequence over* k *generated by the D-DSD algorithm, Eq.* (3.10)*. Then,*

$$\begin{aligned} 2\sum_{k=0}^{\infty} \alpha_k \left(f(\overline{\mathbf{z}}^k) - f^*\right) \leq & n \left\|\overline{\mathbf{z}}^0 - \mathbf{x}^*\right\|^2 + nD^2 \sum_{k=0}^{\infty} \alpha_k^2 \\ & + \frac{4D}{n} \sum_{i=1}^{n} \sum_{k=0}^{\infty} \alpha_k \left\|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\right\|. \end{aligned} \tag{3.19}$$

Proof. Consider Eq. (3.10) and the fact that each column of M sums to one, we have

$$\begin{aligned}\overline{\mathbf{z}}^{k+1} &= \frac{1}{n}\sum_{j=1}^{2n}\left[\sum_{i=1}^{2n}[M]_{ij}\right]\mathbf{z}_j^k - \alpha_k\frac{1}{n}\sum_{i=1}^{2n}\mathbf{g}_i^k,\\ &= \overline{\mathbf{z}}^k - \frac{\alpha_k}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i^k).\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}^*\right\|^2 = \left\|\overline{\mathbf{z}}^k-\mathbf{x}^*\right\|^2 + \left\|\frac{\alpha_k}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{z}_i^k)\right\|^2\\ - 2\frac{\alpha_k}{n}\sum_{i=1}^{n}\left\langle\overline{\mathbf{z}}^k-\mathbf{x}^*,\nabla f_i(\mathbf{z}_i^k)\right\rangle.\end{aligned} \tag{3.20}$$

Denote $\nabla f_i^k = \nabla f_i(\mathbf{z}_i^k)$. Since $\|\nabla f_i^k\| \le D$, we have

$$\begin{aligned}\left\langle\overline{\mathbf{z}}^k-\mathbf{x}^*,\nabla f_i^k\right\rangle &= \left\langle\overline{\mathbf{z}}^k-\mathbf{z}_i^k,\nabla f_i^k\right\rangle + \left\langle\mathbf{z}_i^k-\mathbf{x}^*,\nabla f_i^k\right\rangle\\ &\ge \left\langle\overline{\mathbf{z}}^k-\mathbf{z}_i^k,\nabla f_i^k\right\rangle + f_i(\mathbf{z}_i^k) - f_i(\mathbf{x}^*)\\ &\ge -D\left\|\mathbf{z}_i^k-\overline{\mathbf{z}}^k\right\| + f_i(\mathbf{z}_i^k) - f_i(\overline{\mathbf{z}}^k) + f_i(\overline{\mathbf{z}}^k) - f_i(\mathbf{x}^*)\\ &\ge -2D\left\|\mathbf{z}_i^k-\overline{\mathbf{z}}^k\right\| + f_i(\overline{\mathbf{z}}^k) - f_i(\mathbf{x}^*).\end{aligned} \tag{3.21}$$

By substituting Eq. (3.21) in Eq. (3.20), and rearranging the terms, we obtain that

$$\begin{aligned}2\alpha_k\left(f(\overline{\mathbf{z}}^k)-f^*\right) \le& n\left\|\overline{\mathbf{z}}^k-\mathbf{x}^*\right\|^2 - n\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}^*\right\|^2 + nD^2\alpha_k^2\\ &+ \frac{4D}{n}\sum_{i=1}^{n}\alpha_k\left\|\mathbf{z}_i^k-\overline{\mathbf{z}}^k\right\|.\end{aligned} \tag{3.22}$$

The desired result is achieved by summing Eq. (3.22) over time from $k=0$ to ∞. □

We are ready to present the main result of this paper, by combining all the preceding results.

Theorem 1. *Let the Assumptions A1 hold. Let* $\left\{\mathbf{z}_i^k\right\}$ *be the sequence over* k *generated by the D-DSD algorithm, Eq.* (3.10)*. Then, for any agent* i*, we have*

$$\lim_{k\to\infty} f(\mathbf{z}_i^k) = f^*.$$

Proof. Since that the step-size follows that $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, and $\sum_{k=0}^{\infty} \alpha_k \|\mathbf{z}_i^k - \overline{\mathbf{z}}^k\| < \infty$ from Lemma 1, we obtain from Eq. (3.19) that

$$2\sum_{k=0}^{\infty} \alpha_k \left(f(\overline{\mathbf{z}}^k) - f^*\right) < \infty, \tag{3.23}$$

which reveals that $\lim_{k\to\infty} f(\overline{\mathbf{z}}^k) = f^*$, by considering that $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\left(f(\overline{\mathbf{z}}^k) - f^*\right) > 0$ for all k. In Eq. (3.17), we have already shown that $\lim_{k\to\infty} \mathbf{z}_i^k = \lim_{k\to\infty} \overline{\mathbf{z}}^k$. Therefore, we obtain the desired result. □

Convergence Rate

We now show the convergence rate of D-DSD. Let $f_m := \min_k f(\overline{\mathbf{z}}^k)$, we have

$$(f_m - f^*) \sum_{k=0}^{K} \alpha_k \leq \sum_{k=0}^{K} \alpha_k (f(\overline{\mathbf{z}}^k) - f^*) \tag{3.24}$$

By combining Eqs. (3.16), (3.19) and (3.24), it can be verified that Eq. (3.19) can be represented in the following form:

$$(f_m - f^*) \sum_{k=0}^{K} \alpha_k \leq C_1 + C_2 \sum_{k=0}^{K} \alpha_k^2,$$

or equivalently,

$$(f_m - f^*) \leq \frac{C_1}{\sum_{k=0}^{K} \alpha_k} + \frac{C_2 \sum_{k=0}^{K} \alpha_k^2}{\sum_{k=0}^{K} \alpha_k}, \tag{3.25}$$

where the constants, C_1 and C_2, are given by

$$C_1 = \frac{n}{2} \left\| \overline{\mathbf{z}}^0 - \mathbf{x}^* \right\|^2 - \frac{n}{2} \left\| \overline{\mathbf{z}}^{K+1} - \mathbf{x}^* \right\|^2 + D\Gamma \sum_{j=1}^{2n} \left\| \mathbf{z}_j^0 \right\| \frac{1}{1-\gamma^2},$$

$$C_2 = \frac{nD^2}{2} + 4D^2 + D\Gamma \sum_{j=1}^{2n} \left\| \mathbf{z}_j^0 \right\| + \frac{2D^2\Gamma}{1-\gamma}.$$

Eq. (3.25) actually has the same form as the equations in analyzing the convergence rate of DGD (recall, e.g., [37]). In particular, when $\alpha_k = k^{-1/2}$, the first term in Eq. (3.25) leads to

$$\frac{C_1}{\sum_{k=0}^{K} \alpha_k} = C_1 \frac{1/2}{K^{1/2} - 1} = O\left(\frac{1}{\sqrt{K}}\right),$$

while the second term in Eq. (3.25) leads to

$$\frac{C_2 \sum_{k=0}^{K} \alpha_k^2}{\sum_{k=0}^{K} \alpha_k} = C_2 \frac{\ln K}{2(\sqrt{K} - 1)} = O\left(\frac{\ln K}{\sqrt{K}}\right).$$

It can be observed that the second term dominates, and the overall convergence rate is $O\left(\frac{\ln k}{\sqrt{k}}\right)$. As a result, D-DSD has the same convergence rate as DGD. The restriction of directed graph does not effect the speed.

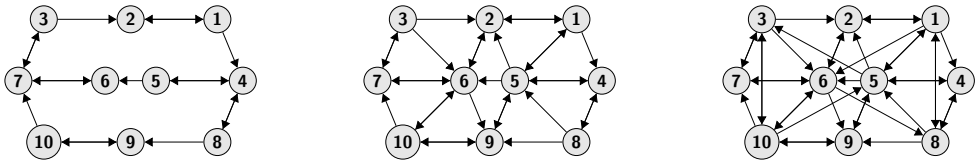
3.4 Numerical Experiment

We consider a distributed least squares problem in a directed graph: each agent owns a private objective function, $\mathbf{s}_i = R_i\mathbf{x} + \mathbf{n}_i$, where $\mathbf{s}_i \in \mathbb{R}^{m_i}$ and $R_i \in$

$\mathbb{R}^{m_i \times p}$ are measured data, $\mathbf{x} \in \mathbb{R}^p$ is unknown states, and $\mathbf{n}_i \in \mathbb{R}^{m_i}$ is random unknown noise. The goal is to estimate $\mathbf{x}$. This problem can be formulated as a distributed optimization problem solving

$$\min\ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \|R_i\mathbf{x} - \mathbf{s}_i\|\,.$$

We consider the network topology as the digraphs shown in Fig. 3.2. We employ identical setting and graphs as [88]. In [88], the value of $\epsilon = 0.7$ is chosen for each $\mathcal{G}_a, \mathcal{G}_b, \mathcal{G}_c$.



$\mathcal{G}_a$ $\mathcal{G}_b$ $\mathcal{G}_c$

Figure 3.2: Three examples of strongly-connected but non-balanced digraphs.

Fig. 3.3 shows the convergence of the D-DSD algorithm for three digraphs displayed in Fig. 3.2. Once the weight matrix, M, defined in Eq. (3.9), converges, the D-DSD ensures the convergence. Moreover, it can be observed that the residuals decrease faster as the number of edges increases, from $\mathcal{G}_a$ to $\mathcal{G}_c$. This indicates faster convergence when there are more communication channels available for information exchange.

Fig. 3.4 shows the convergence of the D-DSD algorithm with three different value of ϵ. We have shown earlier that the value of ϵ plays a key role in the

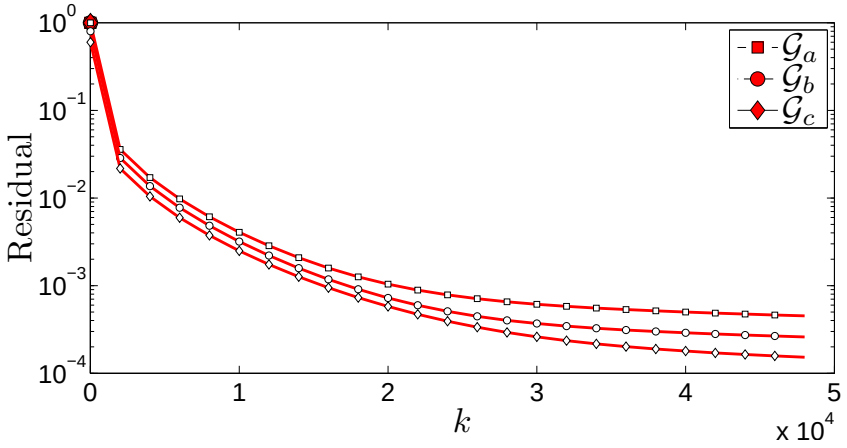


Figure 3.3: Plot of residuals for digraph $\mathcal{G}_a, \mathcal{G}_b, \mathcal{G}_c$ as D-DSD progresses.

convergence of the new weighting matrix, M.It can be find in Fig. 3.4 that in practical experiments, the range of ϵ is much larger than the theoretical range shown earlier.

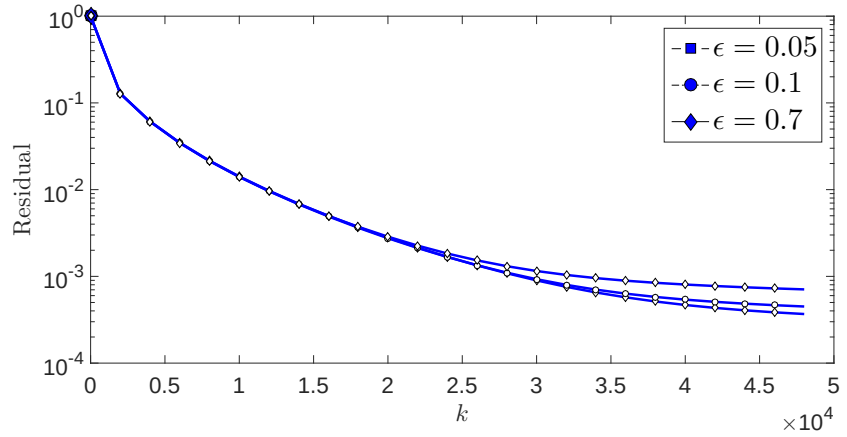


Figure 3.4: Plot of residuals for different ϵ as D-DSD progresses.

In Fig. 3.5, we display the trajectories of both states, $\mathbf{x}$ and $\mathbf{y}$, when the

D-DSD, Eq. (3.10), is applied on digraph $\mathcal{G}_a$ with parameter $\epsilon = 0.7$. Recall that in Eqs. (3.17) and (3.18), we have shown that as times, k, goes to infinity, the state, $\mathbf{x}_i^k$ of all agents will converges to a same accumulation point, $\overline{\mathbf{z}}^k$, which is the optimal solution of the problem, and $\mathbf{y}_i^k$ of all agents converges to zero, which are shown in Fig. 3.5.

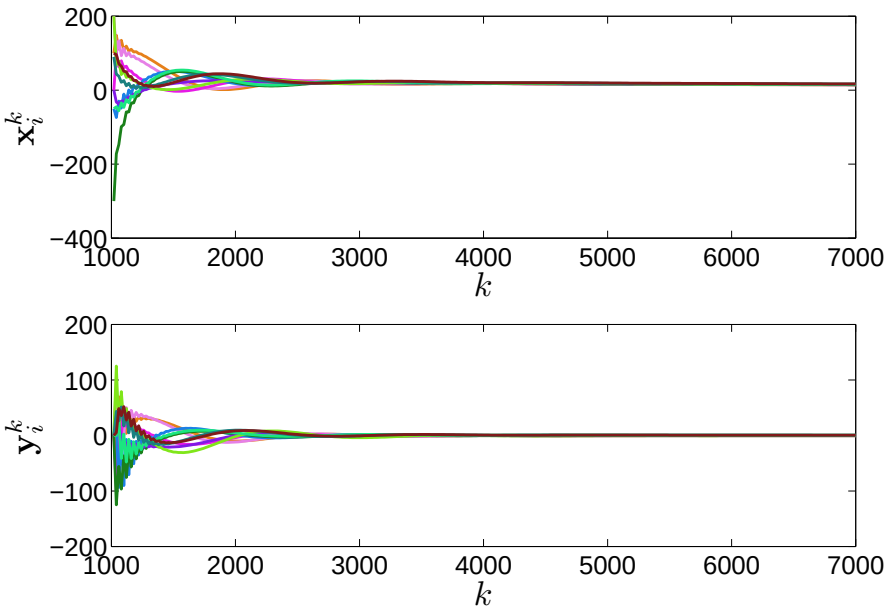


Figure 3.5: Sample paths of states, $\mathbf{x}_i^k$, and $\mathbf{y}_i^k$, on digraphs $\mathcal{G}_a$ with $\epsilon = 0.7$.

In the next experiment, we compare the performance between the D-DSD and others distributed optimization algorithms over directed graphs. The red curve in Fig. 3.6 is the plot of residuals of D-DSD on $\mathcal{G}_a$. In Fig. 3.6, we also shown the convergence behavior of two other algorithms on the same digraph. The blue line is the plot of residuals with a DGD algorithm using a row-

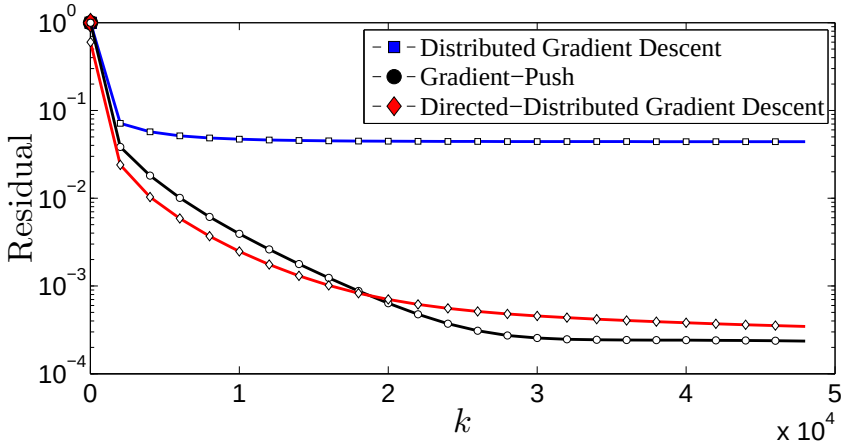


Figure 3.6: Comparison on convergence rate between different algorithms.

stochastic matrix. As we have discussed, when the weight matrix is restricted to be row-stochastic, DGD actually minimizes a new objective function $\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} \pi_i f_i(\mathbf{x})$ where $\boldsymbol{\pi} = \{\pi_i\}$ is the left eigenvector of the weight matrix corresponding to eigenvalue 1. So it does not converge to the true $\mathbf{x}^*$. The black curve shows the convergence behavior of the gradient-push algorithm, proposed in [61, 62]. Our algorithm has the same convergence rate as the gradient-push algorithm, which is $O\left(\frac{\ln k}{\sqrt{k}}\right)$.

3.5 Conclusions and Future Work

In this chapter, we describe a distributed algorithm, called Directed-Distributed Subgradient Descent (D-DSD), to solve the problem of minimizing a sum of convex objective functions over a *directed* graph. Existing distributed algo-

rithms, e.g., Distributed Gradient Descent (DGD), deal with the same problem under the assumption of undirected networks. The primary reason behind assuming the undirected graphs is to obtain a doubly-stochastic weight matrix. The row-stochasticity of the weight matrix guarantees that all agents reach consensus, while the column-stochasticity ensures optimality, i.e., each agents local gradient contributes equally to the global objective. In a directed graph, however, it may not be possible to construct a doubly-stochastic weight matrix in a distributed manner. In each iteration of D-DSD, we simultaneously constructs a row-stochastic matrix and a column-stochastic matrix instead of only a doubly-stochastic matrix. The convergence of the new weight matrix, depending on the row-stochastic and column-stochastic matrices, ensures agents to reach both consensus and optimality. The analysis shows that the D-DSD converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

In the analysis of D-DSD, we stick to the setting of static directed networks. Although we do not pursue this here, D-DSD can be generalized to work over time-varying directed graphs. Numerical experiments illustrate this findings. Extending the analysis to the case of time-varying directed graphs would be important directions for future work.

Chapter 4

D-DPS for Constrained Nonsmooth Convex Optimization

In this chapter, we propose a distributed algorithm, termed Directed-Distributed Projection Subgradient (D-DPS), to solve the distributed optimization problem over directed networks with an additional constrained set. D-DSD can be viewed as a special case of D-DPS when the constrained set is $\mathbb{R}^p$, meaning no constraint. Same as D-DSD, D-DPS converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions in the problem are convex, but not necessarily differentiable. The convergence analysis shows

that D-DPS converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

4.1 Problem, Assumptions, and Algorithm

Consider a strongly-connected network of n agents communicating over a *directed* graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents, and $\mathcal{E}$ is the collection of ordered pairs, $(i, j), i, j \in \mathcal{V}$, such that agent j can send information to agent i. Define $\mathcal{N}_i^{\text{in}}$ to be the collection of in-neighbors that can send information to agent i. Similarly, $\mathcal{N}_i^{\text{out}}$ is defined as the out-neighbors of agent i. We allow both $\mathcal{N}_i^{\text{in}}$ and $\mathcal{N}_i^{\text{out}}$ to include the node i itself. In our case, $\mathcal{N}_i^{\text{in}} \neq \mathcal{N}_i^{\text{out}}$ in general. We focus on solving a constrained convex optimization problem that is distributed over the above multi-agent network. In particular, the network of agents cooperatively solve the following optimization problem:

$$\text{P2}: \quad \text{minimize} \quad f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x}), \qquad \text{subject to} \quad \mathbf{x} \in \mathcal{X},$$

where each local objective function $f_i : \mathbb{R}^p \to \mathbb{R}$ being convex, not necessarily differentiable, is only known by agent i, and the constrained set, $\mathcal{X} \subseteq \mathbb{R}^p$, is convex and closed.

The goal is to solve Problem P2 in a distributed manner such that the agents do not exchange the objective function with each other, but only share their own states with their out-neighbors in each iteration. Note that Problem P1 which we consider in previous chapters is a special case of P2 when $\mathcal{X} = \mathbb{R}^p$.

As a result, D-DPS can be viewed as an extension of D-DSD to constrained optimization. We adopt the same assumptions as the assumptions used in D-DSD. For the sake of argument, we claim these assumptions again here.

Assumption A2. *The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is strongly-connected, i.e., $\forall i, j \in \mathcal{V}$, there exists a directed path from j to i. Assumption A2 ensures that the private information of any agent is disseminated to the whole network.*

Assumption A3. *Each function, f_i, is convex, but not necessarily differentiable. The subgradient, $\nabla f_i(\mathbf{x})$, is bounded, i.e., $\|\nabla f_i(\mathbf{x})\| \leq B_{f_i}$, $\forall \mathbf{x} \in \mathbb{R}^p$. With $B = \max_i\{B_{f_i}\}$, we have for any $\mathbf{x} \in \mathbb{R}^p$,*

$$\|\nabla f_i(\mathbf{x})\| \leq B, \qquad \forall i \in \mathcal{V}. \tag{4.1}$$

We now formally describe the implementation of D-DPS. Let each agent, $j \in \mathcal{V}$, maintain two vectors: $\mathbf{x}_j^k$ and $\mathbf{y}_j^k$, both in $\mathbb{R}^p$, where k is the discrete-time index. At the $k+1$th iteration, agent j sends its state estimate, $\mathbf{x}_j^k$, as well as a weighted auxiliary variable, $b_{ij}\mathbf{y}_j^k$, to each out-neighbor, $i \in \mathcal{N}_j^{\text{out}}$, where all those out-weights, b_{ij}'s, of agent j satisfy:

$$b_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{i=1}^{n} b_{ij} = 1.$$

Agent i then updates the variables, $\mathbf{x}_i^{k+1}$ and $\mathbf{y}_i^{k+1}$, with the information received from its in-neighbors, $j \in \mathcal{N}_i^{\text{in}}$:

$$\mathbf{x}_i^{k+1} = \mathcal{P}_{\mathcal{X}} \left[\sum_{j=1}^{n} a_{ij} \mathbf{x}_j^k + \epsilon \mathbf{y}_i^k - \alpha_k \nabla \mathbf{f}_i^k \right], \tag{4.2a}$$

$$\mathbf{y}_i^{k+1} = \mathbf{x}_i^k - \sum_{j=1}^{n} a_{ij} \mathbf{x}_j^k + \sum_{j=1}^{n} \left(b_{ij} \mathbf{y}_j^k \right) - \epsilon \mathbf{y}_i^k, \tag{4.2b}$$

where the in-weights, a_{ij}'s, of agent i satisfy that:

$$a_{ij} = \begin{cases} > 0, & j \in \mathcal{N}_i^{\text{in}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{j=1}^{n} a_{ij} = 1;$$

$\mathcal{P}_{\mathcal{X}}[\cdot]$ is the projection operator on the set $\mathcal{X}$. The scalar, ϵ, is a small positive constant, of which we will give the range later. The diminishing step-size, $\alpha_k \geq 0$, satisfies the persistence conditions: $\sum_{k=0}^{\infty} \alpha_k = \infty; \sum_{k=0}^{\infty} \alpha_k^2 < \infty$; and $\nabla \mathbf{f}_i^k = \nabla f_i(\mathbf{x}_i^k)$ represents the subgradient of f_i at $\mathbf{x}_i^k$. We provide the proof of D-DPS in next section, where we show that all agents states converge to some common accumulation state, and the accumulation state converges to the optimal solution of the problem, i.e., $\mathbf{x}_i^{\infty} = \mathbf{x}_j^{\infty} = \mathbf{x}^{\infty}$ and $f(\mathbf{x}^{\infty}) = f^*$, $\forall i, j$, where f^* denotes the optimal solution of Problem P2. To facilitate the proof, we present some existing results regarding the convergence of a new weighting matrix, and some inequality satisfied by the projection operator. The first lemma is presented as Lemma 5 in Chapter III. We claim it here again to facilitate the analysis.

Lemma 5. *Let Assumption A2 holds. Let M be the weighting matrix, Eq.* (3.9)*, and the constant ϵ in M satisfy $\epsilon \in (0, \Upsilon)$, where $\Upsilon := \frac{1}{(20+8n)^n}(1-|\lambda_3|)^n$ and λ_3 is the third largest eigenvalue of M by setting $\epsilon = 0$. Then:*

(a) The sequence of $\left\{M^k\right\}$, as k goes to infinity, converges to the following limit:

$$\lim_{k\to\infty} M^k = \begin{bmatrix} \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} & \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} \\ 0 & 0 \end{bmatrix};$$

(b) For all $i, j \in [1, \ldots, 2n]$, the entries $\left[M^k\right]_{ij}$ converge at a geometric rate, i.e., there exist bounded constants, $\Gamma \in \mathbb{R}^+$, and $\gamma \in (0, 1)$, such that

$$\left\| M^k - \begin{bmatrix} \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} & \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} \\ 0 & 0 \end{bmatrix} \right\|_\infty \leq \Gamma\gamma^k.$$

The proof and related discussion can be found in Lemma 5 in Chapter III. The next lemma regarding the projection operator is from [38].

Lemma 6. *Let $\mathcal{X}$ be a non-empty closed convex set in $\mathbb{R}^p$. For any vector $\mathbf{y} \in \mathcal{X}$ and $\mathbf{x} \in \mathbb{R}^p$, it satisfies:*

(a) $\langle \mathbf{y} - \mathcal{P}_\mathcal{X}[\mathbf{x}], \mathbf{x} - \mathcal{P}_\mathcal{X}[\mathbf{x}] \rangle \leq 0.$

(b) $\|\mathcal{P}_\mathcal{X}[\mathbf{x}] - \mathbf{y}\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 - \|\mathcal{P}_\mathcal{X}[\mathbf{x}] - \mathbf{x}\|^2.$

4.2 Convergence Analysis

To analyze D-DPS, we write Eq. (4.2) in a compact form. We denote $\mathbf{z}_i^k \in \mathbb{R}^p$, $\mathbf{g}_i^k \in \mathbb{R}^p$ as

$$\mathbf{z}_i^k = \begin{cases} \mathbf{x}_i^k, & 1 \leq i \leq n, \\ \mathbf{y}_{i-n}^k, & n+1 \leq i \leq 2n, \end{cases}$$
$$\mathbf{g}_i^k = \begin{cases} \mathbf{x}_i^{k+1} - \sum_{j=1}^{n} a_{ij}\mathbf{x}_j^k - \epsilon \mathbf{y}_i^k, & 1 \leq i \leq n, \\ \mathbf{0}_p, & n+1 \leq i \leq 2n, \end{cases} \tag{4.3}$$

and $A = \{a_{ij}\}, B = \{b_{ij}\}$, and $M = \{m_{ij}\}$ collect the weights from Eqs. (4.2) and (3.9). We now represent Eq. (4.2) as follows: for any $i \in \{1, ..., 2n\}$, at $k+1$th iteration,

$$\mathbf{z}_i^{k+1} = \sum_{j=1}^{2n} m_{ij}\mathbf{z}_j^k + \mathbf{g}_i^k, \tag{4.4}$$

where we refer to $\mathbf{g}_i^k$ as the *perturbation*. Eq. (4.4) can be viewed as a distributed subgradient method, [37], where the doubly stochastic matrix is substituted with the new weighting matrix, M, Eq. (3.9), and the subgradient is replaced by the perturbation, $\mathbf{g}_i^k$. We summarize the spirit of the upcoming convergence proof, which consists of proving both the consensus property and the optimality property of D-DPS. As to the consensus property, we show that the disagreement between estimates of agents goes to zero, i.e., $\lim_{k\to\infty} \|\mathbf{x}_i^k - \mathbf{x}_j^k\| = 0$, $\forall i, j \in \mathcal{V}$. More specifically, we show that the limit

of agent estimates converge to some accumulation state, $\overline{\mathbf{z}}^k = \frac{1}{n}\sum_{i=1}^{2n} \mathbf{z}_i^k$, i.e., $\lim_{k\to\infty} \|\mathbf{x}_i^k - \overline{\mathbf{z}}^k\| = 0$, $\forall i$, and the agents additional variables go to zero, i.e., $\lim_{k\to\infty} \|\mathbf{y}_i^k\| = 0$, $\forall i$. Based on the consensus property, we next show the optimality property that the difference between the objective function evaluated at the accumulation state and the optimal solution goes to zero, i.e., $\lim_{k\to\infty} f(\overline{\mathbf{z}}^k) = f^*$.

We formally define the accumulation state $\overline{\mathbf{z}}^k$ as follow,

$$\overline{\mathbf{z}}^k = \frac{1}{n}\sum_{i=1}^{2n} \mathbf{z}_i^k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^k + \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i^k. \tag{4.5}$$

The following lemma regarding $\mathbf{x}_i^k$, $\mathbf{y}_i^k$, and $\overline{\mathbf{z}}^k$ is straightforward. We assume that all of the initial states of agents are zero, i.e., $\mathbf{z}_i^k = \mathbf{0}_p$, $\forall i$, for the sake of simplicity in the representation of proof.

Lemma 7. *Let Assumptions A2, A3 hold. Then, there exist some bounded constants, $\Gamma > 0$ and $0 < \gamma < 1$, such that:*

(a) for all $i \in \mathcal{V}$ and $k \geq 0$, the agent estimate satisfies[1]

$$\left\|\mathbf{x}_i^k - \overline{\mathbf{z}}^k\right\| \leq \Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \sum_{j=1}^{n} \left\|\mathbf{g}_j^{r-1}\right\| + \sum_{j=1}^{n} \left\|\mathbf{g}_j^{k-1}\right\|;$$

(b) for all $i \in \mathcal{V}$ and $k \geq 0$, the additional variable satisfies

$$\left\|\mathbf{y}_i^k\right\| \leq \Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \sum_{j=1}^{n} \left\|\mathbf{g}_j^{r-1}\right\|.$$

[1]In this paper, we allow the notation that the superscript of sum being smaller than its subscript. In particular, for any sequence $\{\mathbf{s}_k\}$, we have $\sum_{k=k_1}^{k_2} \mathbf{s}_k = 0$, if $k_2 < k_1$. Besides, we denote in this paper for convenience that $\mathbf{g}_i^{-1} = \mathbf{0}_p$, $\forall i$

Proof. For any $k \geq 0$, we write Eq. (4.4) recursively

$$\mathbf{z}_i^k = \sum_{r=1}^{k-1} \sum_{j=1}^{n} [M^{k-r}]_{ij} \mathbf{g}_j^{r-1} + \mathbf{g}_i^{k-1}. \tag{4.6}$$

We have $\sum_{i=1}^{2n} [M^k]_{ij} = 1$ for any $k \geq 0$ since each column of M sums up to one. Considering the recursive relation of $\mathbf{z}_i^k$ in Eq. (4.6), we obtain that $\overline{\mathbf{z}}^k$ can be written as

$$\overline{\mathbf{z}}^k = \sum_{r=1}^{k-1} \sum_{j=1}^{n} \frac{1}{n} \mathbf{g}_j^{r-1} + \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i^{k-1}. \tag{4.7}$$

Subtracting Eq. (4.7) from (4.6) and taking the norm, we obtain

$$\begin{aligned} \left\| \mathbf{z}_i^k - \overline{\mathbf{z}}^k \right\| \leq & \sum_{r=1}^{k-1} \sum_{j=1}^{n} \left\| [M^{k-r}]_{ij} - \frac{1}{n} \right\| \left\| \mathbf{g}_j^{r-1} \right\| \\ & + \frac{n-1}{n} \left\| \mathbf{g}_i^{k-1} \right\| + \frac{1}{n} \sum_{j \neq i} \left\| \mathbf{g}_j^{k-1} \right\|. \end{aligned} \tag{4.8}$$

The proof of part (a) follows by applying Lemma 5 to Eq. (4.8) for $1 \leq i \leq n$, whereas the proof of part (b) follows by applying Lemma 5 to Eq. (4.6) for $n+1 \leq i \leq 2n$. □

Convergence of the perturbation

We now show that the perturbation, $\mathbf{g}_i^k$, goes to zero, i.e., at kth iteration, the norm of the perturbation, $\mathbf{g}_i^k$, at any agent can be bounded by the step-size times some positive bounded constant, i.e., there exists some bounded constant $C > 0$ such that $\|\mathbf{g}_i^k\| \leq C\alpha_k, \forall i, k$. The next lemma bounds perturbations by step-sizes in an ergodic sense.

Lemma 8. *Let Assumptions A2, A3 hold. Let ϵ be the small constant used in the algorithm, Eq. (3.8), such that $\epsilon \leq \frac{1-\gamma}{2n\Gamma\gamma}$. Define the variable $g_k = \sum_{i=1}^{n} \|\mathbf{g}_i^k\|$. Then there exists some bounded constant $D > 0$ such that for all $K \geq 2$, g_k satisfies:*

$$\sum_{k=0}^{K} g_k \leq D \sum_{k=0}^{K} \alpha_k; \tag{4.9}$$

$$\sum_{k=0}^{K} \alpha_k g_k \leq D \sum_{k=0}^{K} \alpha_k^2, \tag{4.10}$$

where α_k is the diminishing step-size used in the algorithm.

Proof. Based on the result of Lemma 6(b), we have

$$\left\| \mathcal{P}_{\mathcal{X}} \left[\sum_{j=1}^{n} a_{ij} \mathbf{x}_j^k + \epsilon \mathbf{y}_i^k - \alpha_k \nabla \mathbf{f}_i^k \right] - \sum_{j=1}^{n} a_{ij} \mathbf{x}_j^k \right\| \leq \left\| \epsilon \mathbf{y}_i^k - \alpha_k \nabla \mathbf{f}_i^k \right\|. \tag{4.11}$$

Therefore, we obtain

$$\begin{aligned} \left\| \mathbf{g}_i^k \right\| &\leq \left\| \mathbf{x}_i^{k+1} - \sum_{j=1}^{n} a_{ij} \mathbf{x}_j^k \right\| + \epsilon \left\| \mathbf{y}_i^k \right\|, \\ &\leq \left\| \epsilon \mathbf{y}_i^k - \alpha_k \nabla \mathbf{f}_i^k \right\| + \epsilon \left\| \mathbf{y}_i^k \right\|, \\ &\leq B\alpha_k + 2\epsilon \left\| \mathbf{y}_i^k \right\|, \end{aligned} \tag{4.12}$$

where in the last inequality, we use the relation $\|\nabla \mathbf{f}_i^k\| \leq B$. Applying the result of Lemma 7(b) regarding $\|\mathbf{y}_i^k\|$ to the preceding relation, we have for all i,

$$\left\| \mathbf{g}_i^k \right\| \leq B\alpha_k + 2\epsilon\Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \sum_{j=1}^{n} \left\| \mathbf{g}_j^{r-1} \right\|.$$

By defining $g_k = \sum_{i=1}^{n} \|\mathbf{g}_i^k\|$, and summing the above relation over i, it follows that

$$g_k \leq nB\alpha_k + 2n\epsilon\Gamma \sum_{r=1}^{k-1} \gamma^{k-r} g_{r-1}. \tag{4.13}$$

Summing Eq. (4.13) over time from $k = 0$ to K, we obtain

$$\begin{aligned}\sum_{k=0}^{K} g_k &\leq nB \sum_{k=0}^{K} \alpha_k + 2n\epsilon\Gamma \sum_{k=0}^{K} \sum_{r=1}^{k-1} \gamma^{k-r} g_{r-1}, \\ &\leq nB \sum_{k=0}^{K} \alpha_k + 2n\epsilon\Gamma \frac{\gamma(1-\gamma^{K-2})}{1-\gamma} \sum_{k=0}^{K-2} g_k.\end{aligned}$$

Therefore, it satisfies, for any $K \geq 2$, that

$$\left(1 - \frac{2n\epsilon\Gamma\gamma}{1-\gamma}\right) \sum_{k=0}^{K} g_k \leq nB \sum_{k=0}^{K} \alpha_k.$$

Since ϵ can be arbitrary small, (see Lemma 5), it is achievable that $\epsilon \leq \frac{1-\gamma}{2n\Gamma\gamma}$, which obtains the desired result.

Similarly, it can be derived from Eq. (4.13) that

$$\sum_{k=0}^{K} \alpha_k g_k \leq nB \sum_{k=0}^{K} \alpha_k^2 + 2n\epsilon\Gamma \sum_{k=0}^{K} \alpha_k \sum_{r=1}^{k-1} \gamma^{k-r} g_{r-1}.$$

Noticing that the step-size is diminishing, it follows that

$$\begin{aligned}\sum_{k=0}^{K} \alpha_k \sum_{r=1}^{k-1} \gamma^{k-r} g_{r-1} &\leq \sum_{k=0}^{K} \sum_{r=1}^{k-1} \gamma^{k-r} \alpha_{r-1} g_{r-1}, \\ &\leq \frac{\gamma(1-\gamma^{K-2})}{1-\gamma} \sum_{k=0}^{K-2} \alpha_k g_k.\end{aligned}$$

Therefore, it satisfies, for any $K \geq 2$, that

$$\left(1 - \frac{2n\epsilon\Gamma\gamma}{1-\gamma}\right) \sum_{k=0}^{K} \alpha_k g_k \leq nB \sum_{k=0}^{K} \alpha_k^2,$$

which completes the proofs. □

Based on the result of Lemma 8, we show that at kth iteration, the norm of perturbation, $\mathbf{g}_i^k$, of any agent can be bounded by the step-size times some bounded constant.

Lemma 9. *Let Assumptions A2, A3 hold. Let ϵ be the small constant used in the algorithm, Eq. (3.8), such that $\epsilon \leq \frac{1-\gamma}{2n\Gamma\gamma}$. Define the variable $g_k = \sum_{i=1}^n \|\mathbf{g}_i^k\|$. Then there exists some bounded contant $C > 0$ such that for all $k \geq 0$, g_k satisfies:*

$$g_k \leq C\alpha_k; \tag{4.14}$$

where α_k is the diminishing step-size used in the algorithm.

Proof. Suppose on the contrary that $g_k/\alpha_k = \infty$, for some k. Since $\alpha_k \neq 0$, for any finite k, and we get from Lemma 8 that $\sum_{k=0}^{\infty} \alpha_k g_k \leq \sum_{k=0}^{\infty} \alpha_k^2 < \infty$, we obtain that g_k is bounded for any finite k. Therefore, we only get $g_k/\alpha_k = \infty$ when k goes to infinity, i.e., $\lim_{k\to\infty} \frac{g_k}{\alpha_k} = \infty$. This implies that there exists some finite K such that for all $k \geq K$, we have $g_k > 2D\alpha_k$, where D is the constant in the result of Lemma 8. The preceding relation implies that

$$\sum_{k=K}^{\infty} g_k > 2D \sum_{k=K}^{\infty} \alpha_k.$$

Since $\sum_{k=0}^{\infty} \alpha_k = \infty$, we have $\sum_{k=0}^{K-1} \alpha_k < \sum_{k=K}^{\infty} \alpha_k = \infty$. Therefore, we obtain

$$\sum_{k=0}^{\infty} g_k > \sum_{k=K}^{\infty} g_k > 2D \sum_{k=K}^{\infty} \alpha_k > D \sum_{k=0}^{\infty} \alpha_k,$$

which is a contradiction to the result in Lemma 8(a). □

Lemma 9 shows that the perturbation, $\mathbf{g}_i^k$, goes to zero and the D-DPS converges. We next show that the agents reach consensus and also converge to the optimal solution.

Consensus in Estimates

In Lemma 7, we bound the disagreement between estimates of agent and the accumulation state, $\|\mathbf{x}_i^k - \overline{\mathbf{z}}^k\|$, in terms of the perturbation norm, $\sum_{j=1}^{n} \|\mathbf{g}_j^k\|$. In Lemmas 8 and 9, we bound the perturbation. By combining these results, we show the consensus property of the algorithm in the following lemma.

Lemma 10. *Let Assumptions A2, A3 hold. Let* $\{\mathbf{z}_i^k\}$ *be the sequence over* k *generated by Eq.* (4.4)*. Then, for all* $i \in \mathcal{V}$*:*

(a) the agents reach consensus, i.e., $\lim_{k\to\infty} \left\|\mathbf{x}_i^k - \overline{\mathbf{z}}^k\right\| = 0$;

(b) at each agent, $\lim_{k\to\infty} \left\|\mathbf{y}_i^k\right\| = 0$.

Proof. Considering Lemma 7(a), we have for any $K > 0$

$$\begin{aligned}\sum_{k=1}^{K}\alpha_k\left\|\mathbf{x}_i^k-\overline{\mathbf{z}}^k\right\| \leq & \Gamma\sum_{k=1}^{K}\sum_{r=1}^{k-1}\gamma^{k-r}\alpha_k\sum_{j=1}^{n}\left\|\mathbf{g}_j^{r-1}\right\|+\sum_{k=1}^{K}\alpha_k\sum_{j=1}^{n}\left\|\mathbf{g}_j^{k-1}\right\|, \\ \leq & \Gamma C\sum_{k=1}^{K}\sum_{r=1}^{k-1}\gamma^{k-r}\alpha_k\alpha_{r-1}+\sum_{k=1}^{K}\alpha_k\alpha_{k-1}, \\ \leq & \frac{\Gamma C\gamma(1-\gamma^K)}{1-\gamma}\sum_{k=1}^{K}\alpha_k^2+\sum_{k=1}^{K}\alpha_k^2, \end{aligned} \tag{4.15}$$

where we used Lemma 9 to obtain the second inequality. By letting $K \to \infty$ and noticing that $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, we get

$$\sum_{k=1}^{\infty}\alpha_k\left\|\mathbf{x}_i^k-\overline{\mathbf{z}}^k\right\| < \infty. \tag{4.16}$$

Combined with $\sum_{k=0}^{\infty}\alpha_k = \infty$, the preceding relation implies part (a). The result in part (b) follows a similar argument. □

Optimality Convergence

The result of Lemma 10 reveals the fact that all agents reach consensus. We next show that the accumulation state converges to the optimal solution of the problem.

Theorem 2. *Let Assumptions A2, A3 hold. Let* $\left\{\mathbf{z}_i^k\right\}$ *be the sequence over* k *generated by Eq.* (4.4)*. Then, each agent converges to the optimal solution, i.e.,*

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

Proof. Consider Eq. (4.4) and the fact that each column of M sums to one, we have the accumulation state

$$\overline{\mathbf{z}}^{k+1} = \overline{\mathbf{z}}^{k} + \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i^k.$$

Therefore, we obtain that

$$\begin{aligned}
\left\|\overline{\mathbf{z}}^{k+1} - \mathbf{x}^*\right\|^2 &= \left\|\overline{\mathbf{z}}^{k} - \mathbf{x}^*\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i^k\right\|^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \mathbf{g}_i^k\right\rangle, \\
&= \left\|\overline{\mathbf{z}}^{k} - \mathbf{x}^*\right\|^2 + \frac{1}{n^2}\left\|\sum_{i=1}^{n}\mathbf{g}_i^k\right\|^2 - \frac{2\alpha_k}{n}\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \nabla \mathbf{f}_i^k\right\rangle \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \mathbf{g}_i^k + \alpha_k \nabla \mathbf{f}_i^k\right\rangle. \qquad (4.17)
\end{aligned}$$

Since $\|\nabla \mathbf{f}_i^k\| \leq B$, we have

$$\begin{aligned}
\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \nabla \mathbf{f}_i^k\right\rangle &= \left\langle \overline{\mathbf{z}}^k - \mathbf{x}_i^k, \nabla \mathbf{f}_i^k\right\rangle + \left\langle \mathbf{x}_i^k - \mathbf{x}^*, \nabla \mathbf{f}_i^k\right\rangle, \\
&\geq \left\langle \overline{\mathbf{z}}^k - \mathbf{x}_i^k, \nabla \mathbf{f}_i^k\right\rangle + f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}^*), \\
&\geq -B\left\|\overline{\mathbf{z}}^k - \mathbf{x}_i^k\right\| + f_i(\mathbf{x}_i^k) - f_i(\overline{\mathbf{z}}^k) + f_i(\overline{\mathbf{z}}^k) - f_i(\mathbf{x}^*), \\
&\geq -2B\left\|\overline{\mathbf{z}}^k - \mathbf{x}_i^k\right\| + f_i(\overline{\mathbf{z}}^k) - f_i(\mathbf{x}^*). \qquad (4.18)
\end{aligned}$$

By substituting Eq. (4.18) in Eq. (4.17), we obtain that

$$\begin{aligned}
\frac{2\alpha_k}{n}\left(f(\overline{\mathbf{z}}^k) - f^*\right) \leq \left\|\overline{\mathbf{z}}^k - \mathbf{x}^*\right\|^2 - \left\|\overline{\mathbf{z}}^{k+1} - \mathbf{x}^*\right\|^2 + \frac{4B\alpha_k}{n}\sum_{i=1}^{n}\left\|\overline{\mathbf{z}}^k - \mathbf{x}_i^k\right\| \\
+ \frac{1}{n^2}\left\|\sum_{i=1}^{n}\mathbf{g}_i^k\right\|^2 + \frac{2}{n}\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \mathbf{g}_i^k + \alpha_k \nabla \mathbf{f}_i^k\right\rangle. \qquad (4.19)
\end{aligned}$$

We now analyze the last term in Eq. (4.19).

$$\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \mathbf{x}^*, \mathbf{g}_i^k + \alpha_k \nabla \mathbf{f}_i^k\right\rangle = \sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^k - \overline{\mathbf{z}}^{k+1}, \mathbf{g}_i^k + \alpha_k \nabla \mathbf{f}_i^k\right\rangle$$

$$
\begin{aligned}
&+\sum_{i=1}^{n}\left\langle \overline{\mathbf{z}}^{k+1}-\mathbf{x}_i^{k+1},\mathbf{g}_i^k+\alpha_k\nabla\mathbf{f}_i^k\right\rangle\\
&+\sum_{i=1}^{n}\left\langle \mathbf{x}_i^{k+1}-\mathbf{x}^*,\mathbf{g}_i^k+\alpha_k\nabla\mathbf{f}_i^k\right\rangle\\
:=&s_1+s_2+s_3 \qquad (4.20)
\end{aligned}
$$

where s_1, s_2, and s_3 denote each of RHS terms in Eq. (4.20). We discuss each term in sequence. Since $g_k=\sum_{i=1}^{n}\|\mathbf{g}_i^k\|\le C\alpha_k$ and $\|\nabla\mathbf{f}_i^k\|\le B$, we have

$$
\begin{aligned}
s_1&=-\sum_{i=1}^{n}\left\langle \mathbf{g}_i^k,\mathbf{g}_i^k+\alpha_k\nabla\mathbf{f}_i^k\right\rangle\le B\alpha_k\sum_{i=1}^{n}\left\|\mathbf{g}_i^k\right\|=BC\alpha_k^2;\\
s_2&\le(B+C)\alpha_k\sum_{i=1}^{n}\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}_i^{k+1}\right\|.
\end{aligned}
$$

Using the result of Lemma 6(a), we have for any i

$$
\left\langle \mathbf{x}_i^{k+1}-\mathbf{x}^*,\mathbf{g}_i^k+\alpha_k\nabla\mathbf{f}_i^k\right\rangle\le 0,
$$

which reveals that $s_3\le 0$. Using the upperbound of s_1, s_2, and s_3 in the preceding relations and the fact that $g_k=\sum_{i=1}^{n}\|\mathbf{g}_i^k\|\le C\alpha_k$, we derive from Eq. (4.19) that

$$
\begin{aligned}
\frac{2\alpha_k}{n}\left(f(\overline{\mathbf{z}}^k)-f^*\right)\le&\left\|\overline{\mathbf{z}}^k-\mathbf{x}^*\right\|^2-\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}^*\right\|^2+\frac{4B\alpha_k}{n}\sum_{i=1}^{n}\left\|\overline{\mathbf{z}}^k-\mathbf{x}_i^k\right\|+\frac{C^2}{n^2}\alpha_k^2\\
&+\frac{2BC}{n}\alpha_k^2+\frac{2(B+C)}{n}\alpha_k\sum_{i=1}^{n}\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}_i^{k+1}\right\|.
\end{aligned}
$$

By summing the preceding relation over k, we have that

$$
\sum_{k=1}^{\infty}\frac{2\alpha_k}{n}\left(f(\overline{\mathbf{z}}^k)-f^*\right)\le\left\|\overline{\mathbf{z}}^1-\mathbf{x}^*\right\|^2+\left(\frac{C^2}{n^2}+\frac{2BC}{n}\right)\sum_{k=1}^{\infty}\alpha_k^2
$$

$$+\frac{4B}{n}\sum_{i=1}^{n}\sum_{k=1}^{\infty}\alpha_k\left\|\overline{\mathbf{z}}^k-\mathbf{x}_i^k\right\|+\frac{2(B+C)}{n}\sum_{i=1}^{n}\sum_{k=1}^{\infty}\alpha_k\left\|\overline{\mathbf{z}}^{k+1}-\mathbf{x}_i^{k+1}\right\|. \quad (4.21)$$

Since that the step-size follows $\sum_{k=1}^{\infty}\alpha_k^2<\infty$ and $\sum_{k=1}^{\infty}\alpha_k\|\overline{\mathbf{z}}^k-\mathbf{x}_i^k\|<\infty$, from Eq. (4.16), we obtain that

$$\sum_{k=1}^{\infty}\frac{2\alpha_k}{n}\left(f(\overline{\mathbf{z}}^k)-f^*\right)<\infty, \quad (4.22)$$

which reveals that $\lim_{k\to\infty}f(\overline{\mathbf{z}}^k)=f^*$ as $\sum_{k=1}^{\infty}\alpha_k=\infty$; the proof follows from Lemma 10. □

Convergence Rate

We now characterize the convergence rate with $\alpha_k=\frac{1}{k^a}$, and $a>0$. Let $f_K^*:=\min_{0<k\leq K}f(\overline{\mathbf{z}}^k)$, we have

$$(f_K^*-f^*)\sum_{k=1}^{K}\alpha_k\leq\sum_{k=1}^{K}\alpha_k(f(\overline{\mathbf{z}}^k)-f^*). \quad (4.23)$$

By combining Eqs. (4.15), (4.21) and (4.23), Eq. (4.21) leads to

$$(f_K^*-f^*)\sum_{k=1}^{K}\alpha_k\leq C_1+C_2\sum_{k=1}^{K}\alpha_k^2,$$

or equivalently,

$$(f_K^*-f^*)\leq\frac{C_1}{\sum_{k=1}^{K}\alpha_k}+\frac{C_2\sum_{k=1}^{K}\alpha_k^2}{\sum_{k=1}^{K}\alpha_k}, \quad (4.24)$$

where the constants, C_1 and C_2, are given by

$$C_1=\frac{n}{2}\left\|\overline{\mathbf{z}}^0-\mathbf{x}^*\right\|^2, C_2=\frac{C^2}{2n}+BC+(3B+C)\left(\frac{\Gamma C\gamma}{1-\gamma}+1\right).$$

Assume the diminishing step-size, $\alpha_k = \frac{1}{k^a}$, with $a > 0$.

(i) When $0 < a < \frac{1}{2}$, the first term in Eq. (4.24) leads to

$$\frac{C_1}{\sum_{k=1}^{K} \alpha_k} < C_1 \frac{1-a}{K^{1-a}-1} = O\left(\frac{1}{K^{1-a}}\right),$$

while the second term in Eq. (4.24) leads to

$$\frac{C_2 \sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} < C_2 \frac{(1-a)(K^{1-2a}-2a)}{(1-2a)(K^{1-a}-1)} = O\left(\frac{1}{K^a}\right).$$

Considering that $0 < a < \frac{1}{2}$, we have $O\left(\frac{1}{K^a}\right)$ dominates since it decreases slower than $O\left(\frac{1}{K^{1-a}}\right)$.

(ii) When $\alpha_k = k^{-1/2}$, the first term in Eq. (4.24) leads to

$$\frac{C_1}{\sum_{k=1}^{K} \alpha_k} < C_1 \frac{1/2}{K^{1/2}-1} = O\left(\frac{1}{\sqrt{K}}\right),$$

while the second term in Eq. (4.24) leads to

$$\frac{C_2 \sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} < C_2 \frac{1+\ln K}{2(\sqrt{K}-1)} = O\left(\frac{\ln K}{\sqrt{K}}\right).$$

It can be observed that $O\left(\frac{\ln K}{\sqrt{K}}\right)$ dominates.

(iii) When $\frac{1}{2} < a < 1$, the first term in Eq. (4.24) leads to

$$\frac{C_1}{\sum_{k=1}^{K} \alpha_k} < C_1 \frac{1-a}{K^{1-a}-1} = O\left(\frac{1}{K^{1-a}}\right),$$

while the second term in Eq. (4.24) leads to

$$\frac{C_2 \sum_{k=1}^{K} \alpha_k^2}{\sum_{k=1}^{K} \alpha_k} < C_2 \frac{(1-a)(2a-1/K^{2a-1})}{(2a-1)(K^{1-a}-1)} = O\left(\frac{1}{K^{1-a}}\right).$$

The two terms are in the same order.

(iv) When $a > 1$, the two terms in Eq. (4.24) approach constant values. Therefore, the persistence conditions of step-size are not satisfied, and convergence of D-DPS is not satisfied.

By comparing (i), (ii), and (iii), we have that $O(\frac{\ln K}{\sqrt{K}})$ is the fastest. In conclusion, the optimal convergence rate is achieved by choosing $\alpha_k = \frac{1}{\sqrt{k}}$, and the corresponding convergence rate of D-DPS is $O(\frac{\ln k}{\sqrt{k}})$. This convergence rate is the same as the distributed projected subgradient method, [38], solving constrained optimization over undirected graphs. Therefore, the restriction of directed graphs does not effect the convergence speed.

4.3 Numerical Results

Consider the application of D-DPS for solving a distributed logistic regression problem over a directed graph:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{X}\subset\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left[1+\exp\left(-\left(\mathbf{c}_{ij}^\top \mathbf{x}\right) y_{ij}\right)\right],$$

where $\mathcal{X}$ is a small convex set restricting the value of $\mathbf{x}$ to avoid overfitting. Each agent i has access to m_i training samples, $(\mathbf{c}_{ij}, y_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ includes the p features of the jth training sample of agent i, and y_{ij} is the corresponding label. This problem can be formulated in the form of P2

with the private objective function f_i being

$$f_i(\mathbf{x}) = \sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{c}_{ij}^\top \mathbf{x}\right) y_{ij}\right)\right], \quad \text{s.t. } \mathbf{x} \in \mathcal{X}.$$

In our setting, we have $n = 10$, $m_i = 10$, for all i, and $p = 100$. The constrained set is described by a ball in $\mathbb{R}^p$. We consider the network topology as the digraph shown in Fig. 4.1. We plot the residuals $\frac{\|\mathbf{x}_i^k - \mathbf{x}^*\|_F}{\|\mathbf{x}_i^0 - \mathbf{x}^*\|_F}$ for each

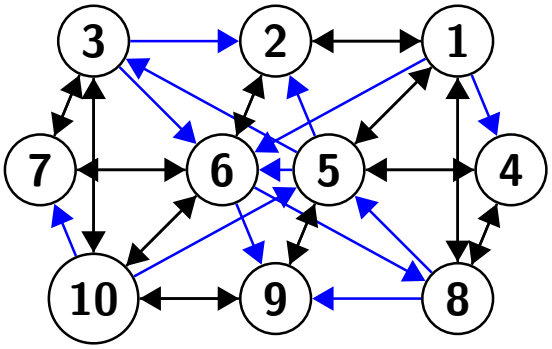


Figure 4.1: A strongly-connected but non-balanced directed graph.

agent i as a function of k in Fig. 4.2. In Fig. 4.3, we show the disagreement between the state estimate of each agent and the accumulation state, and the additional variables of all agents. The experiment follows the results of Lemma 10 that both the disagreements and the additional variables converge to zero. We compare the convergence of D-DPS with others related algorithms, Subgradient-Push (SP), [61], and WeightBalencing Subgradient Descent (WBSD), [74], in Fig. 4.4. Since both SP and WBSD are algorithms for unconstrained problems, we reformulate the problem in an approximate

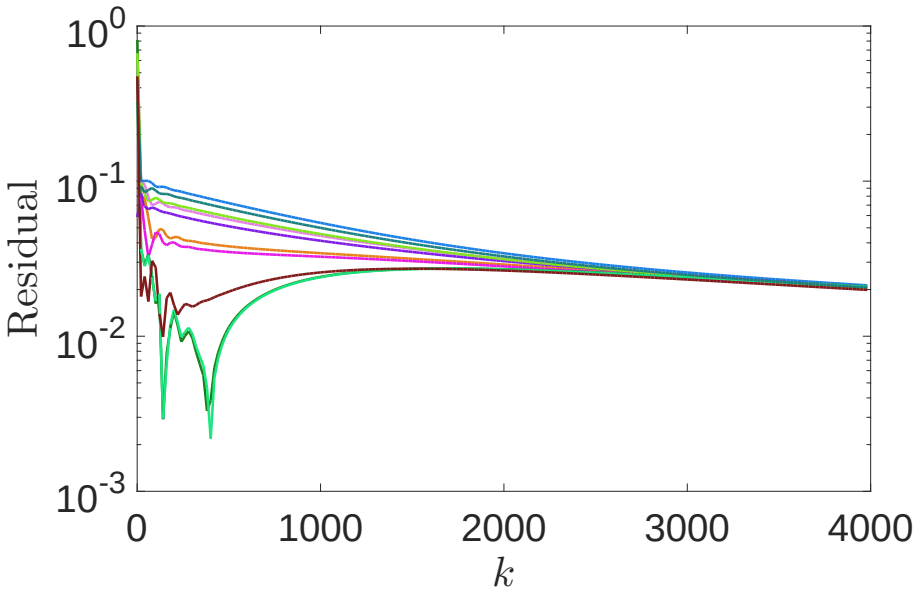


Figure 4.2: D-DPS residuals at 10 agents.

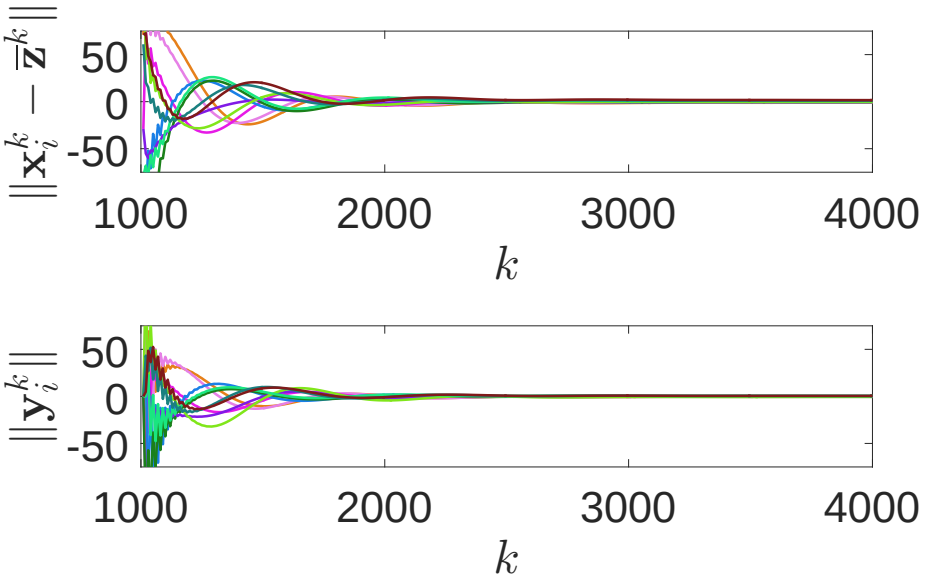


Figure 4.3: Sample paths of states, $\|\mathbf{x}_i^k - \overline{\mathbf{z}}^k\|$, and $\|\mathbf{y}_i^k\|$, for all agents.

form,

$$f_i(\mathbf{x}) = \lambda\|\mathbf{x}\|^2 + \sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{c}_{ij}^\top \mathbf{x}\right) y_{ij}\right)\right],$$

where the regularization term $\lambda\|\mathbf{x}\|^2$ is an approximation to replace the original constrained set to avoid overfitting. It can be observed from Fig. 4.4 that all three algorithms have the same order of convergence rate. However, D-DPS is further suited for the constrained problems.

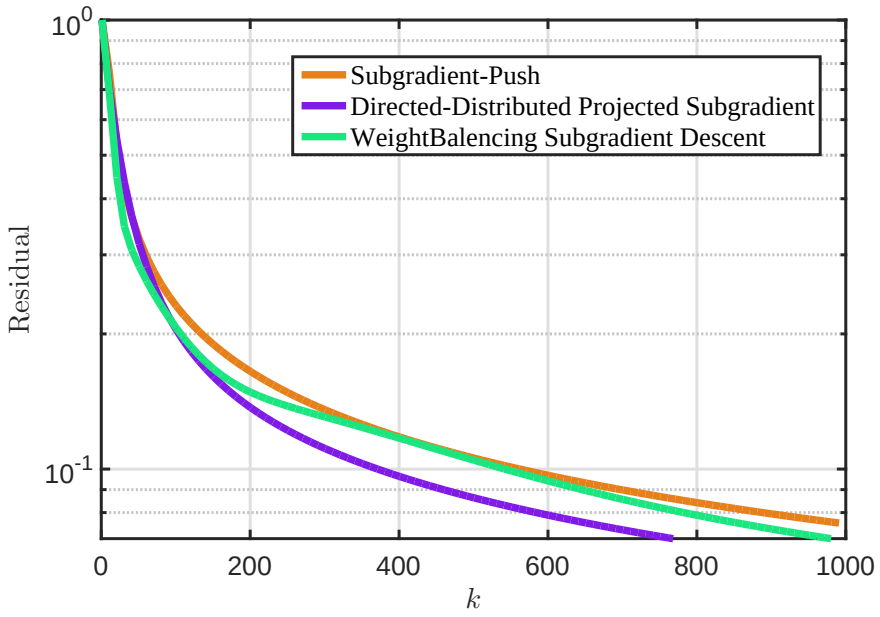


Figure 4.4: Convergence comparison between different algorithms.

4.4 Conclusions

In this Chapter, we present a distributed solution, D-DPS, to the *constrained* optimization problem over *directed* multi-agent networks, where the agents' goal is to collectively minimize the sum of locally known convex functions. D-DSD, proposed in Chapter III, can be viewed as a special case of D-DPS when the constrained set is $\mathbb{R}^p$. Same as D-DSD, D-DPS converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions in Problem P2 are convex, but not necessarily differentiable. Compared to the algorithm solving over undirected networks, the D-DPS simultaneously constructs a row-stochastic matrix and a column-stochastic matrix instead of only a doubly-stochastic matrix. This enables all agents to overcome the asymmetry caused by the directed communication network. We show that D-DPS converges to the optimal solution and the convergence rate is $O(\frac{\log k}{\sqrt{k}})$, where k is the number of iterations.

Chapter 5

DEXTRA for Smooth Convex Optimization

In this chapter, we introduce DEXTRA, to solve the distributed optimization problem, P1, over directed networks. Recall D-DSD and D-DPS, which achieve a sub-linear convergence rate to solve P1. We harness the function smoothness to accelerate the convergence rate. Therefore, DEXTRA converges to the optimal solution in smooth convex optimization, i.e., the local objective functions in Problem P1 are convex and differentiable. We show that, with the appropriate step-size, DEXTRA converges at a linear rate $O(\tau^k)$ for $0 < \tau < 1$, given that the objective functions are restricted strongly-convex. The implementation of DEXTRA requires each agent to know its local out-degree. Simulation examples further illustrate our findings.

5.1 Algorithm

To solve the Problem P1 suited to the case of directed graphs, we propose DEXTRA that can be described as follows. Each agent, $j \in \mathcal{V}$, maintains two vector variables: $\mathbf{x}_j^k,\ \mathbf{z}_j^k \in \mathbb{R}^p$, as well as a scalar variable, $y_j^k \in \mathbb{R}$, where k is the discrete-time index. At the kth iteration, agent j weights its states, $a_{ij}\mathbf{x}_j^k$, $a_{ij}y_j^k$, as well as $\widetilde{a}_{ij}\mathbf{x}_j^{k-1}$, and sends these to each of its out-neighbors, $i \in \mathcal{N}_j^{\text{out}}$, where the weights, a_{ij}, and, $\widetilde{a}_{ij}$,'s are such that:

$$a_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{i=1}^{n} a_{ij} = 1, \forall j, \tag{5.1}$$

$$\widetilde{a}_{ij} = \begin{cases} \theta + (1-\theta)a_{ij}, & i = j, \\ (1-\theta)a_{ij}, & i \neq j, \end{cases} \qquad \forall j, \tag{5.2}$$

where $\theta \in (0, \frac{1}{2}]$. With agent i receiving the information from its in-neighbors, $j \in \mathcal{N}_i^{\text{in}}$, it calculates the state, $\mathbf{z}_i^k$, by dividing $\mathbf{x}_i^k$ over y_i^k, and updates $\mathbf{x}_i^{k+1}$ and y_i^{k+1} as follows:

$$\mathbf{z}_i^k = \frac{\mathbf{x}_i^k}{y_i^k}, \tag{5.3a}$$

$$\begin{aligned} \mathbf{x}_i^{k+1} =& \mathbf{x}_i^k + \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij}\mathbf{x}_j^k\right) - \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(\widetilde{a}_{ij}\mathbf{x}_j^{k-1}\right) \\ & - \alpha \left[\nabla f_i(\mathbf{z}_i^k) - \nabla f_i(\mathbf{z}_i^{k-1})\right], \end{aligned} \tag{5.3b}$$

$$y_i^{k+1} = \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij}y_j^k\right). \tag{5.3c}$$

In the above, $\nabla f_i(\mathbf{z}_i^k)$ is the gradient of the function $f_i(\mathbf{z})$ at $\mathbf{z} = \mathbf{z}_i^k$, and $\nabla f_i(\mathbf{z}_i^{k-1})$ is the gradient at $\mathbf{z}_i^{k-1}$, respectively. The method is initiated with an arbitrary vector, $\mathbf{x}_i^0$, and with $y_i^0 = 1$ for any agent i. The step-size, α, is a positive number within a certain interval. We will explicitly discuss the range of α later. We adopt the convention that $\mathbf{x}_i^{-1} = \mathbf{0}_p$ and $\nabla f_i(\mathbf{z}_i^{-1}) = \mathbf{0}_p$, for any agent i, such that at the first iteration, i.e., $k = 0$, we have the following iteration instead of Eq. (5.3),

$$\mathbf{z}_i^0 = \frac{\mathbf{x}_i^0}{y_i^0}, \tag{5.4a}$$

$$\mathbf{x}_i^1 = \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij}\mathbf{x}_j^0\right) - \alpha \nabla f_i(\mathbf{z}_i^0), \tag{5.4b}$$

$$y_i^1 = \sum_{j \in \mathcal{N}_i^{\text{in}}} \left(a_{ij} y_j^0\right). \tag{5.4c}$$

We note that the implementation of Eq. (5.3) needs each agent to have the knowledge of its out-neighbors (such that it can design the weights according to Eqs. (5.1) and (5.2)). In a more restricted setting, e.g., a broadcast application where it may not be possible to know the out-neighbors, we may use $a_{ij} = |\mathcal{N}_j^{\text{out}}|^{-1}$; thus, the implementation only requires each agent to know its out-degree, [61, 63–65, 74, 79, 82].

To simplify the analysis, we assume from now on that all sequences updated by Eq. (5.3) have only one dimension, i.e., $p = 1$; thus x_i^k, y_i^k, $z_i^k \in \mathbb{R}, \forall i, k$. For $\mathbf{x}_i^k$, $\mathbf{z}_i^k \in \mathbb{R}^p$ being p-dimensional vectors, the proof is the same for ev-

ery dimension by applying the results to each coordinate. Therefore, assuming $p = 1$ is without the loss of generality. We next write DEXTRA, Eq. (5.3), in a matrix form. Let, $A = \{a_{ij}\} \in \mathbb{R}^{n\times n}$, $\widetilde{A} = \{\widetilde{a}_{ij}\} \in \mathbb{R}^{n\times n}$, be the collection of weights, a_{ij}, $\widetilde{a}_{ij}$, respectively. It is clear that both A and $\widetilde{A}$ are column-stochastic matrices. Let $\mathbf{x}^k$, $\mathbf{z}^k$, $\nabla\mathbf{f}(\mathbf{x}^k) \in \mathbb{R}^{np}$, be the collection of all agent states and gradients at time k, i.e., $\mathbf{x}^k \triangleq [x_1^k; \cdots; x_n^k]$, $\mathbf{z}^k \triangleq [z_1^k; \cdots; z_n^k]$, $\nabla\mathbf{f}(\mathbf{x}^k) \triangleq [\nabla f_1(x_1^k); \cdots; \nabla f_n(x_n^k)]$, and $\mathbf{y}^k \in \mathbb{R}^n$ be the collection of agent states, y_i^k, i.e., $\mathbf{y}^k \triangleq [y_1^k; \cdots; y_n^k]$. Note that at time k, $\mathbf{y}^k$ can be represented by the initial value, $\mathbf{y}^0$:

$$\mathbf{y}^k = A\mathbf{y}^{k-1} = A^k\mathbf{y}^0 = A^k \cdot \mathbf{1}_n. \tag{5.5}$$

Define a diagonal matrix, $D^k \in \mathbb{R}^{n\times n}$, for each k, such that the ith element of D^k is y_i^k, i.e.,

$$D^k = \mathrm{diag}\left(\mathbf{y}^k\right) = \mathrm{diag}\left(A^k \cdot \mathbf{1}_n\right). \tag{5.6}$$

Given that the graph, $\mathcal{G}$, is strongly-connected and the corresponding weighting matrix, A, is non-negative, it follows that D^k is invertible for any k. Then, we can write Eq. (5.3) in the matrix form equivalently as follows:

$$\mathbf{z}^k = \left[D^k\right]^{-1}\mathbf{x}^k, \tag{5.7a}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + A\mathbf{x}^k - \widetilde{A}\mathbf{x}^{k-1} - \alpha\left[\nabla\mathbf{f}(\mathbf{z}^k) - \nabla\mathbf{f}(\mathbf{z}^{k-1})\right], \tag{5.7b}$$

$$\mathbf{y}^{k+1} = A\mathbf{y}^k, \tag{5.7c}$$

where both of the weight matrices, A and $\widetilde{A}$, are column-stochastic and satisfy the relationship: $\widetilde{A} = \theta I_n + (1-\theta)A$ with some $\theta \in (0, \frac{1}{2}]$. From Eq. (5.7a), we obtain for any k

$$\mathbf{x}^k = D^k \mathbf{z}^k. \tag{5.8}$$

Therefore, Eq. (5.7) can be represented as a single equation:

$$D^{k+1}\mathbf{z}^{k+1} = (I_n + A)D^k\mathbf{z}^k - \widetilde{A}D^{k-1}\mathbf{z}^{k-1} - \alpha\left[\nabla \mathbf{f}(\mathbf{z}^k) - \nabla \mathbf{f}(\mathbf{z}^{k-1})\right]. \tag{5.9}$$

We refer to the above algorithm as DEXTRA, since Eq. (5.9) has a similar form as EXTRA in Eq. (2.8) and is designed to solve Problem P1 in the case of directed graphs. We later shows that as time goes to infinity, the iteration in Eq. (5.9) pushes $\mathbf{z}^k$ to achieve consensus and reach the optimal solution in a linear rate. Our proof in this paper will based on the form, Eq. (5.9), of DEXTRA.

Interpretation of DEXTRA

In this section, we give an intuitive interpretation on DEXTRA's convergence to the optimal solution. Since A is column-stochastic, the sequence, $\{\mathbf{y}^k\}$, generated by Eq. (5.7c), satisfies $\lim_{k\to\infty} \mathbf{y}^k = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is some vector in the span of A's right-eigenvector corresponding to the eigenvalue 1. We also obtain that $D^\infty = \mathrm{diag}\,(\boldsymbol{\pi})$. For the sake of argument, let us assume that the

sequences, $\{\mathbf{z}^k\}$ and $\{\mathbf{x}^k\}$, generated by DEXTRA, Eq. (5.7) or (5.9), also converge to their own limits, $\mathbf{z}^\infty$ and $\mathbf{x}^\infty$, respectively (not necessarily true). According to the updating rule in Eq. (5.7b), the limit $\mathbf{x}^\infty$ satisfies

$$\mathbf{x}^\infty = \mathbf{x}^\infty + A\mathbf{x}^\infty - \widetilde{A}\mathbf{x}^\infty - \alpha\left[\nabla\mathbf{f}(\mathbf{z}^\infty) - \nabla\mathbf{f}(\mathbf{z}^\infty)\right], \tag{5.10}$$

which implies that $(A-\widetilde{A})\mathbf{x}^\infty = \mathbf{0}_n$, or $\mathbf{x}^\infty = u\boldsymbol{\pi}$ for some scalar, u. It follows from Eq. (5.7a) that

$$\mathbf{z}^\infty = u\left[D^\infty\right]^{-1}\boldsymbol{\pi} = u\mathbf{1}_n, \tag{5.11}$$

where the consensus is reached. The above analysis reveals the idea of DEXTRA, which is to overcome the imbalance of agent states occurred when the graph is directed: both $\mathbf{x}^\infty$ and $\mathbf{y}^\infty$ lie in the span of $\boldsymbol{\pi}$; by dividing $\mathbf{x}^\infty$ over $\mathbf{y}^\infty$, the imbalance is canceled.

Summing up the updates in Eq. (5.7b) over k from 0 to ∞, we obtain that

$$\mathbf{x}^\infty = A\mathbf{x}^\infty - \alpha\nabla\mathbf{f}(\mathbf{z}^\infty) - \sum_{r=0}^{\infty}\left(\widetilde{A}-A\right)\mathbf{x}^r;$$

note that the first iteration is slightly different as shown in Eqs. (5.4). Consider $\mathbf{x}^\infty = u\boldsymbol{\pi}$ and the preceding relation. It follows that the limit, $\mathbf{z}^\infty$, satisfies

$$\alpha\nabla\mathbf{f}(\mathbf{z}^\infty) = -\sum_{r=0}^{\infty}\left(\widetilde{A}-A\right)\mathbf{x}^r. \tag{5.12}$$

Therefore, we obtain that

$$\alpha \mathbf{1}_n^\top \nabla \mathbf{f}(\mathbf{z}^\infty) = -\mathbf{1}_n^\top \left(\widetilde{A} - A\right) \sum_{r=0}^{\infty} \mathbf{x}^r = 0,$$

which is the optimality condition of Problem P1 considering that $\mathbf{z}^\infty = u\mathbf{1}_n$. Therefore, given the assumption that the sequence of DEXTRA iterates, $\left\{\mathbf{z}^k\right\}$ and $\left\{\mathbf{x}^k\right\}$, have limits, $\mathbf{z}^\infty$ and $\mathbf{x}^\infty$, we have the fact that $\mathbf{z}^\infty$ achieves consensus and reaches the optimal solution of Problem P1. In the next section, we state the convergence result of DEXTRA.

5.2 Assumptions and Main Results

Recall D-DSD and D-DPS in previous chapters, which achieve a sub-linear convergence rate to solve P1. We harness the function smoothness to accelerate the convergence rate. In particular, we modify the bounded gradient assumption to a Lipschitz continuous gradient assumption. In other words, DEXTRA converges to the optimal solution in smooth convex optimization, i.e., the local objective functions in Problem P1 are convex and differentiable. With appropriate assumptions, our main result states that DEXTRA converges to the optimal solution of Problem P1 linearly. We state again that from now on we assume that the states of agents have only one dimension, i.e., $p = 1$, which is without the loss of generality. We assume that the agent graph, $\mathcal{G}$, is strongly-connected; each local function, $f_i : \mathbb{R}^p \to \mathbb{R}$, is convex and

differentiable, and the optimal solution of Problem P1 and the corresponding optimal value exist. Formally, we denote the optimal solution by $u \in \mathbb{R}$ and optimal value by f^*, i.e.,

$$f^* = f(u) = \min_{x \in \mathbb{R}} f(x). \tag{5.13}$$

Let $\mathbf{z}^* \in \mathbb{R}^n$ be defined as

$$\mathbf{z}^* = u\mathbf{1}_n. \tag{5.14}$$

Besides the above assumptions, we emphasize some other assumptions regarding the objective functions and weighting matrices, which are formally presented as follows.

Assumption A4 (Functions and Gradients)**.** *Each private function, f_i, is convex and differentiable and satisfies the following assumptions.*

(a) The function, f_i, has Lipschitz gradient with the constant L_{f_i}, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_{f_i}\|x - y\|$, $\forall x, y \in \mathbb{R}$.

(b) The function, f_i, is restricted strongly-convex[1] with respect to point u with a positive constant S_{f_i}, i.e., $S_{f_i}\|x-u\|^2 \leq \langle \nabla f_i(x) - \nabla f_i(u), x-u\rangle$, $\forall x \in \mathbb{R}$, where u is the optimal solution of the Problem P1.

[1]There are different definitions of restricted strong-convexity. We use the same as the one used in EXTRA, [48].

Following Assumption A4, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|, \tag{5.15a}$$

$$S_f \|\mathbf{x} - \mathbf{z}^*\|^2 \leq \langle \nabla \mathbf{f}(\mathbf{x}) - \nabla \mathbf{f}(\mathbf{z}^*), \mathbf{x} - \mathbf{z}^* \rangle, \tag{5.15b}$$

where $\nabla \mathbf{f}(\mathbf{x}) \triangleq [\nabla f_1(x_1); \cdots; \nabla f_n(x_n)]$ for any $\mathbf{x} \triangleq [x_1; \cdots; x_n]$, and the constants $L_f = \max_i\{L_{f_i}\}$, $S_f = \min_i\{S_{f_i}\}$.

Recall the definition of D^k in Eq. (5.6), we formally denote the limit of D^k by D^∞, i.e.,

$$D^\infty = \lim_{k\to\infty} D^k = \operatorname{diag}(A^\infty \cdot \mathbf{1}_n) = \operatorname{diag}(\boldsymbol{\pi}), \tag{5.16}$$

where $\boldsymbol{\pi}$ is some vector in the span of the right-eigenvector of A corresponding to eigenvalue 1. The next assumption is related to the weighting matrices, A, $\widetilde{A}$, and D^∞.

Assumption A5 (Weighting matrices)**.** *The weighting matrices, A and $\widetilde{A}$, used in DEXTRA, Eq.* (5.7) *or* (5.9)*, satisfy the following.*

(a) A is a column-stochastic matrix.

(b) $\widetilde{A}$ is a column-stochastic matrix and satisfies $\widetilde{A} = \theta I_n + (1-\theta)A$, for some $\theta \in (0, \frac{1}{2}]$.

(c) $(D^\infty)^{-1}\widetilde{A} + \widetilde{A}^\top (D^\infty)^{-1} \succ 0$.

One way to guarantee Assumption A5(c) is to design the weighting matrix, $\widetilde{A}$, to be *diagonally-dominant*. For example, each agent j designs the following weights:

$$a_{ij} = \begin{cases} 1 - \zeta(|\mathcal{N}_j^{\text{out}}| - 1), & i = j, \\ \zeta, & i \neq j, \quad i \in \mathcal{N}_j^{\text{out}}, \end{cases},$$

where ζ is some small positive constant close to zero. This weighting strategy guarantees the Assumption A5(c) as we explain in the following. According to the definition of D^{∞} in Eq. (5.16), all eigenvalues of the matrix, $2(D^{\infty})^{-1} = (D^{\infty})^{-1} I_n + I_n^{\top}(D^{\infty})^{-1}$, are greater than zero. Since eigenvalues are a continuous functions of the corresponding matrix elements, [30, 89], there must exist a small constant $\overline{\zeta}$ such that for all $\zeta \in (0, \overline{\zeta})$ the weighting matrix, $\widetilde{A}$, designed by the constant weighting strategy with parameter ζ, satisfies that all the eigenvalues of the matrix, $(D^{\infty})^{-1}\widetilde{A} + \widetilde{A}^{\top}(D^{\infty})^{-1}$, are greater than zero.

Since the weighting matrices, A and, $\widetilde{A}$, are designed to be column-stochastic, they satisfy the following.

Lemma 11. *(Nedic et al. [61]) For any column-stochastic matrix $A \in \mathbb{R}^{n \times n}$, we have*

(a) The limit $\lim_{k\to\infty} \left[A^k\right]$ exists and $\lim_{k\to\infty} \left[A^k\right]_{ij} = \boldsymbol{\pi}_i$, where $\boldsymbol{\pi} = \{\boldsymbol{\pi}_i\}$ is some vector in the span of the right-eigenvector of A corresponding to eigenvalue 1.

(b) For all $i \in \{1, \cdots, n\}$*, the entries* $\left[A^k\right]_{ij}$ *and* $\boldsymbol{\pi}_i$ *satisfy*

$$\left|\left[A^k\right]_{ij} - \boldsymbol{\pi}_i\right| < C\gamma^k, \qquad \forall j,$$

where we can have $C = 4$ *and* $\gamma = (1 - \frac{1}{n^n})$.

As a result, we obtain that for any k,

$$\left\|D^k - D^\infty\right\| \leq nC\gamma^k. \tag{5.17}$$

Eq. (5.17) implies that different agents reach consensus in a linear rate with the constant γ. Clearly, the convergence rate of DEXTRA will not exceed this consensus rate (because the convergence of DEXTRA means both consensus and optimality are achieved). We will show this fact theoretically later in this section. We now denote some notations to simplify the representation in the rest of the paper. Define the following matrices,

$$M = (D^\infty)^{-1}\widetilde{A}, \tag{5.18}$$

$$N = (D^\infty)^{-1}(\widetilde{A} - A), \tag{5.19}$$

$$Q = (D^\infty)^{-1}(I_n + A - 2\widetilde{A}), \tag{5.20}$$

$$P = I_n - A, \tag{5.21}$$

$$L = \widetilde{A} - A, \tag{5.22}$$

$$R = I_n + A - 2\widetilde{A}, \tag{5.23}$$

and constants,

$$d = \max_k \left\{ \|D^k\| \right\}, \tag{5.24}$$

$$d^- = \max_k \left\{ \|(D^k)^{-1}\| \right\}, \tag{5.25}$$

$$d^-_\infty = \|(D^\infty)^{-1}\|. \tag{5.26}$$

We also define some auxiliary variables and sequences. Let $\mathbf{q}^* \in \mathbb{R}^n$ be some vector satisfying

$$L\mathbf{q}^* + \alpha \nabla \mathbf{f}(\mathbf{z}^*) = \mathbf{0}_n; \tag{5.27}$$

and $\mathbf{q}^k$ be the accumulation of $\mathbf{x}^r$ over time:

$$\mathbf{q}^k = \sum_{r=0}^{k} \mathbf{x}^r. \tag{5.28}$$

Based on M, N, D^k, $\mathbf{z}^k$, $\mathbf{z}^*$, $\mathbf{q}^k$, and $\mathbf{q}^*$, we further define

$$G = \begin{bmatrix} M^\top & \\ & N \end{bmatrix}, \mathbf{t}^k = \begin{bmatrix} D^k \mathbf{z}^k \\ \mathbf{q}^k \end{bmatrix}, \mathbf{t}^* = \begin{bmatrix} D^\infty \mathbf{z}^* \\ \mathbf{q}^* \end{bmatrix}. \tag{5.29}$$

It is useful to note that the G-matrix norm, $\|\mathbf{a}\|^2_G$, of any vector, $\mathbf{a} \in \mathbb{R}^{2n}$, is non-negative, i.e., $\|\mathbf{a}\|^2_G \geq 0$, $\forall \mathbf{a}$. This is because $G + G^\top$ is PSD as can be shown with the help of the following lemma.

Lemma 12. *(Chung. [90]) Let $\mathcal{L}_\mathcal{G}$ denote the Laplacian matrix of a directed graph, $\mathcal{G}$. Let U be a transition probability matrix associated to a Markov chain*

described on $\mathcal{G}$ *and* $\boldsymbol{s}$ *be the left-eigenvector of* U *corresponding to eigenvalue* 1. *Then,*

$$\mathcal{L}_{\mathcal{G}} = I_n - \frac{S^{1/2}US^{-1/2} + S^{-1/2}U^\top S^{1/2}}{2},$$

where $S = diag(\boldsymbol{s})$. *Additionally, if* $\mathcal{G}$ *is strongly-connected, then the eigenvalues of* $\mathcal{L}_{\mathcal{G}}$ *satisfy* $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_n$.

Considering the underlying directed graph, $\mathcal{G}$, and let the weighting matrix A, used in DEXTRA, be the corresponding transition probability matrix, we obtain that

$$\mathcal{L}_{\mathcal{G}} = \frac{(D^\infty)^{1/2}(I_n - A^\top)(D^\infty)^{-1/2}}{2} + \frac{(D^\infty)^{-1/2}(I_n - A)(D^\infty)^{1/2}}{2}. \tag{5.30}$$

Therefore, we have the matrix N, defined in Eq. (5.19), satisfy

$$N + N^\top = 2\theta \left(D^\infty\right)^{-1/2} \mathcal{L}_{\mathcal{G}} \left(D^\infty\right)^{-1/2}, \tag{5.31}$$

where θ is the positive constant in Assumption A5(b). Clearly, $N + N^\top$ is PSD as it is a product of PSD matrices and a non-negative scalar. Additionally, from Assumption A5(c), note that $M + M^\top$ is PD and thus for any $\mathbf{a} \in \mathbb{R}^n$, it also follows that $\|\mathbf{a}\|^2_{M^\top} \geq 0$. Therefore, we conclude that $G + G^\top$ is PSD and for any $\mathbf{a} \in \mathbb{R}^{2n}$,

$$\|\mathbf{a}\|^2_G \geq 0. \tag{5.32}$$

We now state the main result of this paper in Theorem 3.

Theorem 3. *Define*

$$\begin{aligned}
C_1 &= d^- \left(d \left\| (I_n + A) \right\| + d \left\| \widetilde{A} \right\| + 2\alpha L_f \right), \\
C_2 &= \frac{\left(\lambda_{\max} \left(N N^\top \right) + \lambda_{\max} \left(N + N^\top \right) \right)}{2 \widetilde{\lambda}_{\min} \left(L^\top L \right)}, \\
C_3 &= \alpha (nC)^2 \left[\frac{C_1^2}{2\eta} + (d_\infty^- d^- L_f)^2 \left(\eta + \frac{1}{\eta} \right) + \frac{S_f}{d^2} \right], \\
C_4 &= 8 C_2 \left(L_f d^- \right)^2, \\
C_5 &= \lambda_{\max} \left(\frac{M + M^\top}{2} \right) + 4 C_2 \lambda_{\max} \left(R^\top R \right), \\
C_6 &= \frac{\frac{S_f}{d^2} - \eta - 2\eta (d_\infty^- d^- L_f)^2}{2}, \\
C_7 &= \frac{1}{2} \lambda_{\max} \left(M M^\top \right) + 4 C_2 \lambda_{\max} \left(\widetilde{A}^\top \widetilde{A} \right), \\
\Delta &= C_6^2 - 4 C_4 \delta \left(\frac{1}{\delta} + C_5 \delta \right),
\end{aligned}$$

where η is some positive constant satisfying that $0 < \eta < \frac{S_f}{d^2 (1 + (d_\infty^- d^- L_f)^2)}$, and $\delta < \lambda_{\min}(M + M^\top)/(2C_7)$ is a positive constant reflecting the convergence rate.

Let Assumptions A4 and A5 hold. Then with proper step-size $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, there exist, $0 < \Gamma < \infty$ and $0 < \gamma < 1$, such that the sequence $\left\{ \mathbf{t}^k \right\}$ defined in Eq. (5.29) *satisfies*

$$\left\| \mathbf{t}^k - \mathbf{t}^* \right\|_G^2 \geq (1 + \delta) \left\| \mathbf{t}^{k+1} - \mathbf{t}^* \right\|_G^2 - \Gamma \gamma^k. \tag{5.33}$$

The constant γ is the same as used in Eq. (5.17)*, reflecting the consensus rate. The lower bound, $\alpha_{\min}$, of α satisfies $\alpha_{\min} \leq \underline{\alpha}$, where*

$$\underline{\alpha} \triangleq \frac{C_6 - \sqrt{\triangle}}{2 C_4 \delta}, \tag{5.34}$$

and the upper bound, $\alpha_{\max}$, of α satisfies $\alpha_{\max} \geq \overline{\alpha}$, where

$$\overline{\alpha} \triangleq \min\left\{\frac{\eta\lambda_{min}\left(M+M^\top\right)}{2(d_\infty^- d^- L_f)^2}, \frac{C_6+\sqrt{\triangle}}{2C_4\delta}\right\}. \tag{5.35}$$

Proof. See next section. □

Theorem 3 is the key result of this paper. We will show the complete proof of Theorem 3 in next section. Note that Theorem 3 shows the relation between $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2$ and $\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2$ but we would like to show that $\mathbf{z}^k$ converges linearly to the optimal point $\mathbf{z}^*$, which Theorem 3 does not show. To this aim, we provide Theorem 4 that describes a relation between $\|\mathbf{z}^k - \mathbf{z}^*\|^2$ and $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2$.

In Theorem 3, we are given specific bounds on $\alpha_{\min}$ and $\alpha_{\max}$. In order to ensure that the solution set of step-size, α, is not empty, i.e., $\alpha_{\min} \leq \alpha_{\max}$, it is sufficient (but not necessary) to satisfy

$$\underline{\alpha} = \frac{C_6-\sqrt{\triangle}}{2C_4\delta} \leq \frac{\eta\lambda_{min}\left(M+M^\top\right)}{2(d_\infty^- d^- L_f)^2} \leq \overline{\alpha}, \tag{5.36}$$

which is equivalent to

$$\eta \geq \frac{\left(\frac{S_f}{2d^2} - \sqrt{\Delta}\right)/\left(2C_4\delta\right)}{\frac{\lambda_{min}\left(M+M^\top\right)}{2L_f^2(d_\infty^- d^-)^2} + \frac{1+2(d_\infty^- d^- L_f)^2}{4C_4\delta}}. \tag{5.37}$$

Recall from Theorem 3 that

$$\eta \leq \frac{S_f}{d^2(1+(d_\infty^- d^- L_f)^2)}. \tag{5.38}$$

We note that it may not always be possible to find solutions for η that satisfy both Eqs. (5.37) and (5.38). The theoretical restriction here is due to the fact that the step-size bounds in Theorem 3 are not tight. However, the representation of $\underline{\alpha}$ and $\overline{\alpha}$ imply how to increase the interval of appropriate step-sizes. For example, it may be useful to set the weights to increase $\lambda_{\min}\left(M+N^{\top}\right)/(2d_{\infty}{}^{-}d^{-})^2$ such that $\overline{\alpha}$ is increased. We will discuss such strategies in the numerical experiments. We also observe that in reality, the range of appropriate step-sizes is much wider. Note that the values of $\underline{\alpha}$ and $\overline{\alpha}$ need the knowledge of the network topology, which may not be available in a distributed manner. Such bounds are not uncommon in the literature where the step-size is a function of the entire topology or global objective functions, see [46, 48]. It is an open question on how to avoid the global knowledge of network topology when designing the interval of α.

Remark 1. *The positive constant δ in Eq.* (5.33) *reflects the convergence rate of $\|\mathbf{t}^k-\mathbf{t}^*\|_G^2$. The larger δ is, the faster $\|\mathbf{t}^k-\mathbf{t}^*\|_G^2$ converges to zero. As $\delta<\lambda_{\max}(M+M^{\top})/(2C_7)$, we claim that the convergence rate of $\|\mathbf{t}^k-\mathbf{t}^*\|_G^2$ can not be arbitrarily large.*

Based on Theorem 3, we now show the r-linear convergence rate of DEXTRA to the optimal solution.

Theorem 4. *Let Assumptions A4 and A5 hold. With the same step-size, α, used in Theorem 3, the sequence, $\{\mathbf{z}^k\}$, generated by DEXTRA, converges exactly to the optimal solution, $\mathbf{z}^*$, at an r-linear rate, i.e., there exist some bounded constants, $T > 0$ and $\max\left\{\frac{1}{1+\delta}, \gamma\right\} < \tau < 1$, where δ and γ are constants used in Theorem 3, Eq.* (5.33)*, such that for any k,*

$$\left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|^2 \leq T\tau^k.$$

Proof. We start with Eq. (5.33) in Theorem 3, which is defined earlier in Eq. (5.29). Since the G-matrix norm is non-negative. recall Eq. (5.32), we have $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq 0$, for any k. Define $\psi = \max\left\{\frac{1}{1+\delta}, \gamma\right\}$, where δ and γ are constants in Theorem 3. From Eq. (5.33), we have for any k,

$$\begin{aligned}\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 &\leq \frac{1}{1+\delta}\left\|\mathbf{t}^{k-1} - \mathbf{t}^*\right\|_G^2 + \Gamma\frac{\gamma^{k-1}}{1+\delta},\\ &\leq \psi\left\|\mathbf{t}^{k-1} - \mathbf{t}^*\right\|_G^2 + \Gamma\psi^k,\\ &\leq \psi^k\left\|\mathbf{t}^0 - \mathbf{t}^*\right\|_G^2 + k\Gamma\psi^k.\end{aligned}$$

For any τ satisfying $\psi < \tau < 1$, there exists a constant Ψ such that $(\frac{\tau}{\psi})^k > \frac{k}{\Psi}$, for all k. Therefore, we obtain that

$$\begin{aligned}\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 &\leq \tau^k\left\|\mathbf{t}^0 - \mathbf{t}^*\right\|_G^2 + (\Psi\Gamma)\frac{k}{\Psi}\left(\frac{\psi}{\tau}\right)^k\tau^k,\\ &\leq \left(\left\|\mathbf{t}^0 - \mathbf{t}^*\right\|_G^2 + \Psi\Gamma\right)\tau^k. \qquad (5.39)\end{aligned}$$

From Eq. (5.29) and the corresponding discussion, we have

$$\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 = \left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|_{M^\top}^2 + \left\|\mathbf{q}^k - \mathbf{q}^*\right\|_N^2.$$

Since $N + N^\top$ is PSD, (see Eq. (5.31)), it follows that

$$\left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|^2_{\frac{M+M^\top}{2}} \leq \left\|\mathbf{t}^k - \mathbf{t}^*\right\|^2_G.$$

Noting that $M + M^\top$ is PD (see Assumption A5(c)), i.e., all eigenvalues of $M + M^\top$ are positive, we obtain that

$$\left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|^2_{\frac{\lambda_{\min}(M+M^\top)}{2}I_{np}} \leq \left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|^2_{\frac{M+M^\top}{2}}.$$

Therefore, we have that

$$\begin{aligned}\frac{\lambda_{\min}(M+M^\top)}{2}\left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|^2 &\leq \left\|\mathbf{t}^k - \mathbf{t}^*\right\|^2_G \\ &\leq \left(\left\|\mathbf{t}^0 - \mathbf{t}^*\right\|^2_G + \Psi\Gamma\right)\tau^k.\end{aligned}$$

By letting

$$T = 2\frac{\|\mathbf{t}^0 - \mathbf{t}^*\|^2_G + \Psi\Gamma}{\lambda_{\min}(M+M^\top)},$$

we obtain the desired result. □

Theorem 4 shows that the sequence, $\{\mathbf{z}^k\}$, converges at an r-linear rate to the optimal solution, $\mathbf{z}^*$, where the convergence rate is described by the constant, τ. During the derivation of τ, we have τ satisfying that $\gamma \leq \max\{\frac{1}{1+\delta}, \gamma\} < \tau < 1$. This implies that the convergence rate (described by the constant τ) is bounded by the consensus rate (described by the constant γ). In next section, we present some basic relations. Based on these relations, we finally state the proof of Theorem 3.

5.3 Auxiliary Relations

We provide several basic relations in this section, which will help in the proof of Theorem 3. We first establish a relation among $D^k\mathbf{z}^k$, $\mathbf{q}^k$, $D^\infty\mathbf{z}^*$, and $\mathbf{q}^*$.

Lemma 13. *Let Assumptions A4 and A5 hold. In DEXTRA, the quadruple sequence* $\{D^k\mathbf{z}^k, \mathbf{q}^k, D^\infty\mathbf{z}^*, \mathbf{q}^*\}$ *obeys, for any* k*,*

$$\begin{aligned} &R\left(D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right) + \widetilde{A}\left(D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right) \\ =& - L\left(\mathbf{q}^{k+1} - \mathbf{q}^*\right) - \alpha\left[\nabla\mathbf{f}(\mathbf{z}^k) - \nabla\mathbf{f}(\mathbf{z}^*)\right], \end{aligned} \tag{5.40}$$

recall Eqs. (5.18)–(5.28) *for notation.*

Proof. We sum DEXTRA, Eq. (5.9), over time from 0 to k,

$$D^{k+1}\mathbf{z}^{k+1} = \widetilde{A}D^k\mathbf{z}^k - \alpha\nabla\mathbf{f}(\mathbf{z}^k) - L\sum_{r=0}^{k} D^r\mathbf{z}^r.$$

By subtracting $LD^{k+1}\mathbf{z}^{k+1}$ on both sides of the preceding equation and rearranging the terms, it follows that

$$RD^{k+1}\mathbf{z}^{k+1} + \widetilde{A}\left(D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right) = -L\mathbf{q}^{k+1} - \alpha\nabla\mathbf{f}\left(\mathbf{z}^k\right). \tag{5.41}$$

Note that $D^\infty\mathbf{z}^* = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is some vector in the span of the right-eigenvector of A corresponding to eigenvalue 1. Since $R\boldsymbol{\pi} = \mathbf{0}_n$, we have

$$RD^\infty\mathbf{z}^* = \mathbf{0}_n. \tag{5.42}$$

By subtracting Eq. (5.42) from Eq. (5.41), and noting that $L\mathbf{q}^* + \alpha\nabla\mathbf{f}(\mathbf{z}^*) = \mathbf{0}_n$, Eq. (5.27), we obtain the desired result. □

Recall Eq. (5.17) that shows the convergence of D^k to D^∞ at a geometric rate. We will use this result to develop a relation between $\left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\|$ and $\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|$, which is in the following lemma. Similarly, we can establish a relation between $\left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\|$ and $\left\|\mathbf{z}^{k+1} - \mathbf{z}^*\right\|$.

Lemma 14. *Let Assumptions A4 and A5 hold and recall the constants d and d^- from Eqs.* (5.24) *and* (5.25)*. If $\mathbf{z}^k$ is bounded, i.e., $\|\mathbf{z}^k\| \leq B < \infty$, then*

(a) $\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| \leq d^- \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\| + 2d^- nCB\gamma^k$;

(b) $\left\|\mathbf{z}^{k+1} - \mathbf{z}^*\right\| \leq d^- \left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\| + d^- nCB\gamma^k$;

(c) $\left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\| \leq d \left\|\mathbf{z}^{k+1} - \mathbf{z}^*\right\| + nCB\gamma^k$;

where C and γ are constants defined in Lemma 11.

Proof. (a)

$$\begin{aligned}
\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| &= \left\|\left(D^{k+1}\right)^{-1}\left(D^{k+1}\right)\left(\mathbf{z}^{k+1} - \mathbf{z}^k\right)\right\|, \\
&\leq \left\|\left(D^{k+1}\right)^{-1}\right\| \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k + D^k\mathbf{z}^k - D^{k+1}\mathbf{z}^k\right\|, \\
&\leq d^- \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\| + d^- \left\|D^k - D^{k+1}\right\| \left\|\mathbf{z}^k\right\|, \\
&\leq d^- \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\| + 2d^- nCB\gamma^k.
\end{aligned}$$

Similarly, we can prove (b). Finally, we have

$$\left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\| = \left\|D^{k+1}\mathbf{z}^{k+1} - D^{k+1}\mathbf{z}^* + D^{k+1}\mathbf{z}^* - D^\infty\mathbf{z}^*\right\|,$$

$$\leq d\left\|\mathbf{z}^{k+1}-\mathbf{z}^*\right\|+\left\|D^{k+1}-D^{\infty}\right\|\|\mathbf{z}^*\|,$$
$$\leq d\left\|\mathbf{z}^{k+1}-\mathbf{z}^*\right\|+nCB\gamma^k.$$

The proof is complete. □

Note that the result of Lemma 14 is based on the prerequisite that the sequence $\{\mathbf{z}^k\}$ generated by DEXTRA at kth iteration is bounded. We will show this boundedness property (for all k) together with the proof of Theorem 3 in the next section. The following two lemmas discuss the boundedness of $\|\mathbf{z}^k\|$ for a fixed k.

Lemma 15. *Let Assumptions A4 and A5 hold and recall* $\mathbf{t}^k$, $\mathbf{t}^*$, *and* G *defined in Eq.* (5.29). *If* $\|\mathbf{t}^k-\mathbf{t}^*\|_G^2$ *is bounded by some constant* F *for some* k, *i.e.,*$\|\mathbf{t}^k-\mathbf{t}^*\|_G^2\leq F$, *we have* $\|\mathbf{z}^k\|$ *be bounded by a constant* B *for the same* k, *defined as follow,*

$$\left\|\mathbf{z}^k\right\|\leq B\triangleq\sqrt{\frac{2(d^-)^2F}{\lambda_{\min}\left(\frac{M+M^\top}{2}\right)}+2(d^-)^2\left\|D^{\infty}\mathbf{z}^*\right\|^2},\tag{5.43}$$

where d^-, M *are constants defined in Eq.* (5.25) *and* (5.18).

Proof. We follow the following derivation,

$$\frac{1}{2}\left\|\mathbf{z}^k\right\|^2\leq\frac{(d^-)^2}{2}\left\|D^k\mathbf{z}^k\right\|^2,$$
$$\leq(d^-)^2\left\|D^k\mathbf{z}^k-D^{\infty}\mathbf{z}^*\right\|^2+(d^-)^2\left\|D^{\infty}\mathbf{z}^*\right\|^2,$$

$$\begin{aligned}
&\leq \frac{(d^-)^2}{\lambda_{\min}\left(\frac{M+M^\top}{2}\right)} \left\|D^k\mathbf{z}^k - D^\infty\mathbf{z}^*\right\|_{M^\top}^2 + (d^-)^2 \left\|D^\infty\mathbf{z}^*\right\|^2, \\
&\leq \frac{(d^-)^2}{\lambda_{\min}\left(\frac{M+M^\top}{2}\right)} \left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 + (d^-)^2 \left\|D^\infty\mathbf{z}^*\right\|^2, \\
&\leq \frac{(d^-)^2 F}{\lambda_{\min}\left(\frac{M+M^\top}{2}\right)} + (d^-)^2 \left\|D^\infty\mathbf{z}^*\right\|^2,
\end{aligned}$$

where the third inequality holds due to $M + M^\top$ being PD (see Assumption A5(c)), and the fourth inequality holds because N-matrix norm has been shown to be nonnegative (see Eq. (5.31)). Therefore, it follows that $\left\|\mathbf{z}^k\right\| \leq B$ for B defined in Eq. (5.43), which is clearly $< \infty$ as long as $F < \infty$. □

Lemma 16. *Let Assumptions A4 and A5 hold and recall the definition of constant C_1 from Theorem 3. If $\|\mathbf{z}^{k-1}\|$ and $\|\mathbf{z}^k\|$ are bounded by a same constant B, we have that $\|\mathbf{z}^{k+1}\|$ is also bounded. More specifically, we have $\|\mathbf{z}^{k+1}\| \leq C_1 B$.*

Proof. According to the iteration of DEXTRA in Eq. (5.9), we can bound $D^{k+1}\mathbf{z}^{k+1}$ as

$$\begin{aligned}
\left\|D^{k+1}\mathbf{z}^{k+1}\right\| &\leq \left\|(I_n + A)D^k\right\| \left\|\mathbf{z}^k\right\| + \left\|\widetilde{A}D^{k-1}\right\| \left\|\mathbf{z}^{k-1}\right\| \\
&\quad + \alpha L_f \left\|\mathbf{z}^k\right\| + \alpha L_f \left\|\mathbf{z}^{k-1}\right\|, \\
&\leq \left[d\left\|(I_n + A)\right\| + d\left\|\widetilde{A}\right\| + 2\alpha L_f\right] B,
\end{aligned}$$

where d is the constant defined in Eq. (5.24). Accordingly, we have $\mathbf{z}^{k+1}$ be

bounded as follow,

$$\left\|\mathbf{z}^{k+1}\right\| \leq d^{-} \left\|D^{k+1}\mathbf{z}^{k+1}\right\| = C_1 B. \tag{5.44}$$

□

5.4 Convergence Analysis

In this section, we first give two propositions that provide the main framework of the proof. Based on these propositions, we use induction to prove Theorem 3. Proposition 2 claims that for all $k \in \mathbb{N}^+$, if $\|\mathbf{t}^{k-1} - \mathbf{t}^*\|_G^2 \leq F_1$ and $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F_1$, for some bounded constant F_1, then, $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k$, for some appropriate step-size.

Proposition 2. *Let Assumptions A4 and A5 hold, and recall the constants C_1, C_2, C_3, C_4, C_5, C_6, C_7, Δ, δ, and γ from Theorem 3. Assume $\|\mathbf{t}^{k-1} - \mathbf{t}^*\|_G^2 \leq F_1$ and $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F_1$, for a same bounded constant F_1. Let the constant B be a function of F_1 as defined in Eq. (5.43) by substituting F with F_1, and we define Γ as*

$$\Gamma = C_3 B^2. \tag{5.45}$$

With proper step-size α, Eq. (5.33) is satisfied at kth iteration, i.e.,

$$\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 \geq (1+\delta)\left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2 - \Gamma\gamma^k,$$

where the range of step-size is given in Eqs. (5.34) *and* (5.35) *in Theorem 3.*

Proof. We first bound $\|\mathbf{z}^{k-1}\|$, $\|\mathbf{z}^{k}\|$, and $\|\mathbf{z}^{k+1}\|$. According to Lemma 15, we obtain that $\|\mathbf{z}^{k-1}\| \leq B$ and $\|\mathbf{z}^{k}\| \leq B$, since $\|\mathbf{t}^{k-1} - \mathbf{t}^{*}\|_G^2 \leq F_1$ and $\|\mathbf{t}^{k} - \mathbf{t}^{*}\|_G^2 \leq F_1$. By applying Lemma 16, we further obtain that $\|\mathbf{z}^{k+1}\| \leq C_1 B$. Based on the boundedness of $\|\mathbf{z}^{k-1}\|$, $\|\mathbf{z}^{k}\|$, and $\|\mathbf{z}^{k+1}\|$, we start to prove the desired result. By applying the restricted strong-convexity assumption, Eq. (5.15b), it follows that

$$\begin{aligned}
2\alpha S_f \left\|\mathbf{z}^{k+1} - \mathbf{z}^{*}\right\|^2 \leq & 2\alpha \left\langle D^{\infty}\left(\mathbf{z}^{k+1} - \mathbf{z}^{*}\right), (D^{\infty})^{-1}\left[\nabla \mathbf{f}(\mathbf{z}^{k+1}) - \nabla \mathbf{f}(\mathbf{z}^{*})\right]\right\rangle, \\
= & 2\alpha \left\langle D^{\infty}\mathbf{z}^{k+1} - D^{k+1}\mathbf{z}^{k+1}, (D^{\infty})^{-1}\left[\nabla \mathbf{f}(\mathbf{z}^{k+1}) - \nabla \mathbf{f}(\mathbf{z}^{*})\right]\right\rangle \\
& + 2\alpha \left\langle D^{k+1}\mathbf{z}^{k+1} - D^{\infty}\mathbf{z}^{*}, (D^{\infty})^{-1}\left[\nabla \mathbf{f}(\mathbf{z}^{k+1}) - \nabla \mathbf{f}(\mathbf{z}^{k})\right]\right\rangle \\
& + 2\left\langle D^{k+1}\mathbf{z}^{k+1} - D^{\infty}\mathbf{z}^{*}, (D^{\infty})^{-1}\alpha\left[\nabla \mathbf{f}(\mathbf{z}^{k}) - \nabla \mathbf{f}(\mathbf{z}^{*})\right]\right\rangle, \\
:= & s_1 + s_2 + s_3, \qquad (5.46)
\end{aligned}$$

where s_1, s_2, s_3 denote each of RHS terms. We show the boundedness of s_1, s_2, and s_3 as follow.

Bounding s_1: By using $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \eta\|\mathbf{a}\|^2 + \frac{1}{\eta}\|\mathbf{b}\|^2$ for any appropriate vectors $\mathbf{a}, \mathbf{b}$, and a positive η, we obtain that

$$s_1 \leq \frac{\alpha}{\eta}\left\|D^{\infty} - D^{K+1}\right\|^2 \left\|\mathbf{z}^{K+1}\right\|^2 + \alpha\eta(d_{\infty}^{-}L_f)^2\left\|\mathbf{z}^{K+1} - \mathbf{z}^{*}\right\|^2. \qquad (5.47)$$

It follows $\left\|D^{\infty} - D^{K+1}\right\| \leq nC\gamma^K$ as shown in Eq. (5.17), and $\|\mathbf{z}^{K+1}\|^2 \leq C_1^2 B^2$ as shown in Eq. (5.44). The term $\|\mathbf{z}^{K+1} - \mathbf{z}^{*}\|$ can be bounded with

applying Lemma 14(b). Therefore,

$$\begin{aligned} s_1 \leq & \alpha(nC)^2 \left[\frac{C_1^2}{\eta} + 2\eta(d_\infty^- d^- L_f)^2\right] B^2\gamma^{2K} \\ & + 2\alpha\eta(d_\infty^- d^- L_f)^2 \left\|D^{K+1}\mathbf{z}^{K+1} - D^\infty \mathbf{z}^*\right\|^2. \end{aligned} \tag{5.48}$$

Bounding s_2: Similarly, we use Lemma 14(a) to obtain

$$\begin{aligned} s_2 \leq & \alpha\eta \left\|D^{K+1}\mathbf{z}^{K+1} - D^\infty \mathbf{z}^*\right\|^2 + \frac{\alpha(d_\infty^- L_f)^2}{\eta} \left\|\mathbf{z}^{K+1} - \mathbf{z}^k\right\|^2, \\ \leq & \alpha\eta \left\|D^{K+1}\mathbf{z}^{K+1} - D^\infty \mathbf{z}^*\right\|^2 + \frac{2\alpha(nCd_\infty^- d^- L_f)^2 B^2}{\eta}\gamma^{2K} \\ & + \frac{2\alpha(d_\infty^- d^- L_f)^2}{\eta} \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\|^2. \end{aligned} \tag{5.49}$$

Bounding s_3: We rearrange Eq. (5.40) in Lemma 13 as follow,

$$\begin{aligned} \alpha\left[\nabla\mathbf{f}(\mathbf{z}^k) - \nabla\mathbf{f}(\mathbf{z}^*)\right] = & R\left(D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right) \\ & + \widetilde{A}\left(D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right) + L\left(\mathbf{q}^{k+1} - \mathbf{q}^*\right). \end{aligned} \tag{5.50}$$

By substituting $\alpha[\nabla\mathbf{f}(\mathbf{z}^k) - \nabla\mathbf{f}(\mathbf{z}^*)]$ in s_3 with the representation in the preceding relation, we represent s_3 as

$$\begin{aligned} s_3 = & \left\|D^{K+1}\mathbf{z}^{K+1} - D^\infty\mathbf{z}^*\right\|_{-2Q}^2 \\ & + 2\left\langle D^{K+1}\mathbf{z}^{K+1} - D^\infty\mathbf{z}^*, M\left(D^K\mathbf{z}^K - D^{K+1}\mathbf{z}^{K+1}\right)\right\rangle \\ & + 2\left\langle D^{K+1}\mathbf{z}^{K+1} - D^\infty\mathbf{z}^*, N\left(\mathbf{q}^* - \mathbf{q}^{k+1}\right)\right\rangle, \\ := & s_{3a} + s_{3b} + s_{3c}, \end{aligned} \tag{5.51}$$

where s_{3b} is equivalent to

$$s_{3b} = 2\left\langle D^{K+1}\mathbf{z}^{K+1} - D^K\mathbf{z}^K, M^\top\left(D^\infty\mathbf{z}^* - D^{K+1}\mathbf{z}^{K+1}\right)\right\rangle,$$

and s_{3c} can be simplified as

$$\begin{aligned} s_{3c} &= 2\left\langle D^{K+1}\mathbf{z}^{K+1}, N\left(\mathbf{q}^{*}-\mathbf{q}^{K+1}\right)\right\rangle \\ &= 2\left\langle \mathbf{q}^{K+1}-\mathbf{q}^{K}, N\left(\mathbf{q}^{*}-\mathbf{q}^{K+1}\right)\right\rangle. \end{aligned}$$

The first equality in the preceding relation holds due to the fact that $N^{\top}D^{\infty}\mathbf{z}^{*} = \mathbf{0}_n$ and the second equality follows from the definition of $\mathbf{q}^k$, see Eq. (5.28). By substituting the representation of s_{3b} and s_{3c} into (5.51), and recalling the definition of $\mathbf{t}^k$, $\mathbf{t}^*$, G in Eq. (5.29), we simplify the representation of s_3,

$$s_3 = \left\|D^{K+1}\mathbf{z}^{K+1}-D^{\infty}\mathbf{z}^{*}\right\|_{-2Q}^{2} + 2\left\langle \mathbf{t}^{K+1}-\mathbf{t}^{K}, G\left(\mathbf{t}^{*}-\mathbf{t}^{K+1}\right)\right\rangle. \tag{5.52}$$

With the basic rule

$$\begin{aligned} &\left\langle \mathbf{t}^{K+1}-\mathbf{t}^{K}, G\left(\mathbf{t}^{*}-\mathbf{t}^{K+1}\right)\right\rangle + \left\langle G\left(\mathbf{t}^{K+1}-\mathbf{t}^{K}\right), \mathbf{t}^{*}-\mathbf{t}^{K+1}\right\rangle \\ &= \left\|\mathbf{t}^{K}-\mathbf{t}^{*}\right\|_{G}^{2} - \left\|\mathbf{t}^{K+1}-\mathbf{t}^{*}\right\|_{G}^{2} - \left\|\mathbf{t}^{K+1}-\mathbf{t}^{K}\right\|_{G}^{2}, \end{aligned} \tag{5.53}$$

We obtain that

$$\begin{aligned} s_3 = &\left\|D^{K+1}\mathbf{z}^{K+1}-D^{\infty}\mathbf{z}^{*}\right\|_{-2Q}^{2} \\ &+ 2\left\|\mathbf{t}^{K}-\mathbf{t}^{*}\right\|_{G}^{2} - 2\left\|\mathbf{t}^{K+1}-\mathbf{t}^{*}\right\|_{G}^{2} - 2\left\|\mathbf{t}^{K+1}-\mathbf{t}^{K}\right\|_{G}^{2} \\ &- 2\left\langle G\left(\mathbf{t}^{K+1}-\mathbf{t}^{K}\right), \mathbf{t}^{*}-\mathbf{t}^{K+1}\right\rangle. \end{aligned} \tag{5.54}$$

We analyze the last two terms in Eq. (5.54):

$$-2\left\|\mathbf{t}^{K+1}-\mathbf{t}^{K}\right\|_{G}^{2} \leq -2\left\|D^{K+1}\mathbf{z}^{K+1}-D^{K}\mathbf{z}^{K}\right\|_{M^{\top}}^{2}, \tag{5.55}$$

where the inequality holds due to N-matrix norm is nonnegative, and

$$\begin{aligned}
&-2\left\langle G\left(\mathbf{t}^{K+1}-\mathbf{t}^{K}\right),\mathbf{t}^{*}-\mathbf{t}^{K+1}\right\rangle \\
=&-2\left\langle M^{\top}\left(D^{K+1}\mathbf{z}^{K+1}-D^{K}\mathbf{z}^{K}\right),D^{\infty}\mathbf{z}^{*}-D^{K+1}\mathbf{z}^{K+1}\right\rangle \\
&-2\left\langle D^{K+1}\mathbf{z}^{K+1}-D^{\infty}\mathbf{z}^{*},N^{\top}\left(\mathbf{q}^{*}-\mathbf{q}^{K+1}\right)\right\rangle, \\
\leq&\delta\left\|D^{K+1}\mathbf{z}^{K+1}-D^{K}\mathbf{z}^{K}\right\|_{MM^{\top}}^{2}+\delta\left\|\mathbf{q}^{*}-\mathbf{q}^{K+1}\right\|_{NN^{\top}}^{2} \\
&+\frac{2}{\delta}\left\|D^{K+1}\mathbf{z}^{K+1}-D^{\infty}\mathbf{z}^{*}\right\|^{2}, \qquad (5.56)
\end{aligned}$$

for some $\delta > 0$. By substituting Eqs. (5.55) and (5.56) into Eq. (5.54), we obtain that

$$\begin{aligned}
s_3 \leq&2\left\|\mathbf{t}^{K}-\mathbf{t}^{*}\right\|_{G}^{2}-2\left\|\mathbf{t}^{K+1}-\mathbf{t}^{*}\right\|_{G}^{2}+\left\|\mathbf{q}^{*}-\mathbf{q}^{K+1}\right\|_{\delta NN^{\top}}^{2} \\
&+\left\|D^{K+1}\mathbf{z}^{K+1}-D^{\infty}\mathbf{z}^{*}\right\|_{\frac{2}{\delta}I_n-2Q}^{2} \\
&+\left\|D^{K+1}\mathbf{z}^{K+1}-D^{K}\mathbf{z}^{K}\right\|_{-2M^{\top}+\delta MM^{\top}}^{2}. \qquad (5.57)
\end{aligned}$$

Next, it follows from Lemma 14(c) that

$$\left\|D^{k+1}\mathbf{z}^{k+1}-D^{\infty}\mathbf{z}^{*}\right\|^{2}\leq 2d^{2}\left\|\mathbf{z}^{k+1}-\mathbf{z}^{*}\right\|^{2}+2(nCB)^{2}\gamma^{2k}.$$

Multiplying both sides of the preceding relation by $\frac{\alpha S_f}{d^2}$ and combining it with Eq. (55), we obtain

$$\frac{\alpha S_f}{d^2}\left\|D^{k+1}\mathbf{z}^{k+1}-D^{\infty}\mathbf{z}^{*}\right\|^{2}\leq s_1+s_2+s_3+\frac{2\alpha S_f(nCB)^2}{d^2}\gamma^{2k}. \qquad (5.58)$$

By plugging the related bounds (s_1 from Eq. (5.48), s_2 from Eq. (5.49), and s_3 from Eq. (5.57)) in Eq. (5.58), it follows that

$$\begin{aligned}\left\|\mathbf{t}^k-\mathbf{t}^*\right\|_G^2-\left\|\mathbf{t}^{k+1}-\mathbf{t}^*\right\|_G^2 \geq & \left\|D^{k+1}\mathbf{z}^{k+1}-D^{\infty}\mathbf{z}^*\right\|^2_{\frac{\alpha}{2}\left[\frac{S_f}{d^2}-\eta-2\eta(d_{\infty}^{-}d^{-}L_f)^2\right]I_n-\frac{1}{\delta}I_n+Q} \\ & +\left\|D^{k+1}\mathbf{z}^{k+1}-D^{k}\mathbf{z}^{k}\right\|^2_{M^{\top}-\frac{\delta}{2}MM^{\top}-\frac{\alpha(d_{\infty}^{-}d^{-}L_f)^2}{\eta}I_n} \\ & -\alpha(nC)^2\left[\frac{C_1^2}{2\eta}+(d_{\infty}^{-}d^{-}L_f)^2\left(\eta+\frac{1}{\eta}\right)+\frac{S_f}{d^2}\right]B^2\gamma^k \\ & -\left\|\mathbf{q}^*-\mathbf{q}^{k+1}\right\|^2_{\frac{\delta}{2}NN^{\top}}. \end{aligned} \tag{5.59}$$

In order to derive the relation that

$$\left\|\mathbf{t}^k-\mathbf{t}^*\right\|_G^2 \geq (1+\delta)\left\|\mathbf{t}^{k+1}-\mathbf{t}^*\right\|_G^2-\Gamma\gamma^k, \tag{5.60}$$

it is sufficient to show that the RHS of Eq. (5.59) is no less than $\delta\left\|\mathbf{t}^{k+1}-\mathbf{t}^*\right\|_G^2-\Gamma\gamma^k$. Recall the definition of G, $\mathbf{t}^k$, and $\mathbf{t}^*$ in Eq. (5.29), we have

$$\begin{aligned}\delta\left\|\mathbf{t}^{k+1}-\mathbf{t}^*\right\|_G^2-\Gamma\gamma^k = & \left\|D^{k+1}\mathbf{z}^{k+1}-D^{\infty}\mathbf{z}^*\right\|^2_{\delta M^{\top}} \\ & +\left\|\mathbf{q}^*-\mathbf{q}^{k+1}\right\|^2_{\delta N}-\Gamma\gamma^k. \end{aligned} \tag{5.61}$$

Comparing Eqs. (5.59) with (5.61), it is sufficient to prove that

$$\begin{aligned}& \left\|D^{k+1}\mathbf{z}^{k+1}-D^{\infty}\mathbf{z}^*\right\|^2_{\frac{\alpha}{2}\left[\frac{S_f}{d^2}-\eta-2\eta(d_{\infty}^{-}d^{-}L_f)^2\right]I_n-\frac{1}{\delta}I_n+Q-\delta M^{\top}} \\ & +\left\|D^{k+1}\mathbf{z}^{k+1}-D^{k}\mathbf{z}^{k}\right\|^2_{M^{\top}-\frac{\delta}{2}MM^{\top}-\frac{\alpha(d_{\infty}^{-}d^{-}L_f)^2}{\eta}I_n} \\ & +\Gamma\gamma^k-\alpha(nC)^2\left[\frac{C_1^2}{2\eta}+(d_{\infty}^{-}d^{-}L_f)^2\left(\eta+\frac{1}{\eta}\right)+\frac{S_f}{d^2}\right]B^2\gamma^k \\ & \geq \left\|\mathbf{q}^*-\mathbf{q}^{k+1}\right\|^2_{\delta\left(\frac{NN^{\top}}{2}+N\right)}. \end{aligned} \tag{5.62}$$

We next aim to bound $\|\mathbf{q}^* - \mathbf{q}^{k+1}\|^2_{\delta(\frac{NN^\top}{2}+N)}$ in terms of $\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\|$ and $\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\|$, such that it is easier to analyze Eq. (5.62). From Lemma 13, we have

$$\begin{aligned}
\left\|\mathbf{q}^* - \mathbf{q}^{k+1}\right\|^2_{L^\top L} =& \left\|L\left(\mathbf{q}^* - \mathbf{q}^{k+1}\right)\right\|^2, \\
=& \Big\| R(D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*) + \alpha[\nabla\mathbf{f}(\mathbf{z}^{k+1}) - \nabla\mathbf{f}(\mathbf{z}^*)] \\
&+ \widetilde{A}(D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k) + \alpha[\nabla\mathbf{f}(\mathbf{z}^k) - \nabla\mathbf{f}(\mathbf{z}^{k+1})]\Big\|^2, \\
\leq& 4\left(\left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\|^2_{R^\top R} + \alpha^2L_f^2\left\|\mathbf{z}^{k+1} - \mathbf{z}^*\right\|^2\right) \\
&+ 4\left(\left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\|^2_{\widetilde{A}^\top\widetilde{A}} + \alpha^2L_f^2\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2\right), \\
\leq& \left\|D^{k+1}\mathbf{z}^{k+1} - D^\infty\mathbf{z}^*\right\|^2_{4R^\top R+8(\alpha L_f d^-)^2 I_n} \\
&+ \left\|D^{k+1}\mathbf{z}^{k+1} - D^k\mathbf{z}^k\right\|^2_{4\widetilde{A}^\top\widetilde{A}+8(\alpha L_f d^-)^2 I_n} \\
&+ 24\left(\alpha nCd^-L_f\right)^2B^2\gamma^k. \qquad (5.63)
\end{aligned}$$

Since that $\lambda\left(\frac{N+N^\top}{2}\right) \geq 0$, $\lambda\left(NN^\top\right) \geq 0$, $\lambda\left(L^\top L\right) \geq 0$, and $\lambda_{\min}\left(\frac{N+N^\top}{2}\right) = \lambda_{\min}\left(NN^\top\right) = \lambda_{\min}\left(L^\top L\right) = 0$ with the same corresponding eigenvector, we have

$$\left\|\mathbf{q}^* - \mathbf{q}^{k+1}\right\|^2_{\delta\left(\frac{NN^\top}{2}+N\right)} \leq \delta C_2\left\|\mathbf{q}^* - \mathbf{q}^{k+1}\right\|^2_{L^\top L}, \qquad (5.64)$$

where C_2 is the constant defined in Theorem 3. By combining Eqs. (5.63) with (5.64), it follows that

$$\left\|\mathbf{q}^* - \mathbf{q}^{k+1}\right\|^2_{\delta\left(\frac{NN^\top}{2}+N\right)} \leq \delta C_2\left\|\mathbf{q}^* - \mathbf{q}^{k+1}\right\|^2_{L^\top L},$$

$$\begin{aligned}
&\leq \left\|D^{k+1}\mathbf{z}^{k+1} - D^{\infty}\mathbf{z}^{*}\right\|^2_{\delta C_2\left(4R^{\top}R+8(\alpha L_f d^{-})^2 I_n\right)} \\
&\quad + \left\|D^{k+1}\mathbf{z}^{k+1} - D^{k}\mathbf{z}^{k}\right\|^2_{\delta C_2\left(4\widetilde{A}^{\top}\widetilde{A}+8(\alpha L_f d^{-})^2 I_n\right)} \\
&\quad + 24\delta C_2\left(\alpha n C d^{-} L_f\right)^2 B^2\gamma^k. \qquad (5.65)
\end{aligned}$$

Consider Eq. (5.62), together with (5.65). Let

$$\Gamma = C_3 B^2, \qquad (5.66)$$

where C_3 is the constant defined in Theorem 3, such that all "γ^k items" in Eqs. (5.62) and (5.65) can be canceled out. In order to prove Eq. (5.62), it is sufficient to show that the LHS of Eq. (5.62) is no less than the RHS of Eq. (5.65), i.e.,

$$\begin{aligned}
&\left\|D^{k+1}\mathbf{z}^{k+1} - D^{\infty}\mathbf{z}^{*}\right\|^2_{\frac{\alpha}{2}\left[\frac{S_f}{d^2}-\eta-2\eta(d^{-}_{\infty}d^{-}L_f)^2\right]I_n-\frac{1}{\delta}I_n+Q-\delta M^{\top}} \\
&\quad + \left\|D^{k+1}\mathbf{z}^{k+1} - D^{k}\mathbf{z}^{k}\right\|^2_{M^{\top}-\frac{\delta}{2}MM^{\top}-\frac{\alpha(d^{-}_{\infty}d^{-}L_f)^2}{\eta}I_n} \\
\geq &\left\|D^{k+1}\mathbf{z}^{k+1} - D^{\infty}\mathbf{z}^{*}\right\|^2_{\delta C_2\left(4R^{\top}R+8(\alpha L_f d^{-})^2 I_n\right)} \\
&\quad + \left\|D^{k+1}\mathbf{z}^{k+1} - D^{k}s\mathbf{z}^{k}\right\|^2_{\delta C_2\left(4\widetilde{A}^{\top}\widetilde{A}+8(\alpha L_f d^{-})^2 I_n\right)}. \qquad (5.67)
\end{aligned}$$

To satisfy Eq. (5.67), it is sufficient to have the following two relations hold simultaneously,

$$\begin{aligned}
&\frac{\alpha}{2}\left[\frac{S_f}{d^2}-\eta-2\eta(d^{-}_{\infty}d^{-}L_f)^2\right]-\frac{1}{\delta}-\delta\lambda_{\max}\left(\frac{M+M^{\top}}{2}\right) \\
\geq &\delta C_2\left[4\lambda_{\max}\left(R^{\top}R\right)+8(\alpha L_f d^{-})^2\right], \qquad (5.68a)
\end{aligned}$$

$$\lambda_{\min}\left(\frac{M+M^\top}{2}\right)-\frac{\delta}{2}\lambda_{\max}\left(MM^\top\right)-\frac{\alpha(d_\infty^- d^- L_f)^2}{\eta}$$

$$\geq \delta C_2\left[4\lambda_{\max}\left(\widetilde{A}^\top\widetilde{A}\right)+8(\alpha L_f d^-)^2\right]. \tag{5.68b}$$

where in Eq. (5.68a) we ignore the term $\frac{\lambda_{\min}\left(Q+Q^\top\right)}{2}$ due to $\lambda_{\min}\left(Q+Q^\top\right)=0$. Recall the definition

$$C_4 = 8C_2\left(L_f d^-\right)^2, \tag{5.69}$$

$$C_5 = \lambda_{\max}\left(\frac{M+M^\top}{2}\right)+4C_2\lambda_{\max}\left(R^\top R\right), \tag{5.70}$$

$$C_6 = \frac{\frac{S_f}{d^2}-\eta-2\eta(d_\infty^- d^- L_f)^2}{2}, \tag{5.71}$$

$$\Delta = C_6^2-4C_4\delta\left(\frac{1}{\delta}+C_5\delta\right). \tag{5.72}$$

The solution of step-size, α, satisfying Eq. (5.68a), is

$$\frac{C_6-\sqrt{\Delta}}{2C_4\delta}\leq\alpha\leq\frac{C_6+\sqrt{\Delta}}{2C_4\delta}, \tag{5.73}$$

where we set

$$\eta<\frac{S_f}{d^2(1+(d_\infty^- d^- L_f)^2)}, \tag{5.74}$$

to ensure the solution of α contains positive values. In order to have $\delta>0$ in Eq. (5.68b), the step-size, α, is sufficient to satisfy

$$\alpha\leq\frac{\eta\lambda_{min}\left(M+M^\top\right)}{2(d_\infty^- d^- L_f)^2}. \tag{5.75}$$

By combining Eqs. (5.73) with (5.75), we conclude it is sufficient to set the step-size $\alpha\in[\underline{\alpha},\overline{\alpha}]$, where

$$\underline{\alpha}\triangleq\frac{C_6-\sqrt{\triangle}}{2C_4\delta}, \tag{5.76}$$

and

$$\overline{\alpha} \triangleq \min\left\{\frac{\eta\lambda_{min}\left(M+M^\top\right)}{2(d_\infty^- d^- L_f)^2}, \frac{C_6+\sqrt{\triangle}}{2C_4\delta}\right\}, \tag{5.77}$$

to establish the desired result, i.e.,

$$\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k. \tag{5.78}$$

Finally, we bound the constant δ, which reflecting how fast $\left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2$ converges. Recall the definition of C_7

$$C_7 = \frac{1}{2}\lambda_{\max}\left(MM^\top\right) + 4C_2\lambda_{\max}\left(\widetilde{A}^\top\widetilde{A}\right). \tag{5.79}$$

To have α's solution of Eq. (5.68b) contains positive values, we need to set

$$\delta < \frac{\lambda_{\min}\left(M+M^\top\right)}{2C_7}. \tag{5.80}$$

□

Note that Proposition 2 is different from Theorem 3 in that: (i) it only proves the result, Eq. (5.33), for a certain k, not for all $k \in \mathbb{N}^+$; and, (ii) it requires the assumption that $\|\mathbf{t}^{k-1} - \mathbf{t}^*\|_G^2 \leq F_1$, and $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F_1$, for some bounded constant F_1. Next, Proposition 3 shows that for all $k \geq K$, where K is some specific value defined later, if $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F$, and $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k$, we have that $\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 \leq F$.

Proposition 3. *Let Assumptions A4 and A5 hold, and recall the constants C_1, C_2, C_3, C_4, C_5, C_6, C_7, Δ, δ, and γ from Theorem 3. Assume that at kth*

iteration, $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F_2$*, for some bounded constant* F_2*, and* $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k$*. Then we have that*

$$\left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2 \leq F_2 \tag{5.81}$$

is satisfied for all $k \geq K$*, where* K *is defined as*

$$K = \left\lceil \log_r \left(\frac{\delta \lambda_{\min}\left(\frac{M+M^\top}{2}\right)}{2\alpha(d^-)^2 C_3} \right) \right\rceil. \tag{5.82}$$

Proof. Since we have $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \leq F_2$, and $\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k$, it follows that

$$\begin{aligned} \left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2 &\leq \frac{\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2}{1+\delta} + \frac{\Gamma\gamma^k}{1+\delta}, \\ &\leq \frac{F_2}{1+\delta} + \frac{\Gamma\gamma^k}{1+\delta}. \end{aligned} \tag{5.83}$$

Given the definition of K in Eq. (5.82), it follows that for $k \geq K$

$$\gamma^k \leq \frac{\delta \lambda_{\min}\left(\frac{M+M^\top}{2}\right) B^2}{2\alpha(d^-)^2 C_3 B^2} \leq \frac{\delta F_2}{\Gamma}, \tag{5.84}$$

where the second inequality follows with the definition of Γ, and F in Eqs. (5.45) and (5.43). Therefore, we obtain that

$$\left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2 \leq \frac{F_2}{1+\delta} + \frac{\delta F_2}{1+\delta} = F_2. \tag{5.85}$$

□

Proof of Theorem 3

We now formally state the proof of Theorem 3.

Proof. Define $F = \max_{1\leq k\leq K}\{\|\mathbf{t}^k - \mathbf{t}^*\|_G^2\}$, where K is the constant defined in Eq. (5.82). The goal is to show that Eq. (5.33) in Theorem 3 is valid for all k with the step-size being in the range defined in Eqs. (5.34) and (5.35).

We first prove the result for $k \in [1, ..., K]$: Since $\left\|\mathbf{t}^k - \mathbf{t}^*\right\|_G^2 \leq F$, $\forall k \in [1, ..., K]$, we use the result of Proposition 2 to have

$$\|\mathbf{t}^k - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k, \quad \forall k \in [1, ..., K].$$

Next, we use induction to show Eq. (5.33) for all $k \geq K$. For F defined above:

(i) Base case: when $k = K$, we have the initial relations that

$$\left\|\mathbf{t}^{K-1} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.86a}$$

$$\left\|\mathbf{t}^{K} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.86b}$$

$$\|\mathbf{t}^{K} - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{K+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^K. \tag{5.86c}$$

(ii) We now assume that the induction hypothesis is true at the kth iteration, for some $k \geq K$, i.e.,

$$\left\|\mathbf{t}^{k-1} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.87a}$$

$$\left\|\mathbf{t}^{k} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.87b}$$

$$\|\mathbf{t}^{k} - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^k, \tag{5.87c}$$

and show that this set of equations also hold for $k+1$.

(iii) Given Eqs. (5.87b) and (5.87c), we obtain $\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 \leq F$ by applying Proposition 3. Therefore, by combining $\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 \leq F$ with (5.87b), we obtain that $\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+2} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^{k+1}$ by Proposition 2. To conclude, we obtain that

$$\left\|\mathbf{t}^{k} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.88a}$$

$$\left\|\mathbf{t}^{k+1} - \mathbf{t}^*\right\|_G^2 \leq F, \tag{5.88b}$$

$$\|\mathbf{t}^{k+1} - \mathbf{t}^*\|_G^2 \geq (1+\delta)\|\mathbf{t}^{k+2} - \mathbf{t}^*\|_G^2 - \Gamma\gamma^{k+1}. \tag{5.88c}$$

hold for $k+1$.

By induction, we conclude that this set of equations holds for all k, which completes the proof. □

5.5 Numerical Experiments

This section provides numerical experiments to study the convergence rate of DEXTRA for a least squares problem over a directed graph. The local objective functions in the least squares problems are strongly-convex. We compare the performance of DEXTRA with other algorithms suited to the case of directed graph: GP as defined by [61, 63–65], and D-DSD as defined by [79]. Our second experiment verifies the existence of $\alpha_{\min}$ and $\alpha_{\max}$, such that the proper step-size α is between $\alpha_{\min}$ and $\alpha_{\max}$. We also consider various net-

work topologies and weighting strategies to see how the eigenvalues of network graphs effect the interval of step-size, α. Convergence is studied in terms of the residual

$$re = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{z}_i^k - \mathbf{u}\right\|,$$

where $\mathbf{u}$ is the optimal solution. The distributed least squares problem is described as follows.

Each agent owns a private objective function, $\mathbf{h}_i = H_i\mathbf{x} + \mathbf{n}_i$, where $\mathbf{h}_i \in \mathbb{R}^{m_i}$ and $H_i \in \mathbb{R}^{m_i \times p}$ are measured data, $\mathbf{x} \in \mathbb{R}^p$ is unknown, and $\mathbf{n}_i \in \mathbb{R}^{m_i}$ is random noise. The goal is to estimate $\mathbf{x}$, which we formulate as a distributed optimization problem solving

$$\min\ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left\|H_i\mathbf{x} - \mathbf{h}_i\right\|.$$

We consider the network topology as the digraph shown in Fig. 5.1. We first apply the local degree weighting strategy, i.e., to assign each agent itself and its out-neighbors equal weights according to the agent's own out-degree, i.e.,

$$a_{ij} = \frac{1}{|\mathcal{N}_j^{\text{out}}|}, \qquad (i,j) \in \mathcal{E}. \tag{5.89}$$

According to this strategy, the corresponding network parameters are shown in Fig. 5.2. We now estimate the interval of appropriate step-sizes. We choose $L_f = \max_i\{2\lambda_{\max}(H_i^\top H_i)\} = 0.14$, and $S_f = \min_i\{2\lambda_{\min}(H_i^\top H_i)\} = 0.1$. We set $\eta = 0.04 < S_f/d^2$, and $\delta = 0.1$. Note that η and δ are estimated values.

According to the calculation, we have $C_1 = 36.6$ and $C_2 = 5.6$. Therefore, we estimate that $\overline{\alpha} = \frac{\eta\lambda_{min}\left(M+M^\top\right)}{2L_f^2(d_\infty^- d^-)^2} = 0.26$, and $\underline{\alpha} < \frac{S_f/(2d^2)-\eta/2}{2C_2\delta} = 9.6 \times 10^{-4}$. We thus pick $\alpha = 0.1 \in [\underline{\alpha},\ \overline{\alpha}]$ for the following experiments.

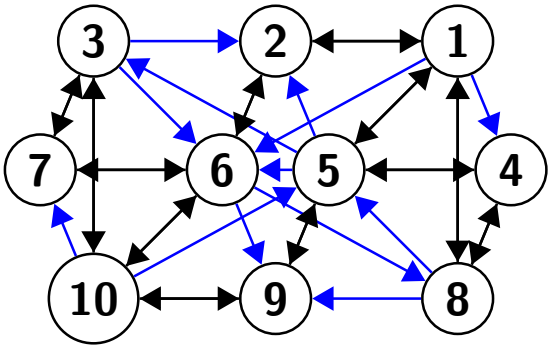


Figure 5.1: Strongly-connected but non-balanced digraphs.

$\lambda_{min}(M + M^\top)$	0.98
d	1.54
d^-	1.38
d_∞^-	1.38

Figure 5.2: The calculated network parameters.

Our first experiment compares several algorithms suited to directed graphs, illustrated in Fig. 5.1. The comparison of DEXTRA, GP, D-DSD and DGD with weighting matrix being row-stochastic is shown in Fig. 5.3. In this exper-

iment, we set $\alpha = 0.1$, which is in the range of our theoretical value calculated above. The convergence rate of DEXTRA is linear as stated in Section 6.3. G-P and D-DSD apply the same step-size, $\alpha = \frac{\alpha}{\sqrt{k}}$. As a result, the convergence rate of both is sub-linear. We also consider the DGD algorithm, but with the weighting matrix being row-stochastic. The reason is that in a directed graph, it is impossible to construct a doubly-stochastic matrix. As expected, DGD with row-stochastic matrix does not converge to the exact optimal solution while other three algorithms are suited to directed graphs.

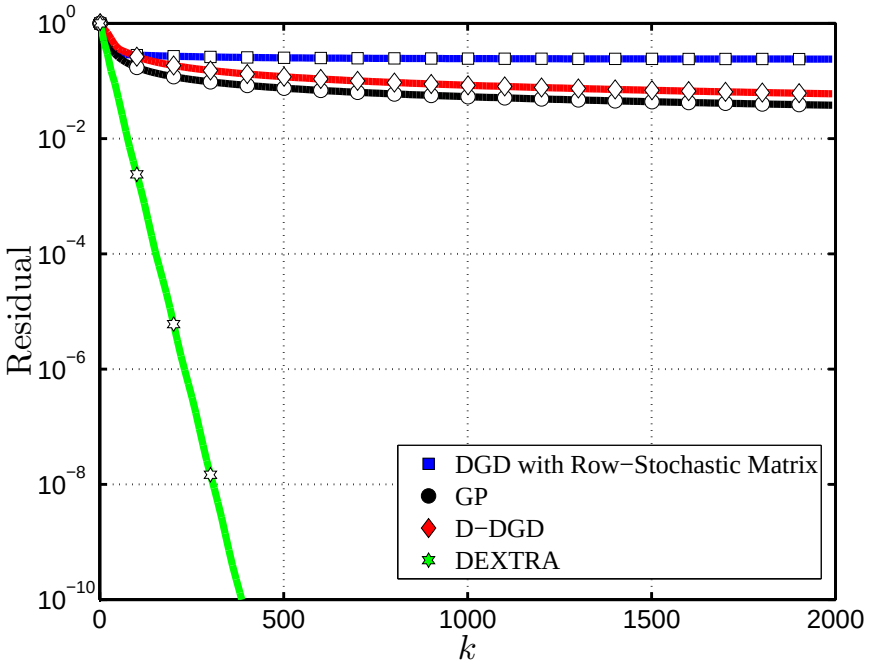


Figure 5.3: Convergence rate comparison between DEXTRA, GP, and D-DSD in a least squares problem over directed graphs.

According to the theoretical value of $\underline{\alpha}$ and $\overline{\alpha}$, we are able to set available

step-size, $\alpha \in [9.6 \times 10^{-4}, 0.26]$. In practice, this interval is much wider. Fig. 5.4 illustrates this fact. Numerical experiments show that $\alpha_{\min} = 0^+$ and $\alpha_{\max} = 0.447$. Though DEXTRA has a much wider range of step-size compared with the theoretical value, it still has a more restricted step-size compared with EXTRA, see [48], where the value of step-size can be as low as any value close to zero in any network topology, i.e., $\alpha_{\min} = 0$, as long as a *symmetric* and doubly-stochastic matrix is applied in EXTRA. The relative smaller range of interval is due to the fact that the weighting matrix applied in DEXTRA can not be symmetric.

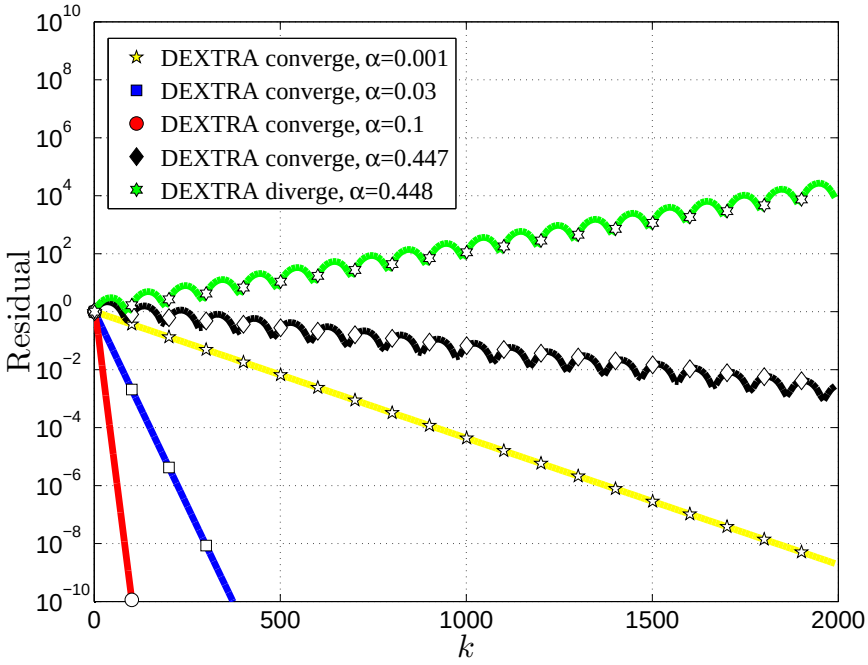


Figure 5.4: DEXTRA convergence w.r.t. different step-sizes.

The explicit representation of $\overline{\alpha}$ and $\underline{\alpha}$ given in Theorem **??** imply the way to increase the interval of step-size, i.e.,

$$\overline{\alpha} \propto \frac{\lambda_{min}(M+M^\top)}{(d_\infty^- d^-)^2}, \qquad \underline{\alpha} \propto \frac{1}{(d^- d)^2}.$$

To increase $\overline{\alpha}$, we increase $\frac{\lambda_{min}(M+M^\top)}{(d_\infty^- d^-)^2}$; to decrease $\underline{\alpha}$, we can decrease $\frac{1}{(d^- d)^2}$. Compared with applying the local degree weighting strategy, Eq. (5.89), as shown in Fig. 5.4, we achieve a wider range of step-sizes by applying the constant weighting strategy, which can be expressed as

$$a_{ij} = \begin{cases} 1 - 0.01(|\mathcal{N}_j^{\text{out}}| - 1), & i = j, \\ 0.01, & i \neq j, \quad i \in \mathcal{N}_j^{\text{out}}, \end{cases} \qquad \forall j,$$

This constant weighting strategy constructs a *diagonal-dominant* weighting matrix, which increases $\frac{\lambda_{min}(M+M^\top)}{(d_\infty^- d^-)^2}$. It may also be observed from Figs. 5.4 and 5.5 that the same step-size generates quiet different convergence speed when the weighting strategy changes. Comparing Figs. 5.4 and 5.5 when step-size $\alpha = 0.1$, DEXTRA with local degree weighting strategy converges much faster.

5.6 Conclusions

In this chapter, we introduce DEXTRA, a distributed algorithm to solve multi-agent smooth optimization problems over *directed* graphs. We have shown

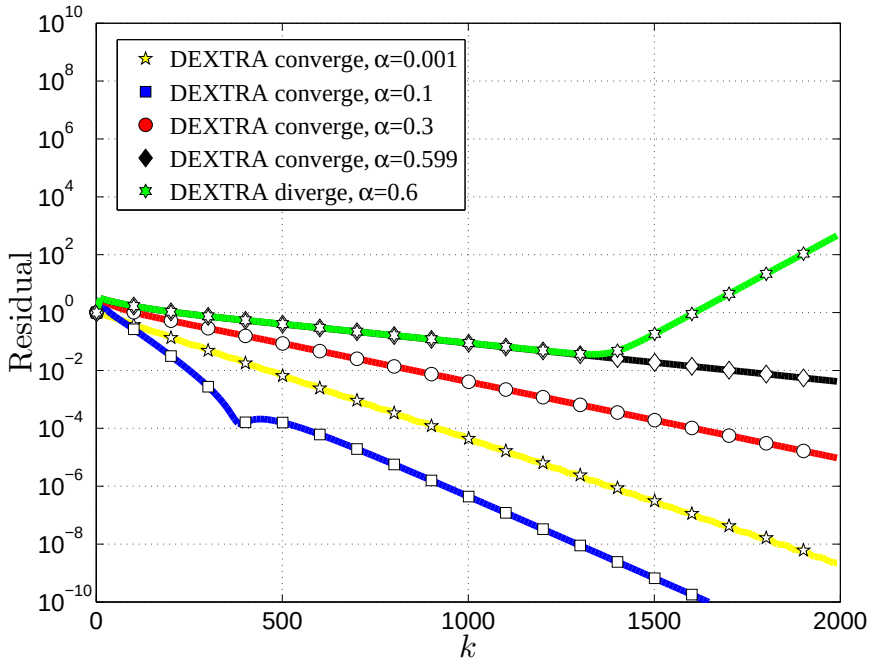


Figure 5.5: DEXTRA convergence using the constant weighting strategy.

that DEXTRA succeeds in driving all agents to the same point, which is the exact optimal solution of the problem, given that the communication graph is strongly-connected and the objective functions are strongly-convex. Moreover, the algorithm converges at a linear rate $O(\tau^k)$ for some constant, $\tau < 1$. This is the best known rate of convergence for this class of problems. The fast convergence rate is achieved because we harness the function smoothness. Numerical experiments on a least squares problem show that DEXTRA is the fastest distributed algorithm among all algorithms applicable to directed graphs.

Chapter 6

ADD-OPT for Smooth Convex Optimization

In this chapter, we propose, *ADD-OPT: Accelerated Distributed Directed OPTimization*, to solve the distributed smooth optimization problem, P1, over directed networks. ADD-OPT achieves the best known rate of convergence for this class of problems, $O(\mu^k)$ for $0 < \mu < 1$ given that the objective functions are strongly-convex, where k is the number of iterations. Compared with DEXTRA, ADD-OPT supports a wider and more realistic range of step-size. In particular, the greatest lower bound of DEXTRA's step-size is strictly greater than zero while that of ADD-OPT's equals exactly to zero. Simulation examples further illustrate the improvements.

6.1 Motivation

In previous chapter, we propose DEXTRA, a fast distributed algorithm over directed network for smooth optimization. By combining the push-sum protocol, [66, 67], and EXTRA, [48], DEXTRA achieves a linear convergence rate given that the objective functions are strongly-convex. However, one drawback of DEXTRA is the restrictive range of step-size. The greatest lower bound of DEXTRA's step-size is strictly greater than zero. In particular, a proper step-size, α, for DEXTRA converging to the optimal solution lies in $\alpha \in (\underline{\alpha}, \overline{\alpha})$, where $\underline{\alpha}$ and $\overline{\alpha}$ denote the lower and upper bound respectively. It is true that $\underline{\alpha} > 0$. Considering that it is hard to estimate $\underline{\alpha}$ in a distributed setting because the expression of $\underline{\alpha}$ requires the global knowledge, it is always an open problem on how to pick a proper step-size, α, in DEXTRA to guarantee the convergence, i.e., $\alpha \in (\underline{\alpha}, \overline{\alpha})$. In contrast if $\underline{\alpha} = 0$, agents can pick whatever small value for α to ensure the convergence.

Therefore, we propose ADD-OPT, aiming to relax the range of step-size while keep achieving a linear convergence rate when the objective functions are strongly-convex. Compared to DEXTRA, ADD-OPT's step-size, α, lies in $\alpha \in (0, \overline{\alpha})$, i.e., $\underline{\alpha} = 0$. This guarantees ADD-OPT to be a more reliable algorithm in distributed setting. We show that, after some derivation, the ADD-OPT has a similar representation as DEXTRA. In this point of view, it

can be regarded as an extended result of DEXTRA.

6.2 ADD-OPT Development

In this section, we first describe the implementation of ADD-OPT. We derive an informal but intuitive proof showing that ADD-OPT pushes the agents to achieve consensus and reach the optimal solution of Problem P1. After proposing ADD-OPT, we derive it to a similar representation of DEXTRA to show the relations between the two. The analysis also helps to reveal how to increase the range of step-size in DEXTRA by adjusting the weighting matrices. We also analyze the performance of both ADD-OPT and DEXTRA when the step-size is zero. Formal convergence results are left to later sections.

ADD-OPT Algorithm

To solve Problem P1, we describe the implementation of ADD-OPT as follows. Each agent, $j \in \mathcal{V}$, maintains three vector variables: $\mathbf{x}_{k,j}$, $\mathbf{z}_{k,j}$, $\mathbf{w}_{k,j} \in \mathbb{R}^p$, as well as a scalar variable, $y_{k,j} \in \mathbb{R}$, where k is the discrete-time index. At kth iteration, agent j weights its states, $a_{ij}\mathbf{x}_{k,j}$, $a_{ij}y_{k,j}$, as well as $a_{ij}\mathbf{w}_{k,j}$, and sends these to each of its out-neighbors, $i \in \mathcal{N}_j^{\text{out}}$, where the weights, a_{ij}'s are such

that:

$$a_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otw.}, \end{cases} \qquad \sum_{i=1}^{n} a_{ij} = 1, \forall j. \tag{6.1}$$

With agent i receiving the information from its in-neighbors, $j \in \mathcal{N}_i^{\text{in}}$, it updates $\mathbf{x}_{k+1,i}$, $y_{k+1,i}$, $\mathbf{z}_{k+1,i}$ and $\mathbf{w}_{k+1,i}$ as follows:

$$\mathbf{x}_{k+1,i} = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \mathbf{x}_{k,j} - \alpha \mathbf{w}_{k,i}, \tag{6.2a}$$

$$y_{k+1,i} = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} y_{k,j}, \tag{6.2b}$$

$$\mathbf{z}_{k+1,i} = \frac{\mathbf{x}_{k+1,i}}{y_{k+1,i}}, \tag{6.2c}$$

$$\mathbf{w}_{k+1,i} = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \mathbf{w}_{k,j} + \nabla f_i(\mathbf{z}_{k+1,i}) - \nabla f_i(\mathbf{z}_{k,i}). \tag{6.2d}$$

In the above, $\nabla f_i(\mathbf{z}_{k,i})$ in the gradient of the function $f_i(\mathbf{z})$ at $\mathbf{z} = \mathbf{z}_{k,i}$, for all $k \geq 0$. The step-size, α, is a positive number within a certain interval. We will explicitly show the range of α later. For any agent i, it is initiated with an arbitrary vector, $\mathbf{x}_{0,i}$, and with $\mathbf{w}_{0,i} = \nabla f_i(\mathbf{z}_{0,i})$ and $y_{0,i} = 1$. We note that the implementation of Eq. (6.2) needs each agent to at least have the knowledge of its out-neighbors degree. See [61, 63–65, 74, 79, 82, 83] for the similar assumptions.

To simplify the analysis, we assume from now on that all sequences updated by Eq. (6.2) have only one dimension, i.e., $p = 1$; thus $x_{k,i}$, $y_{k,i}$, $w_{k,i}$, $z_{k,i} \in \mathbb{R}, \forall i, k$. For $\mathbf{x}_{k,i}$, $\mathbf{w}_{k,i}$, $\mathbf{z}_{k,i} \in \mathbb{R}^p$ being p-dimensional vectors, the

proof is the same for every dimension by applying the results to each coordinate. Therefore, assuming $p = 1$ is without the loss of generality. We next write Eq. (6.2) in a matrix form. Define $\mathbf{x}_k,\ \mathbf{y}_k,\ \mathbf{w}_k,\ \mathbf{z}_k,\ \nabla\mathbf{f}_k \in \mathbb{R}^n$ as $\mathbf{x}_k = [x_{k,1}, \cdots, x_{k,n}]^\top$, $\mathbf{y}_k = [y_{k,1}, \cdots, y_{k,n}]^\top$, $\mathbf{w}_k = [w_{k,1}, \cdots, w_{k,n}]^\top$, $\mathbf{z}_k = [z_{k,1}, \cdots, z_{k,n}]^\top$, and $\nabla\mathbf{f}_k = [\nabla f_1(z_{k,1}), \cdots, \nabla f_n(z_{k,n})]^\top$. Let $A = \{a_{ij}\} \in \mathbb{R}^{n\times n}$ be the collection of weights a_{ij}. It is clear that A is a column-stochastic matrix. Define a diagonal matrix, $Y_k \in \mathbb{R}^{n\times n}$, for each k, as follow

$$Y_k = \operatorname{diag}(\mathbf{y}_k). \tag{6.3}$$

Given that the graph, $\mathcal{G}$, is strongly-connected and the corresponding weighting matrix, A, is non-negative, it follows that Y_k is invertible for any k. Then, we can write Eq. (6.2) in the matrix form equivalently as follows:

$$\mathbf{x}_{k+1} = A\mathbf{x}_k - \alpha\mathbf{w}_k, \tag{6.4a}$$

$$\mathbf{y}_{k+1} = A\mathbf{y}_k, \tag{6.4b}$$

$$\mathbf{z}_{k+1} = Y_{k+1}^{-1}\mathbf{x}_{k+1}, \tag{6.4c}$$

$$\mathbf{w}_{k+1} = A\mathbf{w}_k + \nabla\mathbf{f}_{k+1} - \nabla\mathbf{f}_k, \tag{6.4d}$$

where similarly we have the initial condition $\mathbf{w}_0 = \nabla\mathbf{f}_0$.

Interpretation of ADD-OPT

Based on Eq. (6.4), we now give an intuitive interpretation on the convergence of ADD-OPT to the optimal solution. By combining Eqs. (6.4a) and (6.4d), we obtain that

$$\begin{aligned}\mathbf{x}_{k+1} &= A\mathbf{x}_k - \alpha\left[A\mathbf{w}_{k-1} + \nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}\right] \\ &= A\mathbf{x}_k - \alpha A\left[\frac{A\mathbf{x}_{k-1} - \mathbf{x}_k}{\alpha}\right] - \alpha\left[\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}\right] \\ &= 2A\mathbf{x}_k - A^2\mathbf{x}_{k-1} - \alpha\left[\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}\right]. \end{aligned} \tag{6.5}$$

Assume that the sequences generated by Eq. (6.4) converge to their limits (not necessarily true), denoted by $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{w}_\infty$, $\mathbf{z}_\infty$, $\nabla\mathbf{f}_\infty$, respectively. It follows from Eq. (6.5) that

$$\mathbf{x}_\infty = 2A\mathbf{x}_\infty - A^2\mathbf{x}_\infty - \alpha\left[\nabla\mathbf{f}_\infty - \nabla\mathbf{f}_\infty\right], \tag{6.6}$$

which implies that $(I_n - A)^2\mathbf{x}_\infty = \mathbf{0}_n$, or $\mathbf{x}_\infty \in \text{span}\{\mathbf{y}_\infty\}$, considering that $\mathbf{y}_\infty = A\mathbf{y}_\infty$. Therefore, we obtain that

$$\mathbf{z}_\infty = Y_\infty^{-1}\mathbf{x}_\infty \in \text{span}\{\mathbf{1}_n\}, \tag{6.7}$$

where the consensus is reached.

By summing up the updates in Eq. (6.5) over k from 0 to ∞, we obtain that

$$\mathbf{x}_\infty = A\mathbf{x}_\infty + \sum_{r=1}^{\infty}(A - I_n)\mathbf{x}_r - \sum_{r=0}^{\infty}(A^2 - A)\mathbf{x}_r - \alpha\nabla\mathbf{f}_\infty.$$

Noting that $\mathbf{x}_\infty = A\mathbf{x}_\infty$, it follows

$$\alpha\nabla\mathbf{f}_\infty = \sum_{r=1}^{\infty}(A - I_n)\mathbf{x}_r - \sum_{r=0}^{\infty}(A^2 - A)\mathbf{x}_r.$$

Therefore, we obtain that

$$\alpha\mathbf{1}_n^\top\nabla\mathbf{f}_\infty = \mathbf{1}_n^\top(A - I_n)\sum_{r=1}^{\infty}\mathbf{x}_r - \mathbf{1}_n^\top(A^2 - A)\sum_{r=0}^{\infty}\mathbf{x}_r = 0,$$

which is the optimality condition of Problem P1 considering that $\mathbf{z}_\infty \in \text{span}\{\mathbf{1}_n\}$. To conclude, if we assume the sequences updated in Eq. (6.4) have limits, $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{w}_\infty$, $\mathbf{z}_\infty$, $\nabla\mathbf{f}_\infty$, we have the fact that $\mathbf{z}_\infty$ achieves consensus and reaches the optimal solution of Problem P1. We next discuss the relations between ADD-OPT and DEXTRA.

ADD-OPT and DEXTRA

We consider DEXTRA to solve the corresponding distributed optimization problem over directed graphs. It achieves a linear convergence rate given that the objective functions are strongly-convex. At kth iteration, each agent i keeps and updates three states, $x_{k,i}$, $y_{k,i}$, and $z_{k,i}$. The iteration, in matrix form, is shown as follow.

$$\mathbf{x}_{k+1} = (I_n + A)\,\mathbf{x}_k - \widetilde{A}\mathbf{x}_{k-1} - \alpha\left[\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}\right], \tag{6.8a}$$

$$\mathbf{y}_{k+1} = A\mathbf{y}_k, \tag{6.8b}$$

$$\mathbf{z}_{k+1} = Y_{k+1}^{-1}\mathbf{x}_{k+1}, \tag{6.8c}$$

where $\widetilde{A}$ is a column-stochastic matrix satisfying that $\widetilde{A} = \theta I_n + (1-\theta)A$ with any $\theta \in (0, \frac{1}{2}]$, and all other notations stick to the same definition appeared earlier in the chapter.

By comparing Eqs. (6.5) and (6.8a), (6.4b) and (6.8b), and (6.4c) and (6.8c), it follows that the only difference between ADD-OPT and DEXTRA lies in the weighting matrices used when updating $\mathbf{x}_k$. From DEXTRA to ADD-OPT, we change $(I_n + A)$ in (6.8a) to $2A$ in (6.5), and $\widetilde{A}$ to A^2, respectively. Mathematically, if $A = I_n$, the two algorithms become the same. Therefore, ADD-OPT can be regarded as an extended version of DEXTRA, in that we will show later that it has a wider range of step-size compared to DEXTRA, i.e., the greatest lower bound, $\underline{\alpha}$, of ADD-OPT's step-size is zero while that of DEXTRA's is positve. This also reveals the reason why in DEXTRA we prefer constructing A to be an extremely diagonal dominant matrix (see Assumption A5(C) in the previous chapter). The more similar A is to I_n, the closer $\underline{\alpha}$ approaches zero. However, in DEXTRA $\underline{\alpha}$ can never reach zero since A can not be the identity, I_n, which otherwise means there is no communication between agents. Therefore, ADD-OPT can not be regarded as a special case of DEXTRA since $A \neq I_n$. In this chapter, we provide a totally different, but much more compact and elegant proof, compared to DEXTRA's proof, to show the linear convergence rate of ADD-OPT.

Note that the implementation of Eq. (6.5) requires agents to communicate

with the neighbors of their own neighbors (because of A^2), which takes two iterations for each agent. This explains why ADD-OPT needs to keep and update 4 variables, $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{z}_k$, compared with DEXTRA which only have 3 variables, $\mathbf{x}_k$, $\mathbf{y}_k$, and $\mathbf{z}_k$. It can be considered as a tradeoff of between increasing the step-size range and decreasing the number of variables.

Interpretation of Algorithms when $\alpha = 0$

For any gradient-based method, let $\alpha = 0$ is mathematically equivalent to $\nabla \mathbf{f}_k = \mathbf{0}_n$ for all k, which means all local objective functions are constants. Thus, the methods are simplified to have agents reach consensus only. We now discuss whether DEXTRA can push all agents to consensus if we force the step-size to be zero. Assume that the sequences generated by Eq. (6.8) converge to their limits (not necessarily true), denoted by $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{z}_\infty$, and $\nabla \mathbf{f}_\infty$, respectively. Consider Eq. (6.8a) when $\alpha = 0$,

$$\mathbf{x}_{k+1} = (I_n + A)\mathbf{x}_k - \widetilde{A}\mathbf{x}_{k-1}. \tag{6.9}$$

By summing Eq. (6.9) over k from 0 to ∞, we obtain that

$$\mathbf{x}_\infty = A\mathbf{x}_\infty - \sum_{r=0}^{\infty}(\widetilde{A} - A)\mathbf{x}_r. \tag{6.10}$$

According to the previous analysis from Eqs. (6.6) to (6.7), we know that in order to reach consensus, i.e., $\mathbf{z}_\infty \in \text{span}\{\mathbf{1}_n\}$, it is equivalent to satisfy

$\mathbf{x}_\infty \in \text{span}\{\mathbf{y}_\infty\}$, which results in

$$\sum_{r=0}^{\infty}(\widetilde{A}-A)\mathbf{x}_r = \mathbf{0}_n. \tag{6.11}$$

Since $\mathbf{x}_0$ is arbitrary, and only $\mathbf{x}_\infty$ satisfying that $\mathbf{x}_\infty = A\mathbf{x}_\infty$ or $(\widetilde{A}-A)\mathbf{x}_\infty = \mathbf{0}_n$, we have that $\sum_{r=0}^{\infty}(\widetilde{A}-A)\mathbf{x}_r \neq \mathbf{0}_n$. This reaches a contradiction. Therefore, DEXTRA does not push all agents to the consensus when $\alpha = 0$.

We now consider the performance of ADD-OPT when $\alpha = 0$. We still assume that the sequences generated by Eq. (6.4) converge to their limits (not necessarily true), denoted by $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{w}_\infty$, $\mathbf{z}_\infty$, $\nabla\mathbf{f}_\infty$, respectively. In fact, it is straightforward to observe that Eq. (6.4) with having $\alpha = 0$ is exactly the push-sum consensus, [66, 67], which push agents to reach average consensus in a directed graph. Therefore, ADD-OPT converges to the optimal solution when $\alpha = 0$. To better compare ADD-OPT with DEXTRA, we analyze ADD-OPT using similar derivations for DEXTRA above. By summing up the updates in Eq. (6.5) over k from 0 to ∞, we obtain that

$$\mathbf{x}_\infty = A\mathbf{x}_\infty + A(I_n - A)\mathbf{x}_0 - \sum_{r=1}^{\infty}(I_n - A)^2\mathbf{x}_r, \tag{6.12}$$

which is sufficient to satisfy that

$$\sum_{r=1}^{\infty}(A - I_n)\mathbf{x}_r = A\mathbf{x}_0, \tag{6.13}$$

by considering the condition that $\mathbf{x}_\infty = A\mathbf{x}_\infty$. Compared Eqs. (6.13) from the derivation of ADD-OPT with (6.11) from DEXTRA, we say that ADD-OPT

works when $\alpha = 0$ because there exists a additional term $A\mathbf{x}_0$. The infinit sum $\sum_{r=1}^{\infty}(A - I_n)\mathbf{x}_r$ is accumulated to compensate this initial term $A\mathbf{x}_0$. In the next section, we state the convergence result with appropriate assumptions.

6.3 Assumptions and Main Result

With appropriate assumptions, our main result states that ADD-OPT converges to the optimal solution of Problem P1 linearly. We state again that from now on we assume that the states of agents have only one dimension, i.e., $p = 1$, which is without the loss of generality. In this paper, we assume that the agent graph, $\mathcal{G}$, is strongly-connected; each local function, $f_i(z)$, is convex and differentiable, and the optimal solution of Problem P1 and the corresponding optimal value exist. Formally, we denote the optimal solution by z^* and optimal value by f^*, i.e., $f^* = f(z^*) = \min_{z\in\mathbb{R}} f(z)$. Besides the above assumptions, we formally present assumptions regarding the gradients of objective functions as follows, which is standard for smooth convex functions, [46, 48, 83],

Assumption A6 (Lipschitz continuous gradients and strong convexity)**.** *Each private function, f_i, is differentiable and strongly-convex, and the gradient is Lipschitz continuous, i.e., for any i and $z_1, z_2 \in \mathbb{R}$,*

(a) there exists a positive constant l such that,

$$\|\nabla f_i(z_1) - \nabla f_i(z_2)\| \leq l\|z_1 - z_2\|;$$

(b) there exists a positive constant s such that,

$$s\|z_1 - z_2\|^2 \leq \langle \nabla f_i(z_1) - \nabla f_i(z_2), z_1 - z_2 \rangle.$$

With these assumptions, we are able to present ADD-OPT's convergence result, the representation of which are based on the following notations. Based on earlier notations, $\mathbf{x}_k$, $\mathbf{w}_k$, and $\nabla \mathbf{f}_k$, we further define $\overline{\mathbf{x}}_k$, $\overline{\mathbf{w}}_k$, $\mathbf{z}^*$, $\mathbf{g}_k$, $\mathbf{h}_k \in \mathbb{R}^n$ as

$$\overline{\mathbf{x}}_k = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{x}_k, \tag{6.14}$$

$$\overline{\mathbf{w}}_k = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{w}_k, \tag{6.15}$$

$$\mathbf{z}^* = z^*\mathbf{1}_n, \tag{6.16}$$

$$\mathbf{g}_k = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\nabla\mathbf{f}_k, \tag{6.17}$$

$$\mathbf{h}_k = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\nabla\mathbf{f}(\overline{\mathbf{x}}_k), \tag{6.18}$$

where $\nabla\mathbf{f}(\overline{\mathbf{x}}_k) = [\nabla f_1(\frac{1}{n}\mathbf{1}_n^\top\mathbf{x}_k), ..., \nabla f_n(\frac{1}{n}\mathbf{1}_n^\top\mathbf{x}_k)]^\top$. We denote constants τ, ϵ, and η as

$$\tau = \|A - I_n\|, \tag{6.19}$$

$$\epsilon = \|I_n - A_\infty\|, \tag{6.20}$$

$$\eta = \max\left(\left|1-\alpha l\right|, \left|1-\alpha s\right|\right), \tag{6.21}$$

where A is the column-stochastic weighting matrix used in Eq. (6.4), $A_\infty = \lim_{k\to\infty} A^k$ represents A's limit, α is the step-size, and l and s are respectively the Lipschitz gradient constant and strong-convexity constant in Assumption A6. Let Y_∞ be the limit of Y_k in Eq. (6.3),

$$Y_\infty = \lim_{k\to\infty} Y_k, \tag{6.22}$$

and y and y_- be the maximum of $\|Y_k\|$ and $\|Y_k^{-1}\|$ over k, respectively,

$$y = \max_k \|Y_k\|, \tag{6.23}$$

$$y_- = \max_k \left\|Y_k^{-1}\right\|. \tag{6.24}$$

Moreover, we define two constants, σ, and, γ_1, through the following two lemmas, which is related to the convergence of A and Y_∞. Note that Lemmas 17 and 18 are reformulated with notations introduced above to simplify the proof.

Lemma 17. *(Nedic et al. [61]) Consider Y_k, generated from the column-stochastic matrix, A, and its limit Y_∞. There exist $0 < \gamma_1 < 1$ and $0 < T < \infty$ such that for all k*

$$\|Y_k - Y_\infty\| \leq T\gamma_1^k. \tag{6.25}$$

Lemma 18. *(Olshevsky et al. [**olshevsky**]) Consider Y_∞ in Eq. (6.22), and A being the column-stochastic matrix used in Eq. (6.4). For any $\mathbf{a} \in \mathbb{R}^n$, define $\overline{\mathbf{a}} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{a}$. There exist $0 < \sigma < 1$ such that for all k*

$$\|A\mathbf{a} - Y_\infty\overline{\mathbf{a}}\| \leq \sigma \|\mathbf{a} - Y_\infty\overline{\mathbf{a}}\|. \tag{6.26}$$

Based on the above notations, we finally denote $\mathbf{t}_k$, $\mathbf{s}_k \in \mathbb{R}^3$, and G, $H_k \in \mathbb{R}^{3\times 3}$ for all k as

$$\mathbf{t}_k = \begin{bmatrix} \|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\| \\ \|\overline{\mathbf{x}}_k - \mathbf{z}^*\| \\ \|\mathbf{w}_k - Y_\infty\mathbf{g}_k\| \end{bmatrix}, \mathbf{s}_k = \begin{bmatrix} \|\mathbf{x}_k\| \\ 0 \\ 0 \end{bmatrix},$$

$$G = \begin{bmatrix} \sigma & 0 & \alpha \\ \alpha(ly_-) & \eta & 0 \\ \epsilon l\tau y_- + \alpha(\epsilon l^2 y y_-^2) & \alpha(\epsilon l^2 y y_-) & \sigma + \alpha(\epsilon l y_-) \end{bmatrix},$$

$$H_k = \begin{bmatrix} 0 & 0 & 0 \\ \alpha l y_- T\gamma_1^k & 0 & 0 \\ (\alpha l y + 2)\epsilon l y_-^2 T\gamma_1^k & 0 & 0 \end{bmatrix}, \tag{6.27}$$

We now state the key relation in this chapter.

Theorem 5. *Let Assumption 6.2 holds. The following inequality holds for all $k \geq 1$,*

$$\mathbf{t}_k \leq G\mathbf{t}_{k-1} + H_{k-1}\mathbf{s}_{k-1}. \tag{6.28}$$

Proof. See the section of convergence analysis. □

Eq. (6.28) in Theorem 5 is the key relation of this paper. We leave the complete proof in later sections, with the help of several auxiliary relations. Note that Eq. (6.28) provides a linear iterative relation between $\mathbf{t}_k$ and $\mathbf{t}_{k-1}$ with matrix G and H_k. Thus, the convergence of $\mathbf{t}_k$ is fully determined by G and H_k. More specifically, if we want to prove a linear convergence rate of $\|\mathbf{t}_k\|$ to zero, it is sufficient to show that $\rho(G) < 1$, where $\rho(\cdot)$ denotes the spectral radius, as well as the linear decaying of H_k, which is straightforward since $0 < \gamma_1 < 1$. In Lemma 19, we first show that with appropriate step-size, the spectral radius of G is less than 1. Following Lemma 19, we show the linear convergence rate of G^k and H_k in Lemma 20.

Lemma 19. *Consider the matrix, G_α, defined in Eq.* (6.27) *as a function of the step-size, α. It follows that $\rho(G_\alpha) < 1$ if the step-size, $\alpha \in (0, \overline{\alpha})$, where*

$$\overline{\alpha} = \frac{\sqrt{(\epsilon\tau s)^2 + 4\epsilon y(l+s)s(1-\sigma)^2} - \epsilon\tau s}{2\epsilon l y y_-(l+s)}. \tag{6.29}$$

Proof. It is easy to verify that $\overline{\alpha} \leq \frac{\sqrt{4\epsilon y(l+s)s(1-\sigma)^2}}{2\epsilon l y y_-(l+s)} < \frac{1}{l}$. As a result, we have $\eta = 1 - \alpha s$. When $\alpha = 0$, we have that

$$G_0 = \begin{bmatrix} \sigma & 0 & 0 \\ 0 & 1 & 0 \\ \epsilon l \tau y_- & 0 & \sigma \end{bmatrix}, \tag{6.30}$$

the eigenvalues of which are σ, σ, and 1. Therefore, $\rho(G_0) = 1$, where $\rho(\cdot)$ denotes the spectral radius. We now consider how the eigenvalue 1 is changed if we slightly increase α from 0. We denote $\mathcal{P}_{\mathcal{G}_\alpha}(q) = \det(qI_n - G_\alpha)$ the characteristic polynomial of G_α. By letting $\det(qI_n - G_\alpha) = 0$, we get the following equation.

$$\begin{aligned}((q-\sigma)^2 - \alpha\epsilon l y_-(q-\sigma))(q-1+\alpha s) - \alpha^3 l^3 \epsilon y y_-^2 \\ -\alpha(q-1+\alpha s)(\epsilon l \tau y_- + \alpha(\epsilon l^2 y y_-^2)) = 0. \end{aligned} \tag{6.31}$$

Since we have already shown that 1 is one of the eigenvalues of G_0, Eq. (6.31) is valid when $q = 1$ and $\alpha = 0$. Take the derivative on both sides of Eq. (6.31), and let $q = 1$ and $\alpha = 0$, we obtain that $\frac{dq}{d\alpha}|_{\alpha=0,q=1} = -s < 0$. This is saying that when α increases from 0 slightly, $\rho(G_\alpha)$ will decrease first.

We now calculate all possible values of α for $\lambda(G_\alpha) = 1$. Let $q = 1$ in Eq. (6.31), and solve the step-size, α, we obtain that, $\alpha_1 = 0$, $\alpha_2 < 0$, and

$$\alpha_3 = \overline{\alpha} = \frac{\sqrt{(\epsilon\tau s)^2 + 4\epsilon y(l+s)s(1-\sigma)^2} - \epsilon\tau s}{2\epsilon l y y_-(l+s)}.$$

Since there is no other value of α for $\lambda(G_\alpha) = 1$, we know that all eigenvalues of G_α is less than 1 for $\alpha \in (0, \overline{\alpha})$ by considering the fact that eigenvalues are continuous functions of matrix. Therefore, $\rho(G_\alpha) < 1$ when $\alpha \in (0, \overline{\alpha})$. □

Lemma 20. *With the step-size, $\alpha \in (0, \overline{\alpha})$, where $\overline{\alpha}$ is defined in Eq. (6.29), the following statements hold for all k,*

(a) there exist $0 < \gamma_1 < 1$ *and* $0 < \Gamma_1 < \infty$*, where* γ_1 *is defined in Eq.* (6.25)*, such that*

$$\|H_k\| = \Gamma_1 \gamma_1^k;$$

(b) there exist $0 < \gamma_2 < 1$ *and* $0 < \Gamma_2 < \infty$*, such that*

$$\left\|G^k\right\| \leq \Gamma_2 \gamma_2^k;$$

(c) there exist $\gamma = \max\{\gamma_1, \gamma_2\}$ *and* $\Gamma = \Gamma_1\Gamma_2/\gamma$*, such that for all* $0 \leq r \leq k$*,*

$$\left\|G^{k-r-1} H_r\right\| \leq \Gamma \gamma^k.$$

Proof. (a). This is easy to verify according to Eq. (6.27), and by letting $\Gamma_1 = ly_{-}T\sqrt{\alpha^2 + (\alpha l y + 2)^2 \epsilon^2 y_{-}^2}$.

(b). We represent G^k in the Jordan canonical form as $G^k = PJ^kQ$. According to Lemma 19, we have that all diagonal entries in J are smaller than one. Therefore, there exist $0 < \Gamma_2 < \infty$ and $0 < \gamma_2 < 1$, such that

$$\left\|G^k\right\| \leq \|P\| \, \|Q\| \left\|J^k\right\| \leq \Gamma_2 \gamma_2^k. \tag{6.32}$$

(c). The proof of (c) is achieved by combining (a) and (b). □

We now present the main result of this paper in Theorem 6, which shows the linear convergence rate of ADD-OPT.

Theorem 6. *Let Assumption 6.2 holds. With the step-size, $\alpha \in (0, \overline{\alpha})$, where $\overline{\alpha}$ is defined in Eq.* (6.29)*, the sequence, $\{\mathbf{z}_k\}$, generated by ADD-OPT, converges exactly to the optimal solution, $\mathbf{z}^*$, at a linear rate, i.e., there exist some bounded constants $M > 0$ and $\gamma < \mu < 1$, where γ is used in Lemma 20(c), such that for any k,*

$$\|\mathbf{z}_k - \mathbf{z}^*\| \leq M\mu^k. \tag{6.33}$$

Proof. We write Eq. (6.28) recursively, which results

$$\mathbf{t}_k \leq G^k \mathbf{t}_0 + \sum_{r=0}^{k-1} G^{k-r-1} H_r \mathbf{s}_r. \tag{6.34}$$

By taking the norm on both sides of Eq. (6.34), and considering Lemma 20, we obtain that

$$\begin{aligned} \|\mathbf{t}_k\| \leq & \left\|G^k\right\| \|\mathbf{t}_0\| + \sum_{r=0}^{k-1} \left\|G^{k-r-1} H_r\right\| \|\mathbf{s}_r\| \\ \leq & \Gamma_2 \gamma_2^k \|\mathbf{t}_0\| + \sum_{r=0}^{k-1} \Gamma \gamma^k \|\mathbf{s}_r\|, \end{aligned} \tag{6.35}$$

in which we can bound $\|\mathbf{s}_r\|$ as

$$\begin{aligned} \|\mathbf{s}_r\| \leq & \|\mathbf{x}_r - Y_\infty \overline{\mathbf{x}}_r\| + \|Y_\infty\| \|\overline{\mathbf{x}}_r - \mathbf{z}^*\| + \|Y_\infty\| \|\mathbf{z}^*\| \\ \leq & (1+y) \|\mathbf{t}_r\| + y \|\mathbf{z}^*\|. \end{aligned} \tag{6.36}$$

Therefore, we have that for all k

$$\|\mathbf{t}_k\| \leq \left(\Gamma_2 \|\mathbf{t}_0\| + \Gamma(1+y) \sum_{r=0}^{k-1} \|\mathbf{t}_r\| + \Gamma y k \|\mathbf{z}^*\| \right) \gamma^k. \tag{6.37}$$

Our first step is to show that $\|\mathbf{t}_k\|$ is bounded for all k. It is true that there exists some bounded $K > 0$ such that for all $k > K$ it satisfies that

$$(\Gamma_2 + \Gamma(1+2y)k)\,\gamma^k \leq 1. \tag{6.38}$$

Define $\Phi = \max_{0\leq k\leq K}(\|\mathbf{t}_k\|, \|\mathbf{z}_*\|)$, which is bounded since K is bounded. It is true that $\|\mathbf{t}_k\| \leq \Phi$ for $0 \leq k \leq K$. Consider the case when $k = K+1$. By combining Eqs. (6.37) and (6.38), we have that

$$\|\mathbf{t}_{K+1}\| \leq \Phi\Big(\Gamma_2 + \Gamma(1+2y)(K+1)\Big)\gamma^{K+1} \leq \Phi. \tag{6.39}$$

We repeat the procedures to show that $\|\mathbf{t}_k\| \leq \Phi$ for all k.

The next step is to show that $\|\mathbf{t}_k\|$ decays linearly. For any μ satisfying $\gamma < \mu < 1$, there exist a constant U such that $(\frac{\mu}{\gamma})^k > \frac{k}{U}$ for all k. Therefore, by bounding all $\|\mathbf{t}_k\|$ and $\|\mathbf{z}^*\|$ by Φ in Eq. (6.37), we obtain that for all k

$$\begin{aligned}
\|\mathbf{t}_k\| \leq& \Phi\Big(\Gamma_2 + \Gamma(1+2y)k\Big)\gamma^k \\
\leq& \Phi\left(\Gamma_2 + \Gamma(1+2y)U\frac{k}{U}\left(\frac{\gamma}{\mu}\right)^k\right)\mu^k \\
\leq& \Phi\Big(\Gamma_2 + \Gamma(1+2y)U\Big)\mu^k.
\end{aligned} \tag{6.40}$$

It follows that $\|\mathbf{z}_k - \mathbf{z}^*\|$ and $\|\mathbf{t}_k\|$ satisfy the relation that

$$\begin{aligned}
\|\mathbf{z}_k - \mathbf{z}^*\| \leq& \left\|Y_k^{-1}\mathbf{x}_k - Y_k^{-1}Y_\infty\overline{\mathbf{x}}_k\right\| + \left\|Y_k^{-1}Y_\infty\mathbf{z}^* - \mathbf{z}^*\right\| \\
&+ \left\|Y_k^{-1}Y_\infty\overline{\mathbf{x}}_k - Y_k^{-1}Y_\infty\mathbf{z}^*\right\| \\
\leq& y_-(1+y)\,\|\mathbf{t}_k\| + y_-T\gamma_1^k\,\|\mathbf{z}^*\|,
\end{aligned} \tag{6.41}$$

where in the second inequality we use the relation $\|Y_k^{-1}Y_\infty - I_n\| \leq \|Y_k^{-1}\|\|Y_\infty - Y_k\| \leq y_- T\gamma_1^k$ achieved from Eq. (6.25). By combining Eqs. (6.40) and (6.41), we obtain that

$$\|\mathbf{z}_k - \mathbf{z}^*\| \leq y_- \Phi \left[(1+y)(\Gamma_2 + \Gamma(1+2y)U) + T\right] \mu^k.$$

The desired result is obtained by letting $M = y_-\Phi[(1+y)(\Gamma_2 + \Gamma(1+2y)U) + T]$. □

Themorem 6 shows the linear convergence rate of ADD-OPT. In next twp sections, we prove Theorem 5 with the help of some auxiliary relations.

6.4 Auxiliary Relations

We provide several basic relations in this section, which will help the proof of Theorem 5. Lemma 21 derives iterative equations that govern the average sequence $\overline{\mathbf{x}}_k$ and $\overline{\mathbf{w}}_k$. Lemma 22 gives inequalities that are direct consequences of Eq. (6.25). Lemma 23 can be found in standard optimization literature, [91]. It states that if we perform a gradient descent step with a fixed step-size for a strongly convex and smooth function, then the distance to optimizer shrinks by at least a fixed ratio.

Lemma 21. *The following equations hold for all k,*

(a) $\overline{\mathbf{w}}_k = \mathbf{g}_k$;

(b) $\overline{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_k - \alpha \mathbf{g}_k$.

Proof. Since A is column-stochastic, satisfying $\mathbf{1}_n^\top A = \mathbf{1}_n^\top$, we obtain that

$$\begin{aligned}\overline{\mathbf{w}}_k &= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\left(A\mathbf{w}_{k-1} + \nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}\right)\\ &= \overline{\mathbf{w}}_{k-1} + \mathbf{g}_k - \mathbf{g}_{k-1}.\end{aligned}$$

Do this recursively, and we have that

$$\overline{\mathbf{w}}_k = \overline{\mathbf{w}}_0 + \mathbf{g}_k - \mathbf{g}_0.$$

Recall that we have the initial condition that $\mathbf{w}_0 = \nabla\mathbf{f}_0$, which is equivalently to $\overline{\mathbf{w}}_0 = \mathbf{g}_k$. Hence, we achieve the result of (a). The proof of (b) follows the following derivation,

$$\begin{aligned}\overline{\mathbf{x}}_{k+1} &= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\left(A\mathbf{x}_k - \alpha\mathbf{w}_k\right)\\ &= \overline{\mathbf{x}}_k - \alpha\overline{\mathbf{w}}_k,\\ &= \overline{\mathbf{x}}_k - \alpha\mathbf{g}_k,\end{aligned}$$

where the last equation use result of (a). The proof is done. □

Lemma 22. *The following inequalities hold for all* $k \geq 1$*,*

(a) $\left\|Y_{k-1}^{-1}Y_\infty - I_n\right\| \leq y_- T\gamma_1^{k-1}$*;*

(b) $\left\|Y_k^{-1} - Y_{k-1}^{-1}\right\| \leq 2y_-^2 T\gamma_1^{k-1}$.

Proof. By considering Eq. (6.25), it follows that

$$\left\|Y_{k-1}^{-1}Y_\infty - I_n\right\| \leq \left\|Y_{k-1}^{-1}\right\| \|Y_\infty - Y_{k-1}\| \leq y_- T\gamma_1^{k-1};$$

The proof of (b) follows

$$\begin{aligned}\left\|Y_k^{-1} - Y_{k-1}^{-1}\right\| &\leq \left\|Y_{k-1}^{-1}\right\| \||Y_{k-1} - Y_k\| \left\|Y_k^{-1}\right\| \\ &\leq 2y_-^2 T\gamma_1^{k-1}.\end{aligned}$$

This finishes the proof. □

Lemma 23. *Let Assumption A6 hold for the objective function, $f(z)$, in P1, and s and l are the strong-convexity constant and Lipschitz continuous gradient constant, respectively. For any $z \in \mathbb{R}$, define $z_+ = z - \alpha\nabla f(z)$, where $0 < \alpha < \frac{2}{l}$. Then*

$$\|z_+ - z^*\| \leq \eta \|z - z^*\|,$$

where $\eta = \max(|1 - \alpha l|, |1 - \alpha s|)$.

6.5 Convergence Analysis

The proof of Theorem 5 is provided in this section. We will bound $\|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\|$, $\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|$, and $\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\|$ by the linear combinations of their past values, i.e., $\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\|$, $\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|$, and $\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|$, as well as $\|\mathbf{x}_{k-1}\|$. The coefficients will be shown to be the entries of G and H_{k-1}.

Step 1: Bound $\|\mathbf{x}_k - Y_\infty \overline{\mathbf{x}}_k\|$.

According to Eq. (6.4a) and Lemma 21(b), we obtain that

$$\|\mathbf{x}_k - Y_\infty \overline{\mathbf{x}}_k\| \leq \|A\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\| + \alpha \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| . \tag{6.42}$$

Note that $\|A\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\| \leq \sigma \|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\|$ from Eq. (6.26), we have

$$\|\mathbf{x}_k - Y_\infty \overline{\mathbf{x}}_k\| \leq \sigma \|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\| + \alpha \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| . \tag{6.43}$$

Step 2: Bound $\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|$.

By considering Lemma 21(b), we obtain that

$$\overline{\mathbf{x}}_k = [\overline{\mathbf{x}}_{k-1} - \alpha \mathbf{h}_{k-1}] - \alpha [\mathbf{g}_{k-1} - \mathbf{h}_{k-1}] . \tag{6.44}$$

Let $\mathbf{x}_+ = \overline{\mathbf{x}}_{k-1} - \alpha \mathbf{h}_{k-1}$, which is performing a (centralized) gradient descent to minimize the objective function in Problem P1. Therefore, we have that, according to Lemma 23,

$$\|\mathbf{x}_+ - \mathbf{z}^*\| \leq \eta \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\| . \tag{6.45}$$

By applying the Lipschitz continuous gradient assumption, Assumption 6.2(a), we obtain

$$\|\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\| \leq \left\| \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right\| l \|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\| . \tag{6.46}$$

Therefore, it follows that

$$\|\overline{\mathbf{x}}_k - \mathbf{z}^*\| \leq \|\mathbf{x}_+ - \mathbf{z}^*\| + \alpha \|\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\|$$

$$\leq \eta \left\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\right\| + \alpha l \left\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\right\|. \tag{6.47}$$

Notice by Eq. (6.4c) and Lemma 22(a), it follows that

$$\begin{aligned}\left\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\right\| \leq& \left\|Y_{k-1}^{-1}\left(\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\right)\right\| + \left\|\left(Y_{k-1}^{-1} Y_\infty - I_n\right)\overline{\mathbf{x}}_{k-1}\right\| \\ \leq& y_- \left\|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\right\| + y_- T \gamma_1^{k-1} \left\|\mathbf{x}_{k-1}\right\|, \end{aligned} \tag{6.48}$$

where in the second inequality we also make use of the relation $\|\overline{\mathbf{x}}_{k-1}\| \leq \|\mathbf{x}_{k-1}\|$. By substituting Eq. (6.48) into Eq. (6.47), we obtain that

$$\begin{aligned}\left\|\overline{\mathbf{x}}_k - \mathbf{z}^*\right\| \leq& \alpha l y_- \left\|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\right\| + \eta \left\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\right\| \\ &+ \alpha l y_- T \gamma_1^{k-1} \left\|\mathbf{x}_{k-1}\right\|. \end{aligned} \tag{6.49}$$

Step 3: Bound $\|\mathbf{w}_k - Y_\infty \mathbf{g}_k\|$.

According to Eq. (6.4d), we have

$$\left\|\mathbf{w}_k - Y_\infty \mathbf{g}_k\right\| \leq \left\|A\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\right\| + \left\|\left(\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1}\right) - \left(Y_\infty \mathbf{g}_k - Y_\infty \mathbf{g}_{k-1}\right)\right\|.$$

With Lemma 21(a) and Eq. (6.26), we obtain that

$$\begin{aligned}\left\|A\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\right\| &= \left\|A\mathbf{w}_{k-1} - Y_\infty \overline{\mathbf{w}}_{k-1}\right\| \\ &\leq \sigma \left\|\mathbf{w}_{k-1} - Y_\infty \overline{\mathbf{w}}_{k-1}\right\|. \end{aligned} \tag{6.50}$$

It follows from the definition of $\mathbf{g}_k$ that

$$\left\|\left(\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1}\right) - \left(Y_\infty \mathbf{g}_k - Y_\infty \mathbf{g}_{k-1}\right)\right\| = \left\|\left(I_n - \frac{1}{n} Y_\infty \mathbf{1}_n \mathbf{1}_n^\top\right)\left(\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1}\right)\right\|. \tag{6.51}$$

Since $\frac{1}{n}Y_\infty \mathbf{1}_n \mathbf{1}_n^\top = A_\infty$, where $A_\infty = \lim_{k\to\infty} A^k$, we obtain that

$$\|(\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1}) - (Y_\infty \mathbf{g}_k - Y_\infty \mathbf{g}_{k-1})\| \leq \epsilon l \|\mathbf{z}_k - \mathbf{z}_{k-1}\|,$$

where in the preceding relation we use the Lipschitz continuous gradient assumption, Assumption A6(a). Therefore, we have

$$\|\mathbf{w}_k - Y_\infty \mathbf{g}_k\| \leq \sigma \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| + \epsilon l \|\mathbf{z}_k - \mathbf{z}_{k-1}\|. \tag{6.52}$$

We now bound $\|\mathbf{z}_k - \mathbf{z}_{k-1}\|$. Note that

$$\|\mathbf{h}_{k-1}\| = \left\| \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \nabla \mathbf{f}(\overline{\mathbf{x}}_{k-1}) \right\| \leq l \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|. \tag{6.53}$$

As a result, we have

$$\begin{aligned}
\left\|Y_k^{-1}\mathbf{w}_{k-1}\right\| \leq& \left\|Y_k^{-1}\left(\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\right)\right\| + \left\|Y_k^{-1} Y_\infty \mathbf{h}_{k-1}\right\| \\
&+ \left\|Y_k^{-1} Y_\infty \left(\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\right)\right\| \\
\leq& y_- \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| + y_- y l \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\| + y_- y l \|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\| \\
\leq& y_- \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| + y_- y l \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\| + y_-^2 y l \|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\| \\
&+ y_-^2 y l T \gamma_1^{k-1} \|\mathbf{x}_{k-1}\|,
\end{aligned} \tag{6.54}$$

where the last inequality is valid by considering Eq. (6.48). With the upper bound of $\|Y_k^{-1}\mathbf{w}_{k-1}\|$ provided in the preceding relation and note that $(A - I_n)Y_\infty \overline{\mathbf{x}}_{k-1} = \mathbf{0}_n$, we can bound $\|\mathbf{z}_k - \mathbf{z}_{k-1}\|$ as follow.

$$\|\mathbf{z}_k - \mathbf{z}_{k-1}\| \leq \left\|Y_k^{-1}\left(\mathbf{x}_k - \mathbf{x}_{k-1}\right)\right\| + \left\|\left(Y_k^{-1} - Y_{k-1}^{-1}\right)\mathbf{x}_{k-1}\right\|$$

$$\begin{aligned}
&\leq \left\|Y_k^{-1}\left(A-I_n\right)\mathbf{x}_{k-1}\right\| + \alpha\left\|Y_k^{-1}\mathbf{w}_{k-1}\right\| + \left\|Y_k^{-1}-Y_{k-1}^{-1}\right\|\|\mathbf{x}_{k-1}\| \\
&\leq (y_-\tau+\alpha y_-^2 yl)\|\mathbf{x}_{k-1}-Y_\infty\overline{\mathbf{x}}_{k-1}\| + \alpha y_-\|\mathbf{w}_{k-1}-Y_\infty\mathbf{g}_{k-1}\| \\
&\quad + \alpha y_- yl\|\overline{\mathbf{x}}_{k-1}-\mathbf{z}^*\| + (\alpha yl+2)y_-^2 T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|. \qquad (6.55)
\end{aligned}$$

By substituting Eq. (6.55) in Eq. (6.52), we obtain that

$$\begin{aligned}
\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\| \leq &(\epsilon l\tau y_- + \alpha\epsilon l^2 y y_-^2)\|\mathbf{x}_{k-1}-Y_\infty\overline{\mathbf{x}}_{k-1}\| + \alpha\epsilon l^2 y y_-\|\overline{\mathbf{x}}_{k-1}-\mathbf{z}^*\| \\
&+ (\sigma+\alpha\epsilon l y_-)\|\mathbf{w}_{k-1}-Y_\infty\mathbf{g}_{k-1}\| + (\alpha yl+2)\epsilon l y_-^2 T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|. \qquad (6.56)
\end{aligned}$$

Step 4: By combining Eqs. (6.43) in step 1, (6.49) in step 2, and (6.56) in step 3, we complete the proof.

6.6 Numerical Experiments

In this section, we compare the performances of algorithms solving the distributed consensus optimization problem over directed graphs, including ADD-OPT, DEXTRA [83], Gradient-Push [61], Directed-Distributed Subgradient Descent [79], and the Weighting Balancing-Distributed Gradient Descent [74]. Our numerical experiments are based on the distributed logistic regression problem over a directed graph:

$$\mathbf{z}^* = \underset{\mathbf{z}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{\beta}{2}\|\mathbf{z}\|^2 + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\ln\left[1+\exp\left(-\left(\mathbf{c}_{ij}^\top\mathbf{z}\right)b_{ij}\right)\right],$$

where for any agent i, it is accessible to m_i training examples, $(\mathbf{c}_{ij}, b_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ includes the p features of the jth training example of agent i, and b_{ij} is the corresponding label. This problem can be formulated in the form of P1 with the private objective function f_i being

$$f_i = \frac{\beta}{2n}\|\mathbf{z}\|^2 + \sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{c}_{ij}^\top \mathbf{z}\right) b_{ij}\right)\right].$$

In our setting, we have $n = 10$, $m_i = 10$, for all i, and $p = 3$. The network topology is described in Fig. 6.1.

In the implementation of algorithms, we apply to all algorithms the local degree weighting strategy, i.e., to assign each agent itself and its out-neighbors equal weights according to the agents's own out-degree.

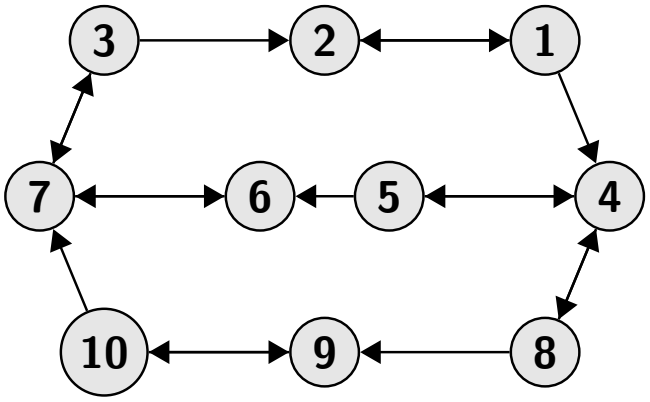


Figure 6.1: A strongly-connected directed network.

In our first experiment, we compare the convergence rates between ADD-OPT and other methods that designed for directed graphs. we apply the same

local degree weighting strategy to all methods. The step-size used in Gradient-Push, Directed-Distributed Subgradient Descent, and WeightingBalencing-Distributed Gradient Descent is $\alpha_k = 1/\sqrt{k}$. The constant step-size used in DEXTRA and ADD-OPT is $\alpha = 1$. It can be found that ADD-OPT and DEXTRA has a fast linear convergence rate, while other methods are sub-linear. The convergence rate performances between different algorithms are found in Fig. 6.2.

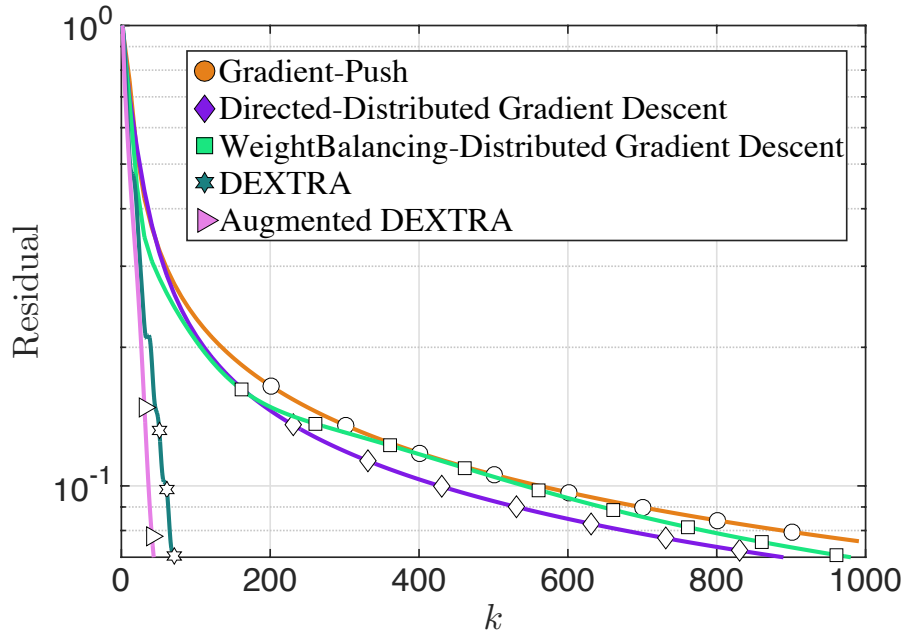


Figure 6.2: Convergence rates between optimization methods for directed networks.

The second experiment compares ADD-OPT and DEXTRA in terms of their step-size ranges. We stick to the same local degree weighting strategy

for both ADD-OPT and DEXTRA. It is shown in Fig. 6.3 that the greatest lower bound of DEXTRA is round $\underline{\alpha} = 0.2$. Since the value of $\underline{\alpha}$ requires the global knowledge, it is hard for agents to estimate this value in the distributed implementation. Once agents pick some value for $\alpha < 0.2$, DEXTRA diverges. In contrast, agents that implementing ADD-OPT can pick whatever small values to ensure the convergence.

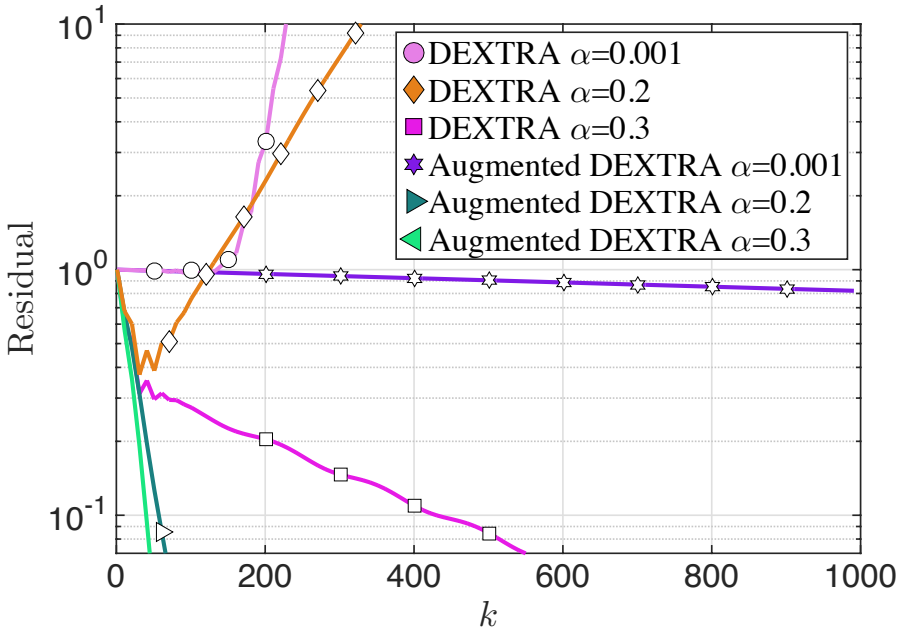


Figure 6.3: Comparison between ADD-OPT and DEXTRA in terms of step-size ranges.

According to the setting, we can calculate that $\tau = 1.25$, $\epsilon = 1.11$, $y = 1.96$, $y_{-} = 2.2$, and $l = 1$. It is also satisfied that $\sigma < 1$. Therefore, we can estimate $\overline{\alpha} < \frac{\sqrt{8.7}}{9.57} = 0.3$. It can be found in Fig. 6.4 that the practical upper bound of

step-size is much bigger, i.e., $\overline{\alpha} = 1.12$. In the implementation of ADD-OPT when we are trying to estimate $\overline{\alpha}$, we set $\overline{\alpha} \approx \frac{1}{10l}$ assuming that l is not a global knowledge. This is an experienced estimation. Finally, we also show the relation between convergence speed and step-size does not simply satisfy any linear function. This can be found in Fig. 6.4.

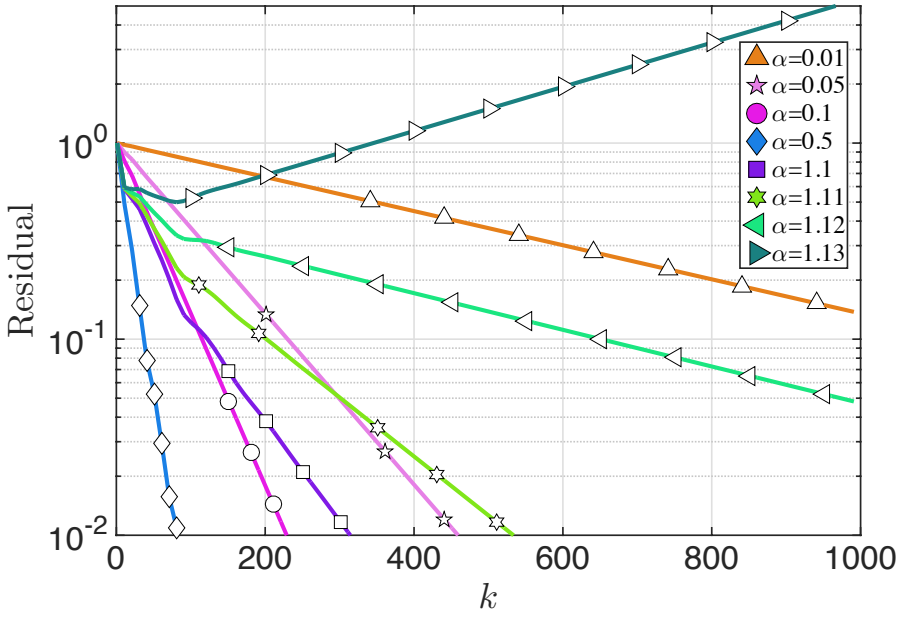


Figure 6.4: The range of ADD-OPT 's step-size.

6.7 Conclusions

In this chapter, we focus on solving the distributed consensus optimization problem over directed graphs. The proposed algorithm, termed ADD-OPT,

can be viewed as an improvement of our recent work, DEXTRA. ADD-OPT converges at a linear rate $O(\mu^k)$ for $0 < \mu < 1$ given the assumption that the objective functions are strongly-convex, where k is the number of iterations. Compared to DEXTRA, ADD-OPT owns a wider and more realistic range of step-size for the convergence. In particular, the greatest lower bound of DEXTRA's step-size is strictly greater than zero while that of ADD-OPT's equals exactly to zero. This guarantees the convergence of ADD-OPT in the distributed implementation as long as agents picking some small step-size. Therefore, ADD-OPT is more reliable in distributed setting. We provide a much more compact proof compared with DEXTRA to show the linear convergence rate of ADD-OPT. Simulation examples further illustrate the improvements.

Chapter 7

Epilogue

In this chapter, we conclude our contribution, and discuss some possible directions for future work.

In this thesis, we focus on solving optimization problems where information is distributed over multi-agents networks. Each agent in the network owns local information that is private. They cooperatively solve a global optimization problem through local computations and information exchange over the network. In particular, we consider the problem of minimizing a sum of objectives, $\sum_{i=1}^{n} f_i(\mathbf{x})$, where $f_i : \mathbb{R}^p \to \mathbb{R}$ is a private objective function at the ith agent of the network. Existing distributed methods mostly deal with this class of problem under the assumption that the multi-agents network is strongly-connected and undirected, i.e., if agent i can send information to agent j, then agent j can also send information to agent i. We relax the assumption

of network topologies to directed networks. The main contribution of this work lies in that we propose four algorithms, Directed-Distributed Subgradient Descent (D-DSD), Directed-Distributed Projection Subgradient (D-DPS), DEXTRA, and ADD-OPT to overcome the challenges. We now summarize each algorithm in the following.

- **Directed-Distributed Subgradient Descent (D-DSD):** D-DSD is a subgradient based method that combines surplus-consensus techniques and DGD [37] to minimize the sum of local objective functions when the network topology among agents is described by a directed graph. It converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions are convex, but not necessary to be differentiable. It can shown that D-DSD converges at a rate of $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

- **Directed-Distributed Projection Subgradient (D-DPS):** D-DPS solves the distributed optimization problem over directed networks with an additional constrained set. It can be viewed as a generalization of D-DSD when the constrained set changes from $\mathbb{R}^p$, meaning no constraint, to a convex constrained set, $\mathcal{X} \subseteq \mathbb{R}^p$. Same as D-DSD, D-DPS converges to the optimal solution in nonsmooth convex optimization, i.e., the local objective functions in the problem are convex, but not necessary to be

differentiable. The convergence rate is $O(\frac{\ln k}{\sqrt{k}})$, where k is the number of iterations.

- **DEXTRA:** DEXTRA harness the smoothness to obtain a fast convergence rate. In other words, DEXTRA converges to the optimal solution in smooth convex optimization, i.e., the local objective functions are convex and differentiable. We show that, with the appropriate step-size, DEXTRA converges at a linear rate $O(\tau^k)$ for $0 < \tau < 1$, given that the objective functions are restricted strongly-convex.

- **ADD-OPT:** ADD-OPT is the other algorithm that solves the distributed smooth optimization problem over directed networks. Same as EXTRA, it achieves the best known rate of convergence for this class of problems, $O(\mu^k)$ for $0 < \mu < 1$ given that the objective functions are strongly-convex, where k is the number of iterations. Compared with DEXTRA, ADD-OPT supports a wider and more realistic range of step-size. In particular, the greatest lower bound of DEXTRA's step-size is strictly greater than zero while that of ADD-OPT's equals exactly to zero.

In the analysis of D-DSD, we stick to the setting of static directed networks. Although we do not pursue this in this thesis, D-DSD can be generalized to work over time-varying directed graphs. Numerical experiments illustrate this

findings. Extending the analysis to the case of time-varying directed graphs would be important directions for future work.

Bibliography

[1] J. N. Tsitsiklis. *Problems in Decentralized Decision making and Computation.* Tech. rep. DTIC Document, 1984.

[2] D. P Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Vol. 23. Prentice hall Englewood Cliffs, NJ, 1989.

[3] J. Manyika et al. "Big data: The next frontier for innovation, competition, and productivity". In: (2011).

[4] D. P. Bertsekas, A. Nedi, and A. E. Ozdaglar. "Convex analysis and optimization". In: (2003).

[5] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] R. Bekkerman, M. Bilenko, and J. Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[7] S. Boyd et al. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundation and Trends in Maching Learning* 3.1 (Jan. 2011), pp. 1–122. ISSN: 1935-8237.

[8] V. Cevher, S. Becker, and M. Schmidt. "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics". In: *IEEE Signal Processing Magazine* 31.5 (2014), pp. 32–43.

[9] G. Mateos, J. A. Bazerque, and G. B. Giannakis. "Distributed sparse linear regression". In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5262–5276.

[10] J. B. Predd, S. R. Kulkarni, and H. V. Poor. "A collaborative training algorithm for distributed learning". In: *IEEE Transactions on Information Theory* 55.4 (2009), pp. 1856–1871.

[11] A. G. Dimakis et al. "Gossip algorithms for distributed signal processing". In: *Proceedings of the IEEE* 98.11 (2010), pp. 1847–1864.

[12] L. Xiao, S. Boyd, and S. Kim. "Distributed average consensus with least-mean-square deviation". In: *Journal of Parallel and Distributed Computing* 67.1 (2007), pp. 33–46.

[13] I. Necoara and J. A. K. Suykens. "Application of a Smoothing Technique to Decomposition in Convex Optimization". In: *IEEE Transactions on Automatic Control* 53.11 (2008), pp. 2674–2679.

[14] S. H. Low and D. E. Lapsley. "Optimization flow controlI: basic algorithm and convergence". In: *IEEE/ACM Transactions on Networking (TON)* 7.6 (1999), pp. 861–874.

[15] G. Mateos, J. A. Bazerque, and G. B. Giannakis. "Distributed Sparse Linear Regression". In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5262–5276.

[16] J. A. Bazerque and G. B. Giannakis. "Distributed Spectrum Sensing for Cognitive Radio Networks by Exploiting Sparsity". In: *IEEE Transactions on Signal Processing* 58.3 (2010), pp. 1847–1862. ISSN: 1053-587X.

[17] A. Ribeiro. "Ergodic stochastic optimization algorithms for wireless communication and networking". In: *IEEE Transactions on Signal Processing* 58.12 (2010), pp. 6369–6386.

[18] Y. Xu et al. "Distributed subgradient-based coordination of multiple renewable generators in a microgrid". In: *IEEE Transactions on Power Systems* 29.1 (2014), pp. 23–33.

[19] A. Jadbabaie, J. Lin, and A. S. Morse. "Coordination of groups of mobile autonomous agents using nearest neighbor rules". In: *IEEE Transactions on automatic control* 48.6 (2003), pp. 988–1001.

[20] S. Yang, S. Tan, and J. Xu. "Consensus based approach for economic dispatch problem in a smart grid". In: *IEEE Transactions on Power Systems* 28.4 (2013), pp. 4416–4426.

[21] U. A. Khan, S. Kar, and J. M. F. Moura. "DILAND: An Algorithm for Distributed Sensor Localization With Noisy Distance Measurements". In: *IEEE Transactions on Signal Processing* 58.3 (2010), pp. 1940–1947.

[22] M. Rabbat and R. Nowak. "Distributed optimization in sensor networks". In: *3rd International Symposium on Information Processing in Sensor Networks*. 2004, pp. 20–27.

[23] Q. Ling, Z. Wen, and W. Yin. "Decentralized jointly sparse optimization by reweighted minimization". In: *IEEE Transactions on Signal Processing* 61.5 (2013), pp. 1165–1170.

[24] K. Yuan et al. "A linearized Bregman algorithm for decentralized basis pursuit". In: *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE. 2013, pp. 1–5.

[25] Q. Ling et al. "Decentralized low-rank matrix completion". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2012, pp. 2925–2928.

[26] Y. Liao et al. "DMFSGD: A decentralized matrix factorization algorithm for network distance prediction". In: *IEEE/ACM Transactions on Networking* 21.5 (2013), pp. 1511–1524.

[27] C. L and L. Li. "A distributed multiple dimensional QoS constrained resource scheduling optimization policy in computational grid". In: *Journal of Computer and System Sciences* 72.4 (2006), pp. 706 –726.

[28] G. Neglia, G. Reina, and S. Alouf. "Distributed gradient optimization for epidemic routing: A preliminary evaluation". In: *2nd IFIP in IEEE Wireless Days*. 2009, pp. 1–6.

[29] C. Xi and U. A. Khan. "On the impact of low-rank interference on distributed multi-agent optimization". In: *48th Asilomar Conference on Signals, Systems and Computers*. 2014, pp. 1511–1514.

[30] R. Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013.

[31] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. "Distributed asynchronous deterministic and stochastic gradient optimization algorithms". In: *1984 American Control Conference*. 1984, pp. 484–489.

[32] D. P. Bertsekas. "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey". In: *Optimization for Machine Learning* 2010 (2011), pp. 1–38.

[33] A. Nedić and D. Bertsekas. "Convergence rate of incremental subgradient algorithms". In: *Stochastic optimization: algorithms and applications*. Springer, 2001, pp. 223–264.

[34] A. Nedic and D. P. Bertsekas. "Incremental subgradient methods for nondifferentiable optimization". In: *SIAM Journal on Optimization* 12.1 (2001), pp. 109–138.

[35] A. Nedic, D. P. Bertsekas, and V. S. Borkar. "Distributed asynchronous incremental subgradient methods". In: (2000).

[36] S. S. Ram, A. Nedic, and V. V. Veeravalli. "Incremental stochastic subgradient algorithms for convex optimization". In: *SIAM Journal on Optimization* 20.2 (2009), pp. 691–717.

[37] A. Nedic and A. Ozdaglar. "Distributed Subgradient Methods for Multi-Agent Optimization". In: *IEEE Transactions on Automatic Control* 54.1 (2009), pp. 48–61.

[38] A. Nedic, A. Ozdaglar, and P. A. Parrilo. "Constrained Consensus and Optimization in Multi-Agent Networks". In: *IEEE Transactions on Automatic Control* 55.4 (2010), pp. 922–938.

[39] I. Lobel, A. Ozdaglar, and D. Feijer. "Distributed multi-agent optimization with state-dependent communication". English. In: *Mathematical Programming* 129.2 (2011), pp. 255–284.

[40] I. Lobel and A. Ozdaglar. "Distributed Subgradient Methods for Convex Optimization Over Random Networks". In: *IEEE Transactions on Automatic Control* 56.6 (2011), pp. 1291–1306.

[41] S. S. Ram, A. Nedic, and V. V. Veeravalli. "Distributed subgradient projection algorithm for convex optimization". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 3653–3656.

[42] S. S. Ram, A. Nedic, and V. V. Veeravalli. "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization". English. In: *Journal of Optimization Theory and Applications* 147.3 (2010), pp. 516–545.

[43] S. Lee and A. Nedic. "Distributed Random Projection Algorithm for Convex Optimization". In: *IEEE Journal of Selected Topics in Signal Processing* 7.2 (2013), pp. 221–229.

[44] B. Johansson et al. "Subgradient methods and consensus algorithms for solving convex optimization problems". In: *47th IEEE Conference on Decision and Control*. 2008, pp. 4185–4190.

[45] J. C. Duchi, A. Agarwal, and M. J. Wainwright. "Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling". In: *IEEE Transactions on Automatic Control* 57.3 (2012), pp. 592–606.

[46] K. Yuan, Q. Ling, and W. Yin. "On the convergence of decentralized gradient descent". In: *arXiv preprint arXiv:1310.7063* (2013).

[47] D. Jakovetic, J. Xavier, and J. M. F. Moura. "Fast distributed gradient methods". In: *IEEE Transactions on Automatic Control* 59.5 (2014), pp. 1131–1146.

[48] W. Shi et al. "EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization". In: *SIAM Journal on Optimization* 25.2 (2015), pp. 944–966.

[49] J. F. C. Mota et al. "D-ADMM: A Communication-Efficient Distributed Algorithm for Separable Optimization". In: *IEEE Transactions on Signal Processing* 61.10 (2013), pp. 2718–2723.

[50] E. Wei and A. Ozdaglar. "Distributed Alternating Direction Method of Multipliers". In: *51st IEEE Annual Conference on Decision and Control*. 2012, pp. 5445–5450.

[51] W. Shi et al. "On the Linear Convergence of the ADMM in Decentralized Consensus Optimization". In: *IEEE Transactions on Signal Processing* 62.7 (2014), pp. 1750–1761. ISSN: 1053-587X.

[52] Q. Ling and A. Ribeiro. "Decentralized linearized alternating direction method of multipliers". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2014, pp. 5447–5451.

[53] M. Zargham et al. "Accelerated dual descent for network flow optimization". In: *IEEE Transactions on Automatic Control* 59.4 (2014), pp. 905–920.

[54] A. Jadbabaie, A. Ozdaglar, and M. Zargham. "A distributed newton method for network optimization". In: *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE. 2009, pp. 2736–2741.

[55] E. Wei, A. Ozdaglar, and A. Jadbabaie. "A distributed Newton method for network utility maximization–I: algorithm". In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2162–2175.

[56] E. Wei, A. Ozdaglar, and A. Jadbabaie. "A Distributed Newton Method for Network Utility MaximizationPart II: Convergence". In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2176–2188.

[57] A. Mokhtari et al. "Dqm: Decentralized quadratically approximated alternating direction method of multipliers". In: *arXiv preprint arXiv:1508.02073* (2015).

[58] A. Mokhtari, Q. Ling, and A. Ribeiro. "Network newton-part i: Algorithm and convergence". In: *arXiv preprint arXiv:1504.06017* (2015).

[59] A. Mokhtari, Q. Ling, and A. Ribeiro. "Network Newton-Part II: Convergence Rate and Implementation". In: *arXiv preprint arXiv:1504.06020* (2015).

[60] A. Mokhtari et al. "A decentralized second-order method with exact linear convergence rate for consensus optimization". In: *arXiv preprint arXiv:1602.00596* (2016).

[61] A. Nedic and A. Olshevsky. "Distributed optimization over time-varying directed graphs". In: *IEEE Transactions on Automatic Control* PP.99 (2014), pp. 1–1.

[62] A. Nedic and A. Olshevsky. "Distributed optimization over time-varying directed graphs". In: *52nd IEEE Annual Conference on Decision and Control*. 2013, pp. 6855–6860.

[63] K. I. Tsianos, S. Lawlor, and M. G. Rabbat. "Push-Sum Distributed Dual Averaging for convex optimization". In: *51st IEEE Annual Conference on Decision and Control*. 2012, pp. 5453–5458.

[64] K. I. Tsianos. "The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays". PhD thesis. Dept. Elect. Comp. Eng. McGill University, 2013.

[65] K. I. Tsianos, S. Lawlor, and M. G. Rabbat. "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning". In: *50th Annual Allerton Conference on Communication, Control, and Computing*. 2012, pp. 1543–1550.

[66] D. Kempe, A. Dobra, and J. Gehrke. "Gossip-based computation of aggregate information". In: *44th Annual IEEE Symposium on Foundations of Computer Science*. 2003, pp. 482–491.

[67] F. Benezit et al. "Weighted Gossip: Distributed Averaging using non-doubly stochastic matrices". In: *IEEE International Symposium on Information Theory*. 2010, pp. 1753–1757.

[68] A. Jadbabaie, J. Lim, and A. S. Morse. "Coordination of groups of mobile autonomous agents using nearest neighbor rules". In: *IEEE Transactions on Automatic Control* 48.6 (2003), pp. 988–1001.

[69] C. W. Reynolds. "Flocks, Herds and Schools: A Distributed Behavioral Model". In: *14th Annual Conference on Computer Graphics and Interactive Techniques*. New York, NY, USA: ACM, 1987, pp. 25–34. ISBN: 0-89791-227-6.

[70] R. Olfati-Saber and R. M. Murray. "Consensus problems in networks of agents with switching topology and time-delays". In: *IEEE Transactions on Automatic Control* 49.9 (2004), pp. 1520–1533.

[71] R. Olfati-Saber and R. M. Murray. "Consensus protocols for networks of dynamic agents". In: *IEEE American Control Conference*. Vol. 2. 2003, pp. 951–956.

[72] R. Olfati-Saber, J. A. Fax, and R. M. Murray. "Consensus and Cooperation in Networked Multi-Agent Systems". In: *Proceedings of the IEEE* 95.1 (2007), pp. 215–233.

[73] L. Xiao, S. Boyd, and S. J. Kim. "Distributed average consensus with least-mean-square deviation". In: *Journal of Parallel and Distributed Computing* 67.1 (2007), pp. 33 –46.

[74] A. Makhdoumi and A. Ozdaglar. "Graph Balancing for Distributed Subgradient Methods over Directed Graphs". In: *54th IEEE Annual Conference on Decision and Control* (2015).

[75] L. Hooi-Tong. "On a class of directed graphswith an application to traffic-flow problems". In: *Operations Research* 18.1 (1970), pp. 87–94.

[76] N. Derbinsky et al. "An improved three-weight message-passing algorithm". In: *arXiv preprint arXiv:1305.1961* (2013).

[77] Q. Ling et al. "Weighted ADMM for Fast Decentralized Network Optimization". In: *IEEE Transactions on Signal Processing* 64.22 (2016), pp. 5930–5942. ISSN: 1053-587X. DOI: 10.1109/TSP.2016.2602803.

[78] A. Ozdaglar. "Distributed Alternating Direction Method of Multipliers for Multi-agent Optimization". In: *Lund Workshop on Dynamics and Control in Networks* (Oct. 2014).

[79] C. Xi, Q. Wu, and U. A. Khan. "Distributed gradient descent over directed graphs". In: *arXiv preprint arXiv:1510.02146* (2015).

[80] C. Xi and U. A. Khan. "Directed-Distributed Gradient Descent". In: *53rd Annual Allerton Conference on Communication, Control, and Computing*. 2015, pp. 1022–1026.

[81] C. Xi and U. A. Khan. "Distributed Dynamic Optimization over Directed Graphs". In: *55th IEEE Conference on Decesion and Control*. 2016.

[82] C. Xi and U. A. Khan. "Distributed Subgradient Projection Algorithm over Directed Graphs". In: *arXiv preprint arXiv:1602.00653* (2016).

[83] C. Xi and U. A. Khan. "On the linear convergence of distributed optimization over directed graphs". In: *arXiv preprint arXiv:1510.02149* (2015).

[84] C. Xi, Q. Wu, and U. A. Khan. "Fast distributed optimization over directed graphs". In: *35th American Control Conference*. 2016, pp. 6507–6512.

[85] C. Xi and U. A. Khan. "ADD-OPT: Accelerated Distributed Directed Optimization". In: *arXiv preprint arXiv:1607.04757* (2016).

[86] C. Xi and U. A. Khan. "Augmented DEXTRA for Fast Distributed Optimization over Directed Graphs". In: *54th Annual Allerton Conference on Communication, Control, and Computing*. 2016.

[87] H. J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*. Vol. 35. Springer Science & Business Media, 2003.

[88] K. Cai and H. Ishii. "Average consensus on general strongly connected digraphs". In: *Automatica* 48.11 (2012), pp. 2750 –2761.

[89] G. W. Stewart. "Matrix perturbation theory". In: (1990).

[90] F. Chung. "Laplacians and the Cheeger inequality for directed graphs". In: *Annals of Combinatorics* 9.1 (2005), pp. 1–19.

[91] S. Bubeck. "Convex optimization: Algorithms and complexity". In: *arXiv preprint arXiv:1405.4980* (2014).