UNIVERSITY OF CALIFORNIA
RIVERSIDE

Distributed Algorithms to Convex Optimization Problems

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Peng Wang

December 2017

Dissertation Committee:

Dr. Wei Ren, Chairperson
Dr. Fabio Pasqualetti
Dr. Nanpeng Yu

The Dissertation of Peng Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Dr. Wei Ren, for the great opportunities to work with him. His insightful guidance benefits me in not only academic but also everyday life. His rigourous academic attitude and painstaking attention to our work are always affecting me during the past few years. I would like to thank Dr. Nanpeng Yu for his brainy advice on topics related to power systems. I would also like to thank all my lab colleagues and friends in UCR, who make my life easier. Among them, I want to express my special thanks to Shouxu Zhang, who helps me overcome financial hardships and Xuewei Qi, who helps me to broaden my academic view. I would like to thank my parents who always support and encourage me. Finally, I would like to acknowledge IEEE Transactions on Automatic Control, 2017 American Control Conference, and IEEE 55th Conference on Decision and Control, where I published my work used in this dissertation.

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Distributed Algorithms to Convex Optimization Problems

by

Peng Wang

Doctor of Philosophy, Graduate Program in Electrical and Computer Engineering
University of California, Riverside, December 2017
Dr. Wei Ren, Chairperson

This dissertation studies first a distributed algorithm to solve general convex optimization problems and then designs distributed algorithms to solve special optimization problems related to a system of linear equations.

First, a wider selection of step sizes is explored for the distributed subgradient algorithm for multi-agent optimization problems with time-varying and balanced communication topologies. The square summable requirement of the step sizes commonly adopted in the literature is removed. The step sizes are only required to be positive, vanishing and non-summable, which provides the possibility for better convergence rates. Both unconstrained and constrained optimization problems are considered. It is proved that the agents' estimates reach a consensus and converge to the minimizer of the global objective function with the more general choice of step sizes. The best convergence rate is shown to be the reciprocal of the square root of iterations for the best record of the function value at the average of the agents' estimates for the unconstrained case with the wider selection of step sizes.

Then we design a distributed algorithm for a special optimization problem to find the solution of the linear equations $Ax = b$ with minimum energy, i.e. the minimum weighted norm associated with the weighted inner product. We first prove that for a special case when the norm is two-norm, the algorithm can make multiple agents reach the minimum two-norm solution of the global linear equations $Ax = b$ if the agents are initialized at the minimum two-norm solutions of their local equations. We then prove that if there are bounded initialization errors, the final convergence of the algorithm is also bounded away from the minimum two-norm solution of the global linear equations. Next, we prove the case with the two-norm replaced with a weighted norm associated with the weighted inner product.

Next, we solve a system of linear equations $Ax = b$ in a distributed way motivated by a special subgradient algorithm. A discrete-time distributed algorithm to solve a system of linear equations $Ax = b$ is proposed. The algorithm can find a solution of $Ax = b$ from arbitrary initializations at a geometric rate when $Ax = b$ has either unique or multiple solutions. When $Ax = b$ has a unique solution, the geometric convergence rate of the algorithm is proved by analyzing the mixed norm of homogeneous M-Fejer type mappings from the subgradient update. Then when $Ax = b$ has multiple solutions, the geometric convergence rate is proved through orthogonal decompositions of the agents' estimates onto the row space and null space of A, and the relationship between the initializations and the final convergence point is also specified. Quantitative upper bounds of the convergence rate for two special cases are given.

Finally, we propose a communication-efficient distributed algorithm to solve a sys-

tem of linear equations $Ax = b$. We suppose that the matrix A has a similar sparsity structure to the Laplacian matrix of the communication topology. Every agent only transmits through a communication link its own state and that of the neighbor connected by the communication link, instead of the states of all agents. The distributed algorithm is designed based on gradient descent method with constant step size and is proved to converge at a linear rate. We also provide a way to select the step sizes in a distributed way.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview and Related Works

1.1.1 Distributed Subgradient Algorithm for General Convex Optimization Problems

With the emergence of large-scale networks and complex large systems, distributed optimization arises in many areas such as distributed model predictive control [27], distributed signal processing [6], optimal network flow [55] and network utility maximization [51] and has attracted significant attention. The distributed optimization problems can be roughly classified into two categories. In the first category, each agent has a local objective function and sometimes a local constraint, both unknown to others, but different agents share the same optimization variable. This means that different agents' estimates of the optimizer should be the same at last [12, 18, 22, 33, 34, 44, 56, 57]. The problems in this category can be regarded as distributed potential problems. In the second category, every

agent has a local objective function unknown to others, the constraints of the agents are coupled, and every agent knows only a part of the coupled constraints [5, 27, 51, 55]. The problems in this category can be regarded as distributed network flow problems. In this dissertation, we will focus on the problems in the first category.

Various algorithms have been developed to solve the problems in the first category. In [33], a distributed subgradient algorithm is designed for an unconstrained distributed optimization problem, with the assumption of uniformly bounded subgradients, and non-degenerate, time-varying and balanced communication topologies. In [34], a distributed optimization problem with identical local constraints or non-identical local constraints in the context of a complete graph is considered through a projected distributed subgradient algorithm. Ref. [22] considers non-identical local constraints for balanced and state-dependent switching graphs. Then [18] proves the convergence of the distributed subgradient algorithm with non-identical local constraints and time delays under time-varying balanced and fixed unbalanced graphs. Some accelerated algorithms are proposed in [12], in which two distributed Nesterov gradient methods are designed and these algorithms are shown to converge faster than the distributed subgradient algorithm in [33]. Ref. [44] develops a distributed algorithm with a constant step size using the gradients of last two iterations and shows that the algorithm can guarantee a faster convergence rate. A zero-gradient-sum algorithm is developed in [25], in which each agent starts from its local minimizer and the sum of the gradients is kept at zero. On the other hand, some dual or primal-dual subgradient algorithms are developed for distributed optimization problems with equality and inequality constraints. Ref. [56] proposes a distributed primal-dual subgradient algorithm

to deal with identical affine equality and convex inequality constraints. A projected subgradient method is designed to find the saddle point of the Lagrangian of the primal problem. Then in [57], a similar idea is adopted to develop a distributed dual subgradient algorithm to solve a non-convex problem approximately, with the consensus requirement relaxed.

In the above papers on subgradient-related distributed solutions to the optimization problem [18,33,34,56,57], the step sizes for the subgradient should be positive, vanishing, non-summable but square summable. However, such selection of step sizes excludes an important class of step sizes that is not square summable in the form of $\frac{D}{\sqrt{k+1}}$, where D is a positive constant and k denotes the kth iteration. Actually, in the centralized subgradient algorithm [36], such selection guarantees convergence and provides the fastest convergence rate for the best record of the objective function values. It would be interesting to explore whether a similar result holds in the distributed context. That will be what to be done in Chapter 3.

Next, we turn to designing distributed algorithm to some special optimization problems that are related to linear equations because solving a group of linear equations $Ax = b$ is probably among the most important problems in numerical computations of real numbers.

1.1.2 Distributed Algorithms to Find Minimum Norm Solution of $Ax = b$

In Chapter 4 of this dissertation, we will focus our attention on finding the solution of a system of linear equations $Ax = b$ with minimum norm in the distributed way by assuming that each agent knows one or several rows of the augmented matrix $[A\ b]$, as in [1, 29, 43]. [29] proposes a distributed algorithm to solve $Ax = b$ from locally feasible

initializations with synchronous and asynchronous updates and also considers time-varying linear equations and a least square solution to unsolvable equations. Ref. [1] provides a continuous-time distributed algorithm by using a result from differential geometry and removes the requirement of feasible initializations. Then two continuous-time algorithms are proposed in [43], with convergence guarantees, explicit formulations of relationships between the initial values and final convergence value, and considerations of least square problems with additional requirements.

When the equations $Ax = b$ have multiple solutions, one might be more concerned about obtaining a specific solution with some special properties, e.g., minimum norm, rather than merely obtaining a solution. Although the algorithms in [1, 20, 29, 43] are proved to arrive at some solution for solvable linear equations with multiple solutions, but it is not clear how to obtain a solution of $Ax = b$ with special properties from these results. In Chapter 4, we will design a distributed algorithm to find the special solution with minimum norms associated with inner product of a system of linear equations $Ax = b$.

If we look further into the discrete-time distributed algorithms to solve a system of linear equations in [30, 49, 53], we will see that they require some unnecessary conditions. Next, we will remove those conditions motivated by a special subgradient algorithm in [46].

1.1.3 Distributed Algorithm to Solve $Ax = b$

Many distributed algorithms are proposed in the literature (see e.g., [1, 19, 30, 38–40, 43, 49, 52, 53]), assuming that each agent knows one or several rows of the augmented matrix $\begin{bmatrix} A & b \end{bmatrix}$. The algorithms in [1, 19, 43, 52] are continuous-time ones, which require communication at every time. In contrast, in this part, we focus on discrete-time distributed

algorithms to solve a system of linear equations $Ax = b$. In [30], a distributed algorithm with locally feasible initializations is proposed for both synchronous and asynchronous cases and is proved to converge at a geometric rate under repeatedly jointly strongly connected graphs. Then in [49], when $Ax = b$ has a unique solution, a distributed algorithm is designed to allow arbitrary initializations with feedback of the deviation from local systems of linear equations and the geometric convergence rate is proved. Also, when $Ax = b$ has a unique solution, [53] proposes an algorithm with an adaptation of the subgradient algorithm and proves the linear convergence rate. In [38–40], a distributed algorithm is also proposed to solve a system of linear equations. The algorithm requires agents to share the kernel of local equations with their neighbors, which leads to heavy communication burden because the kernels of local equations are usually of large volumes of data.

The algorithms in [30, 49, 53] are shown to be effective, but they require either special initializations in [30] or the unique solution of $Ax = b$ in [49, 53]. In Chapter 5, we will design a distributed algorithm to solve a system of linear equations $Ax = b$, which allows arbitrary initialization and can handle the cases when $Ax = b$ has multiple solutions.

1.1.4 Distributed Algorithm to Solve $Ax = b$ for sparse A

The distributed algorithms in [30, 49, 53] requires that every agent store and estimate the whole x vector, which leads to a lot of overhead when the matrix A is sparse. Large-scale systems of linear equations with sparse matrices widely exists in many fields such as power system, e.g., the Ybus, and finite element method to partial differential equations. In these problems, there are more numerous nodes, e.g., buses in power systems, in the system, but each node is connected to only a few other buses, which makes the

matrix sparse. It is neither practical to make every node know the whole vector due to the system scale nor necessary due to the sparsity of the connections. So it is important to investigate communication-efficient distributed algorithms to solve $Ax = b$ with sparse A. Such an algorithm is proposed in [28], but it requires that nodes share the information of their common neighbors, which may not be available to the nodes and are not used by the nodes. In Chapter 6, we will design a communication-efficient distributed algorithm to solve $Ax = b$ employing the sparsity structure of matrix A. The connected nodes share only their estimations of their own states and those of one of their neighbors', which requires less communication than existing results.

1.2 Organization

In Chapter 3, we will show that the square summability of the step sizes is not necessary for the distributed subgradient method, which provides the possibility for better convergence rates. We will prove that in both unconstrained and constrained distributed optimization problems, the positive, vanishing and non-summable step sizes can make the agents' estimates converge to the minimizer of the global objective function under time-varying balanced graphs. This step size selection is actually the same as that required by the centralized subgradient method [46] for the unconstrained optimization problems, including $\frac{D}{\sqrt{k+1}}$ as a special case. We also show that when the step sizes are chosen as $\frac{D}{\sqrt{k+1}}$, the best record of the function value at the average of the agents' estimates converges at the rate of $O(\frac{1}{\sqrt{k}})$ for the unconstrained case. The convergence rate $O(\frac{1}{\sqrt{k}})$ is the same as the optimal one in the centralized subgradient algorithm in [36]. For the unconstrained

optimization problem, we first show the convergence to the minimizer of a subsequence of the average of the agents' estimates by investigating the distance change from the average to the optimal set. Then we show that as time goes by, the average stays in a neighborhood, vanishing with the step sizes, of arbitrary level sets of the global objective function. Next, with consensus, we prove that the estimates of all agents approach the same minimizer. For the constrained optimization problem under the assumption of bounded constraint sets, we perform a similar analysis on the summation of distances of the agents' estimates to a minimizer. As the unconstrained case cannot be treated as a special case of the constrained case under the assumption of bounded constraint sets and vice versa, we deal with them separately. The above results hold for time-varying balanced graphs that are jointly strongly connected. We then show the convergence rate $O(\frac{1}{\sqrt{k}})$ from the distance change mentioned above when the step sizes are selected as $\frac{D}{\sqrt{k+1}}$ for the unconstrained case.

In Chapter 4, we will focus on finding the solution of $Ax = b$ with the minimum weighted norm associated with the weighted inner product in a distributed way as this kind of solution represents the solution with minimum energy. We propose a discrete-time distributed algorithm to find such a solution of $Ax = b$, including the algorithm in [20,29,31] as a special case. We first prove that when the norm is the two-norm, if all agents start from the minimum two-norm solution of their local equations and travel along the sequence generated by the algorithm, they will finally converge to the minimum two-norm solution of the global linear equations $Ax = b$. We also prove that if there are bounded initialization errors, the final solution of the algorithm is also bounded away from the global minimum two-norm solution. Then we extend the results to weighted norms associated with the

weighted inner products.

In Chapter 5, we propose a discrete-time distributed algorithm motivated by a special subgradient algorithm to solve $Ax = b$ with either unique or multiple solutions from arbitrary initializations. The algorithm can guarantee convergence to a solution of $Ax = b$ at a geometric rate. When the system of linear equations $Ax = b$ has a unique solution, we show the geometric convergence rate by analyzing the mixed norms of homogeneous M-Fejer type mappings related to the dynamics of the algorithm. When the system of linear equations $Ax = b$ has multiple solutions, we first perform an orthogonal decomposition of the agents' estimates onto the row space and null space of matrix A, and then show that the part in the row space admits a unique solution and that in the null space is a consensus algorithm. Combining the two parts, we can obtain the geometric convergence rate for a system of linear equations $Ax = b$ with more than one solution. When $Ax = b$ has multiple solutions, the limit point of the agents' estimates is proved to be determined by the agents' initializations, communication topologies, and the minimum 2-norm solution of the linear equations. We also obtain quantitative upper bounds related to the condition number of A and a parameter in the algorithm for the convergence rate for two special cases. To the best of our knowledge, this is the first quantitative result on the convergence rate of such an algorithm.

In Chapter 6, we propose a communication-efficient distributed algorithm to solve $Ax = b$ when the matrix A has a sparsity structure to the Laplacian matrix. The proposed algorithm is based on gradient descent method with constant step size. Agents connected by an edge only share their estimates of the states of the two agents connected by the same

edge. The convergence to a solution of $Ax = b$ is proved to be either in finite time or at a linear rate.

Notations We use $\mathbb{R}$ for the set of real numbers, $\mathbb{R}^n$ for the set of $n \times 1$ real vectors, $\mathbb{R}^{m\times n}$ for the set of $m \times n$ real matrices, and I for the identity matrix, whose dimension is determined by contexts. Let 0 represent zero vectors, matrices or mappings, of which the dimension is determined by contexts. The symbol $\mathbb{N}^+$ represents the set of positive integers, i.e., $\mathbb{N}^+ = \{1, 2, 3, \cdots\}$, and the symbol $\mathbb{N}$ represents the set of natural numbers, i.e. $\mathbb{N} = \{0\} \bigcup \mathbb{N}^+$. A sequence of real numbers or vectors $x(k),\ k = 1, 2, \cdots$, is represented by $\{x(k)\}$. We say a sequence $\{x(k)\}$ converges to $x^\star$ at a linear rate or at a geometric rate if there exists $N \in \mathbb{N}$, such that for all $k > N$, $\|x(k+1)-x^\star\| \leq c\|x(k)-x^\star\|$, where $0 < c < 1$ is a constant. $\|\cdot\|$ denotes the general norm, $\|\cdot\|_2$ means the 2-norm, and $\|\cdot\|_\infty$ denotes the ∞-norm. The distance between a point x and some set X is $d(x, X) = \inf_{p\in X} \|x - p\|_2$, and the distance between two sets X and Y is defined as $d(X, Y) = \inf_{x\in X,\, y\in Y} \|x - y\|_2$. The transpose of a vector x is represented by x^T. We let $\mathbf{1}_n$ be the $n \times 1$ vector of all ones. We use $\left(x_j\right)_n$ to denote a vector whose jth entry is x_j. We use $P_X(x)$ to denote the projection of a point x onto a closed convex set X: $P_X(x) = \arg\min_{p\in X} \|x - p\|_2$. $\lfloor x \rfloor$ represents the largest integer that is less than or equal to x. $O(\cdot)$ is used for infinitesimals of the same order, i.e., y is $O(x)$ if there exists a constant C such that $\|y\| \leq C\|x\|$ as $x \to 0$. If $M \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, we use $\|x\|_M$ to represent the weighted norm associated with the weighted inner product defined by M, i.e. $\|x\|_M = \sqrt{x^T M x}$. We use rank(A) to represent the rank of A, span(A) to represent the column space of A, and ker(A) to denote the kernel of A, which is the vector space $\{x|Ax = 0\}$. We use A^T for

the transpose of a matrix A. The dimension of a space E is represented by $\dim(E)$. The orthogonal complement of a subspace E is denoted as $E^{\perp}$. The singular value of a matrix A is represented by $\sigma(A)$, and the maximal one is denoted as $\sigma_{\max}(A)$ while the smallest one $\sigma_{\min}(A)$. We use a_{ij} to represent the ijth entry of matrix A and A_{ij} for the ijth block of A. We also use $\left(A_{ij}\right)_{m\times n}$ ($\left(a_{ij}\right)_{m\times n}$) to denote an matrix whose ijth block (entry) is A_{ij}, which is composed of $m \times n$ such blocks (entries). We also use $\otimes$ for the Kronecker product. A matrix is positive if all its entries are positive.

Chapter 2

Preliminaries

In this chapter, we present some preliminary knowledge that will pave the way to the main results in the following chapters.

2.1 Graph Theory

An $\bar{m}$th order directed graph, denoted by $\mathcal{G}(V, E, A)$, is composed of a vertex set $V = \{1, \cdots, \bar{m}\}$, an edge set $E \subseteq V \times V$ and a weight matrix A. We use the pair (j, i) to denote the edge from vertex j to vertex i. We suppose that $(i, i) \in E,\ \forall i \in V$. The weight matrix $W = [w_{ij}]_{\bar{m}\times\bar{m}} \in \mathbb{R}^{\bar{m}\times\bar{m}}$ associated with the graph $\mathcal{G}$ is defined such that w_{ij} is positive if $(j, i) \in E$, and $w_{ij} = 0$ otherwise. We assume that A is row stochastic, i.e., $\sum_{j=1}^{\bar{m}} w_{ij} = 1,\ \forall i \in V$. The graph $\mathcal{G}$ is balanced if $\sum_{j=1}^{\bar{m}} w_{ij} = \sum_{j=1}^{\bar{m}} w_{ji},\ \forall i \in V$. The neighbor set of vertex i is defined as $N_i = \{j : (j, i) \in E\}$. A directed path from i to j is a sequence of edges $(i, i_1), (i_1, i_2), \cdots, (i_p, j)$, starting from vertex i and sinking at vertex j. The directed graph $\mathcal{G}$ is strongly connected if, for any pair of vertices i and j, there is a directed path

from i to j. The union of a collection of graphs is a graph with the vertex and edge sets being the unions of the vertex and edge sets of the graphs in the collection.

Let $\mathcal{G}_k$ denote the graph at discrete-time index k. In the rest of the dissertation, we suppose the communication topologies of the agent network are jointly strongly connected as in the following assumption:

Assumption 1 *There exists a constant integer B and an infinite sequence $k_0, k_1, \cdots, k_t, \cdots$ where $0 < k_{j+1} - k_j \leq B, j = 0, 1, 2, \cdots$ such that the union of $\mathcal{G}_{k_j}, \mathcal{G}_{k_j+1}, \cdots, \mathcal{G}_{k_{j+1}-1}$ is strongly connected.*

This assumption ensures that agents can influence each other over a sufficient long time.

We also suppose that the weight matrix associated with $\mathcal{G}_k$ are non-degenerate as in the following assumption:

Assumption 2 *There exists a positive constant $\eta > 0$ such that $w_{ij}(k) > \eta$ if $w_{ij}(k) > 0$, $k = 0, 1, 2, \cdots$.*

With Assumption 2, the edge weights in the weight matrix is either zero or greater than a positive constant η. This assumption shows that if agent i receives information from agent j, then the edge weight w_{ij} is uniformly bounded away from zero. This assumption guarantees that the influence of agents on others, if any, lasts and does not vanish over time.

With the above assumptions, we have the following lemma on the product of the weight matrices:

Lemma 1 *[18] Let*

$$\Phi(t : s) = W(t)W(t-1)\cdots W(s). \tag{2.1}$$

Then under Assumptions 1 and 2, there exists a constant $\mu \in (0, \frac{1}{n})$ *such that for all* i, j *and all* q, p *that* $q - p \geq (\bar{m} + 2)B$, $\Phi_{ij}(q : p) > \mu$.

2.2 Convex Optimization

A set C is convex if for all $x,\ y \in C$, $\alpha x + (1 - \alpha)y \in C$, for all $\alpha \in [0, 1]$. That is, the line segment is in the set C if the two endpoints are. A function f is convex if its domain is convex and for all x and y in its domain and for all $\alpha \in [0, 1]$, $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$. A convex function is proper if $f(x) < +\infty$ for at least one x in its domain and $f(x) > -\infty$ for every x in its domain. In this dissertation, we only consider proper convex functions. For a proper convex function, it is closed if it is lower semi-continuous [42]. For a closed proper convex function, we have the following lemma:

Lemma 2 *[42] Let* f *be a closed proper convex function. If the level set* $\{x : f(x) \leq \alpha\}$ *is non-empty and bounded for one* α*, it is bounded for every* α.

An optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in X \end{aligned}$$

is a convex optimization problem if the objective function $f(x)$ is a convex function and the constraint set X is a convex set.

A vector g is a subgradient of a convex function f at the point x_0 if for all x in the domain of f,

$$f(x) - f(x_0) \geq g^T(x - x_0). \tag{2.2}$$

We denote the set of subgradients of f at x_0 as $\partial f(x_0)$.

For a projection operator onto a closed convex set, we have the following non-expansiveness property.

Lemma 3 *[34] Let $X \subset \mathbb{R}^m$ be a closed convex set. For any pair of points x and y in $\mathbb{R}^m$, we have $\|P_X(x) - P_X(y)\|_2 \leq \|x - y\|_2$.*

2.3 Affine space

We have the following definitions and lemmas on affine spaces.

Definition 1 *[47] Let E be a vector space of dimension n over a field $\mathbb{F}$. An affine space over E is a set $\mathbb{A}$ together with the map*

$$\mathbb{A} \times E \to \mathbb{A}$$

$$P, v \to P + v$$

such that:

1. *$P + \vec{0} = P$ for all $P \in \mathbb{A}$, where $\vec{0}$ is the identity element of E;*
2. *$P + (v + w) = (P + v) + w$ for all $P \in \mathbb{A}$ and $v, w \in E$;*
3. *given $P, Q \in \mathbb{A}$, there exists a unique $v \in E$ such that $P + v = Q$. Denote v as $\vec{PQ}$, or $Q - P$.*

Here E is called the associated vector space of $\mathbb{A}$. We define $\dim(\mathbb{A})$ as $\dim(E)$. For each point $P \in \mathbb{A}$ and for each vector subspace F of E, define

$$P + [F] = \{Q \in \mathbb{A} : Q = P + v, v \in F\}.$$

Definition 2 *[47] A subset $\mathbb{B}$ of an affine space $\mathbb{A}$ is an affine subspace of $\mathbb{A}$, with associated vector space being a vector subspace F of E, if*

1. *for all $P \in \mathbb{B}$ and for all $v \in F$, one has $P + v \in \mathbb{B}$. Moreover, the map*

$$\mathbb{B} \times F \to \mathbb{B}$$

$$P, v \to P + v$$

satisfies

2. *$P + \vec{0} = P$ for all $P \in \mathbb{B}$, $\vec{0} \in F$;*

3. *$P + (v + w) = (P + v) + w$ for all $P \in \mathbb{B}$ and $v, w \in F$;*

4. *for each pair of points $P, Q \in \mathbb{B}$, $\vec{PQ} \in F$.*

If $\mathbb{B}$ is an affine subspace of $\mathbb{A}$, with associated vector space F, and $P \in \mathbb{B}$, then $\mathbb{B} = P + [F]$ [47]. If $\mathbb{B} = P + [F]$, we say that $\mathbb{B}$ is a plane through P directed by F, and F is the direction of $\mathbb{B}$. If F is 1-dimensional, we say that $\mathbb{B}$ is a straight line, or simply a line.

Definition 3 *[47] Two planes $L_1 = P_1 + [F_1]$ and $L_2 = P_2 + [F_2]$ are parallel if $F_1 \subset F_2$ or $F_2 \subset F_1$.*

Definition 4 *[47] A Euclidean affine space is an affine space $\mathbb{A}$ such that the associated vector space E is a Euclidean vector space.*

Definition 5 *[47] An affine frame in an affine space $\mathbb{A}$ is a set $R = \{P; (e_1, \cdots, e_n)\}$ formed by a point $P \in \mathbb{A}$ and a basis $(e_1, \cdots, e_n)$ of the associated vector space E. The point P is called the origin of the affine frame.*

For any point $Q \in \mathbb{A}$, if $\vec{PQ} = \sum_{i=1}^{n} q_i e_i, q_i \in \mathbb{F}$, we say that Q has affine coordinates, or simply coordinates $(q_1, \cdots, q_n)$. In this dissertation, we suppose that the affine frame is always fixed with its origin being 0 and do not distinguish the points in $\mathbb{A}$ from their coordinates.

Lemma 4 *[15,47] Let $L = Q + [F]$ be a plane of dimension r in an affine space $\mathbb{A}$ and let $R = \{P; (e_1, \cdots, e_n)\}$ be an affine frame. Then the coordinates of the points of L in R are a solution of a linear system $Ax = b$, of n unknowns and rank $n - r$. Conversely, given a linear system $Ax = b$ and an affine frame R, we can interpret the solutions of this system as a plane L.*

We can see from this lemma that when the affine frame is fixed, the linear equations and planes in an affine space can be regarded as the same.

Lemma 5 *[15] A subset $\mathbb{B} \subset \mathbb{A}$ is an affine subspace, or a plane, if and only if it contains the straight lines that pass two different points of $\mathbb{B}$.*

Definition 6 *[47] Let $\mathbb{A}$ be a Euclidean affine space. The distance $d(P, Q)$ between the points $P, Q \in \mathbb{A}$ is given by*

$$d(P, Q) = \|\vec{PQ}\|_2.$$

Definition 7 *[47] Two planes $L_1 = P_1 + [F_1]$ and $L_2 = P_2 + [F_2]$ of a Euclidean affine space $\mathbb{A}$ is said to be orthogonal if the vector spaces F_1 and F_2 are orthogonal.*

Lemma 6 *[47] [Pythagoras' Thoerem] If the triangle $\triangle PQR$ is right angled at P, then*

$$d(Q, R)^2 = d(P, Q)^2 + d(P, R)^2.$$

Remark 1 *From Lemma 6, we know that the minimum distance between any point Q and a plane $\mathbb{B}$ in the affine space $\mathbb{A}$ is the distance between Q and its pedal on $\mathbb{B}$.*

2.4 M-Fejer Mapping

In this part, we will introduce some knowledge on M-Fejer mappings.

Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping, and define the set of fixed points of T as $\text{Fix}(T) = \{x \in \mathbb{R}^n : T(x) = x\}$. Then we have a special class of mappings with a nonempty fixed point set defined as follows:

Definition 8 *[48] A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is called M-Fejer, if $M = Fix(T) \neq \varnothing$ and*

$$\|T(x) - z\| < \|x - z\|, \forall x \in \mathbb{R}^n, x \notin M, \forall z \in M.$$

In the rest of the dissertation, we use Tx for $T(x)$ for short.

If M in the above definition contains 0, we have the following lemma:

Lemma 7 *If T is an M-Fejer mapping and $0 \in M$, then $\|Tx\| = \|x\|$ if and only if $Tx = x$.*

Proof. The sufficiency part is obvious.

We now prove the necessity part. For all $x \notin M$, i.e. $Tx \neq x$, we have that $\|Tx\| = \|Tx - 0\| < \|x - 0\| = \|x\|$. Thus, $\|Tx\| = \|x\|$ implies $x \in M$. ■

Next, we introduce the mixed vector norms that are used in the distributed algorithm to find a common fixed point of a family of M-Fejer mappings.

Definition 9 *[9] Let $x_i \in \mathbb{R}^n$, $i = 1, 2, \cdots, \bar{m}$, and*

$$\boldsymbol{x} = \begin{pmatrix} x_1^T & \cdots & x_{\bar{m}}^T \end{pmatrix}^T. \tag{2.3}$$

Then the mixed norm $\|\cdot\|_{p,\infty}$ *is defined as* $\|\boldsymbol{x}\|_{p,\infty} = \max_i \|x_i\|_p$ *for* $1 < p < \infty$.

With the mixed vector norm, we have that

Lemma 8 *[9] Let* S *be a stochastic matrix. Then* $\|(S\otimes I)x - \mathbf{1}_{\bar{m}}\otimes y\|_{p,\infty} \leq \|x - \mathbf{1}_{\bar{m}}\otimes y\|_{p,\infty}$ *for any* $x \in \mathbb{R}^{\bar{m}n}$ *and* $y \in \mathbb{R}^n$.

With the above definition of mixed norms, we introduce the following result on a stack of M-Fejer mappings.

Lemma 9 *[9] Let* $T_i,\ i = 1, \cdots, \bar{m}$ *be continuous* M_i*-Fejer mappings with respect to p-norm in* $\mathbb{R}^n$, $1 < p < \infty$, *and* $\bigcap_{i=1}^{\bar{m}} M_i \neq \varnothing$. *Also define* $\boldsymbol{T} : \mathbb{R}^{\bar{m}n} \to \mathbb{R}^{\bar{m}n}$ *as a stack of all* $T_i,\ i = 1, \cdots, \bar{m}$, *i.e.,*

$$\boldsymbol{T}\boldsymbol{x} = \begin{pmatrix} (T_1 x_1)^T & \cdots & (T_{\bar{m}} x_{\bar{m}})^T \end{pmatrix}^T, \tag{2.4}$$

where $\boldsymbol{x}$ *is defined as in (2.3). Let* $W(1), W(2), \cdots, W(q)$ *be a set of* $\bar{m} \times \bar{m}$ *stochastic matrices. If every entry of the matrix product* $W(q)W(q-1)\cdots W(1)$ *is positive, then the composed map* $\boldsymbol{x} \longrightarrow (W(q)\otimes I)\boldsymbol{T}\cdots(W(1)\otimes I)\boldsymbol{T}\boldsymbol{x}$ *is a continuous* M*-Fejer mapping with respect to the mixed vector norm* $\|\cdot\|_{p,\infty}$, *and its fixed point set is* $\{\mathbf{1}_{\bar{m}} \otimes y : y \in \bigcap_{i=1}^{n} M_i\}$.

Chapter 3

Distributed Subgradient-based Multi-agent Optimization with More General Step Sizes

In this chapter, we will provide a wider choice of step sizes for the distributed subgradient algorithm to solve multi-agent optimization problems. We will first prove the convergence of the algorithm. And then we will show that the distributed subgradient algorithm achieves the fastest convergence rate with the wider selection of step sizes.

3.1 Problem Statement

For a multi-agent system with n agents, we regard each agent as a vertex. There is an edge (j, i) if agent i receives information from agent j. The corresponding entry w_{ij} in the weight matrix A denotes the weight assigned by agent i to the received information

from agent j.

For the distributed optimization problems, we will focus on the first category described in Chapter 1. In this category, each agent has a local objective function and sometimes a local constraint, both unknown to others, but different agents share the same optimization variable. Each agent has a private local objective function unknown to the other agents but shares the same optimization variable. Also it may have its private local constraint. The goal of the multi-agent system is to cooperatively figure out a minimizer of the average of all local objective functions in the common part of all local constraints:

$$\begin{aligned} \text{minimize} \quad & f(x) = \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} f_i(x) \\ \text{subject to} \quad & x \in X = \bigcap_{i=1}^{\bar{m}} X_i, \end{aligned} \tag{3.1}$$

where $x \in \mathbb{R}^n$ is the variable of the multi-agent system, $f_i,\ i \in V$, are the local objective functions and $X_i \subseteq \mathbb{R}^n,\ i \in V$, are the local constraints. For an unconstrained optimization problem, we let $X_i = \mathbb{R}^n,\ i \in V$.

One of the distributed ways to solve the convex optimization problem (3.1) is to use the distributed subgradient method [18, 33, 34]

$$x_i(k+1) = P_{X_i}(\sum_{j=1}^{\bar{m}} w_{ij}(k)x_j(k) - \alpha(k)g_i(k)), \tag{3.2}$$

where $x_i(k)$ is agent i's estimate of the minimizer of the global objective function f at the kth iteration, $w_{ij}(k)$ is the (i,j)th entry of the weight matrix $A(k)$ at the kth iteration, $\alpha(k)$ is the step size, $g_i(k)$ is the subgradient of the local objective function f_i at $\sum_{j=1}^{\bar{m}} w_{ij}(k)x_j(k)$, and P_{X_i} is the projection operator onto X_i.

It is proved in [18, 34] that the algorithm (3.2) can guarantee that the agents' estimates converge to a minimizer of the global objective function with step sizes $\alpha(k)$

satisfying $\alpha(k) > 0$, $\sum_{k=0}^{\infty} \alpha(k) = +\infty$ and $\sum_{k=0}^{\infty} \alpha(k)^2 < \infty$. But this choice of step sizes excludes an important class of step sizes that is not square summable, which could achieve better convergence rates. In this chapter, we prove that the algorithm (3.2) can also converge without the square summable condition:

Assumption 3 *The step sizes $\alpha(k)$ are positive, vanishing and non-summable, i.e., $\alpha(k) > 0$, $\lim_{k\to\infty} \alpha(k) = 0$ and $\sum_{k=0}^{\infty} \alpha(k) = \infty$.*

Remark 2 *The step sizes under Assumption 3 include*

$$\alpha(k) = D/\sqrt{k+1}, \tag{3.3}$$

which is not square summable, as a special case, where D is a positive constant. This selection of step sizes is proved to achieve the optimal convergence rate for the centralized subgradient algorithm in [36], which might promise a better convergence rate in the distributed case.

We also have the following assumptions for the distributed optimization problem (3.1). The assumption on the objective functions is given as follows:

Assumption 4 *Each local objective function f_i, $i \in V$, is proper convex in its domain.*

For the optimal set that we plan to find, we have the following assumption:

Assumption 5 *The problem (3.1) has a bounded nonempty set of minimizers, denoted by $X^\star$.*

For the constraint sets X_i, $i \in V$, we have the following assumption: for the unconstrained problem $X_i = \mathbb{R}^n$, $\forall i \in V$, and for the constrained problem

Assumption 6 *Each local constraint set X_i, $\forall i \in V$, is bounded, closed and convex if $X_i \neq \mathbb{R}^n$. Their intersection $X = \bigcap_{i=1}^{\bar{m}} X_i$ has interior points.*

In the rest of the chapter, we only consider the constrained case under Assumption 6. Note that the unconstrained case cannot be treated as a special case of the constraint case under Assumption 4. For convenience, when we refer to the constrained case, we actually mean the constrained case under Assumption 4.

As the average of convex functions is also convex, the global objective function f is convex from Assumption 4. With Assumption 6, the constraint set X_i is convex and so is the intersection X. Then the problem (3.1) is a convex optimization problem.

For the subgradients of the objective functions, we suppose that they are bounded, as in the following assumption.

Assumption 7 *The subgradients of f_i, $\forall i \in V$, are uniformly bounded in X_i, i.e., there exists $G > 0$ such that for all $g \in \partial f_i(x)$, $\|g\|_2 \leq G$, $\forall x \in X_i$.*

It is easy to see that Assumption 5 is redundant under Assumption 6. But Assumption 5 is required for the unconstrained case. Assumption 7 is also redundant under Assumption 6, because the subgradients of a convex function is uniformly bounded in a bounded set [2]. But Assumption 7 is required for the unconstrained case. The assumption of uniformly bounded subgradients can be found in many references [12, 18, 33, 34, 56, 57], and plays an important role in the consensus and convergence of the distributed subgradient method.

Remark 3 *In the unconstrained case, with Assumption 7, the local objective functions f_i, $\forall i \in V$, are continuous and thus lower semi-continuous and so is the global objective*

function. As a result, in the unconstrained case, the global objective function is closed. In the constrained case, the local objective functions are also continuous because the subgradients are also bounded due to Assumption 6.

Remark 4 *From Lemma 2, the level sets of a function under Assumptions 4, 5, and 7 are always bounded, which plays an important role in the convergence analysis of the unconstrained case in this chapter.*

For the multi-agent network, we suppose that the communication topologies are balanced as detailed below in addition to the assumption in Chapter 2.

Assumption 8 *The communication topology* $\mathcal{G}(k),\ k = 0, 1, 2, \cdots$ *at each time instant is balanced, i.e.,* $\sum_{j=1}^{\bar{m}} w_{ij}(k) = \sum_{j=1}^{\bar{m}} w_{ji}(k) = 1,\ k = 0, 1, 2, \cdots$.

The balanced graph is necessary for the agents' estimates in the algorithm (3.2) to converge to the minimizer of the problem (3.1). If the communication topology is not balanced, the agents' estimates might not converge to the minimizer of (3.1), as shown in [18].

3.2 Main Results

In this section, we prove that all agents' estimates of the minimizer of the distributed optimization problem (3.1) generated by the distributed subgradient algorithm (3.2) converge to a minimizer of (3.1), with the step sizes in Assumption 3. We also show that the convergence rate is $O(\frac{1}{\sqrt{k}})$ with the step sizes in (3.3) for the best record of the function value at the average of the agents' estimates in the unconstrained case.

For the consensus of the algorithm (3.2) under the wider selection of step sizes, we have the following lemma:

Lemma 10 *(Lemma 8(a) in [34], Proposition 1 in [18]) For a graph sequence* $\mathcal{G}(k),\ k = 0, 1, 2, \cdots$, *satisfying Assumptions 1, 2, and 8 and the optimization problem (3.1) satisfying Assumptions 4, 5, 7 with* $X_i = \mathbb{R}^n,\ i \in V$, *or Assumptions 4 and 6, the agent's estimates* $x_i,\ \forall i \in V$, *in the distributed subgradient algorithm (3.2) reach a consensus, i.e.,* $\lim_{k\to\infty} \|x_i(k) - x_j(k)\|_2 = 0,\ \forall i, j \in V$ *under Assumption 3.*

Remark 5 *Let*

$$\phi^i(k) = P_{X_i}[\sum_{j=1}^{\bar{m}} w_{ij}(k)x_j(k) - \alpha(k)g_i(k)] - [\sum_{j=1}^{\bar{m}} w_{ij}(k)x_j(k) - \alpha(k)g_i(k)]. \tag{3.4}$$

For the unconstrained optimization problem, we can exploit the proof of Lemma 8(a) in [34] by replacing $\phi^i(k)$ *with* 0. *For the constrained optimization problem, we can exploit the proof of Proposition 1 in [18] by noting that the condition* $\sum_{k=0}^{\infty} \alpha(k)^2 < \infty$ *is only used to conclude* $\lim_{k\to\infty} \alpha(k) = 0$ *in the proofs of Lemmas 11, 12 and Proposition 1 in [18]. The proofs of Lemmas 11 and 12 and Proposition 1 in [18] are hence still valid with* $\sum_{k=0}^{\infty} \alpha(k)^2 < \infty$ *replaced with* $\lim_{k\to\infty} \alpha(k) = 0$. *In summary, the step sizes in Assumption 3 can guarantee the consensus result in both unconstrained and constrained cases.*

However, it is not clear whether the agents' estimates will converge to the minimizer with Assumption 3. Next, we will prove the convergence of (3.2) to the minimizer of (3.1) in both the unconstrained and constrained cases. The rigorous statement is as follows:

Theorem 1 *For a graph sequence* $\mathcal{G}(k),\ k = 0, 1, 2, \cdots$, *satisfying Assumptions 1, 2, and 8 and the optimization problem (3.1) satisfying Assumptions 3, 4, 5, and 7 with* $X_i =$

$\mathbb{R}^n$, $\forall i \in V$, *or Assumptions 3, 4, and 6, the agents' estimates* x_i, $\forall i \in V$, *in the distributed subgradient algorithm (3.2) converge to the optimal set* $X^\star$ *of (3.1).*

Theorem 1 shows the convergence of the distributed subgradient algorithm (3.2) under the wider selections of step sizes. Moreover, the fastest convergence rate can be achieved with a special form of the step sizes in Assumption 3, as stated in the following theorem.

Theorem 2 *Let*

$$e_N = \min_{M \le k \le N} f(y(k)) - f(x^\star), \tag{3.5}$$

where $M = \lfloor \frac{N}{2} \rfloor$, f *is defined in (3.1),* $x^\star \in X^\star$, *and*

$$y(k) = \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} x_i(k). \tag{3.6}$$

Then under Assumptions 4, 5, 7, 1, 8, and 2 and $X_i = \mathbb{R}^n$, $e_N = O(\frac{1}{\sqrt{N}})$ *with* $\alpha(k)$ *selected in (3.3).*

Remark 6 *The convergence rate in Theorem 2 is the same as the optimal one in the centralized case in [36], which implies that Theorem 2 might also provide the best convergence rate in the distributed case.*

Next, we will prove Theorem 1 for the unconstrained case in Section 3.2.1, for the constrained case in Section 3.2.2, and Theorem 2 in Section 3.2.3.

3.2.1 Proof of Convergence for Unconstrained Case

In this part we will first provide some necessary lemmas and then prove Theorem 1 for the unconstrained case.

Let $x^\star$ be some point in the optimal set $X^\star$, and $y(k)$ be as in (3.6), and

$$v_i(k) = \sum_{j=1}^{\bar{m}} w_{ij}(k) x_j(k) \tag{3.7}$$

be the local weighted average of the estimates of agent i's neighbors. Then we have the following lemma on the change of the distance from $y(k)$ to $x^\star$.

Lemma 11 *Let $y(k)$ be as in (3.6), $v_i(k)$ be as in (3.7), and $x_i(k)$ be generated by (3.2). Under Assumptions 4, 5, 7, and 8, we have that*

$$\begin{aligned} \|y(k+1) - x^\star\|_2^2 \leq & \|y(k) - x^\star\|_2^2 + 4\alpha(k) \sum_{j=1}^{\bar{m}} \frac{1}{\bar{m}} G \|y(k) - v_j(k)\|_2 \\ & + \alpha(k)^2 G^2 - 2\alpha(k)[f(y(k)) - f(x^\star)], \end{aligned} \tag{3.8}$$

where f is defined in (3.1).

Proof. we have

$$\begin{aligned} y(k+1) &= \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} x_i(k+1) \\ &= \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} (\sum_{j=1}^{\bar{m}} w_{ij}(k) x_j(k) - \alpha(k) g_i(k)) \\ &= \frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} (\sum_{i=1}^{\bar{m}} w_{ij}(k)) x_j(k) - \alpha(k) \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} g_i(k) \\ &= \frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} x_j(k) - \alpha(k) \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} g_i(k) \\ &= y(k) - \alpha(k) \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} g_i(k), \end{aligned}$$

where we use Assumption 8 to obtain the fourth equality.

Then the distance between the global weighted average $y(k)$ and the point $x^\star$ in

the optimal set evolves as follows

$$\begin{aligned}\|y(k+1)-x^\star\|^2 &= \|y(k)-\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j(k)-x^\star\|^2 \\ &= \|y(k)-x^\star\|^2+\alpha(k)^2\|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j(k)\|^2-2\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j^T(k)(y(k)-x^\star).\end{aligned}$$

According to Assumption 7, $\|g_j(k)\|\leq G$. Note that $g_j^T(k)(y(k)-v_j(k))\geq -\|g_j(k)\|\|y(k)-v_j(k)\|\geq -G\|y(k)-v_j(k)\|$ and $f_j(v_j(k))-f_j(y(k))\geq g_j^T(y(k))(v_j(k)-y(k))$ from the definition of subgradients in (2.2), we have

$$\begin{aligned}&\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j^T(k)(y(k)-x^\star)\\ =&\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j^T(k)(y(k)-v_j(k))+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j^T(k)(v_j(k)-x^\star)\\ \geq&-\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} G\|y(k)-v_j(k)\|+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(v_j(k))-f_j(x^\star))\\ =&-\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} G\|y(k)-v_j(k)\|+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(v_j(k))-f_j(y(k)))+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(y(k))-f_j(x^\star))\\ \geq&-\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} G\|y(k)-v_j(k)\|+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j^T(y(k))(v_j(k)-y(k))+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(y(k))-f_j(x^\star))\\ \geq&-2G\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\|y(k)-v_j(k)\|+\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(y(k))-f_j(x^\star)).\end{aligned}$$

Combining with the fact that $\|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} g_j(k)\|^2\leq\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\|g_j(k)\|^2\leq G^2$, we have

$$\begin{aligned}\|y(k+1)-x^\star\|^2\leq&\|y(k)-x^\star\|^2+\alpha(k)^2G^2+4\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} G\|y(k)-v_j(k)\|\\ &-2\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(y(k))-f_j(x^\star)).\end{aligned}$$

■

Using Lemma 11, we can obtain the following lemma:

Lemma 12 *Under Assumptions 1, 2, 3, 4, 5, 7, and 8, there exists a subsequence* $\{y(k_p)\}$ *of* $\{y(k)\}$*, such that* $\lim_{k_p\to\infty} f(y(k_p)) = f(x^\star)$*, where* $f(\cdot)$ *is defined in (3.1).*

Proof. We first prove by contradiction that

$$\liminf_{k\to\infty} f(y(k)) - f(x^\star) \leq 0.$$

Suppose not. Then there exist $\epsilon > 0$ and $K_\epsilon \in \mathbb{N}^+$, such that for all $k > K_\epsilon$,

$$f(y(k)) - f(x^\star) \geq \epsilon. \tag{3.9}$$

Then

$$\begin{aligned}
&\|y(k+1) - x^\star\|_2^2 \\
\leq &\|y(k) - x^\star\|_2^2 + \alpha(k)^2 G^2 + 4\frac{1}{\bar{m}} G\alpha(k) \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2 - 2\alpha(k)\epsilon \\
= &\|y(k) - x^\star\|_2^2 - \alpha(k)\epsilon + G^2[\alpha(k)^2 + \frac{1}{G^2}(4\frac{1}{\bar{m}} G \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2 - \epsilon)\alpha(k)].
\end{aligned}$$

Under Assumptions 1, 2, 3, 4, 7, and 8, it follows from Lemma 10 that $\lim_{k\to\infty} \|x_i(k) - x_j(k)\|_2 = 0$ and thus

$$\lim_{k\to\infty} \|v_i(k) - y(k)\|_2 = 0. \tag{3.10}$$

It follows that there exists $K_c \in \mathbb{N}^+$, such that for all $k > K_c$,

$$\frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2 \leq \frac{\epsilon}{8G}.$$

Then we have

$$\|y(k+1) - x^\star\|_2^2 \leq \|y(k) - x^\star\|_2^2 - \alpha(k)\epsilon + G^2(\alpha(k)^2 - \frac{\epsilon}{2G^2}\alpha(k)).$$

As $\alpha(k)$ vanishes from Assumption 3, there exists $K_\alpha \in \mathbb{N}^+$, such that for all $k > K_\alpha$, $\alpha(k) \leq \frac{\epsilon}{2G^2}$ and hence $\alpha(k)^2 - \frac{\epsilon}{2G^2}\alpha(k) \leq 0$. Denote $K_0 = \max\{K_\epsilon, K_c, K_\alpha\}$. Then for all $k > K_0$, (3.9) holds and it follows that

$$\|y(k+1) - x^\star\|_2^2 \leq \|y(k) - x^\star\|_2^2 - \alpha(k)\epsilon. \tag{3.11}$$

It then follows that

$$\|y(K_0+m) - x^\star\|_2^2 \leq \|y(K_0+1) - x^\star\|_2^2 - \epsilon \sum_{t=K_0+1}^{K_0+m-1} \alpha(t).$$

As $\sum_{k=1}^{\infty} \alpha(k) = \infty$ from Assumption 3,

$$\|y(K_0+m) - x^\star\|_2^2 \leq \|y(K_0+1) - x^\star\|_2^2 - \epsilon \sum_{t=K_0+1}^{K_0+m-1} \alpha(t) < 0$$

when m is sufficiently large. This contradicts with the fact that $\|y(K_0+m) - x^\star\|_2^2 \geq 0$. So $\liminf_{k\to\infty} f(y(k)) - f(x^\star) \leq 0$.

Note that $f(y(k)) - f(x^\star) \geq 0$ because $x^\star$ is in the optimal set $X^\star$. As a result, $\liminf_{k\to\infty} f(y(k)) - f(x^\star) = 0$.

We can also know that f is continuous form Assumption 7 (see Remark 3). So there exists a subsequence $\{y(k_p)\}$ of $\{y(k)\}$ such that

$$\lim_{k_p\to\infty} f(y(k_p)) = f(x^\star). \tag{3.12}$$

■

Then we have the following lemma on the convergence of $\{y(k_p)\}$ itself:

Lemma 13 *Under Assumptions 1, 2, 3, 4, 5, 7, and 8, there exists a subsequence $\{y(k_q)\}$ of $\{y(k_p)\}$, such that $\{y(k_q)\}$ converges to some point in $X^\star$, where $\{y(k)\}$ is defined in (3.6) and $\{y(k_p)\}$ is the subsequence in Lemma 12.*

Proof. If $\{y(k_p)\}$ is bounded or has a bounded subsequence, the existence of a convergent subsequence is obvious.

If $\|y(k_p)\| \to \infty$, let $k_p = N$ for some N and $\epsilon_N = f(y(N)) - f(x^\star)$. Define $U_{\epsilon_N} = \{x : f(x) = f(x^\star) + \epsilon_N\}$. Note that U_{ϵ_N} is bounded from Remark 4, which is a result from Lemma 2, Assumptions 4, 5, and 7. As $\|y(k_p)\| \to \infty$ and $\lim_{k_p \to \infty} f(y(k_p)) = f(x^\star)$ from (3.12) under Assumptions 1, 2, 3, 4, 5, 7, and 8, we can find $k_p = N_1$ for some sufficiently large N_1 such that $y(k_p)$ is outside the level curve U_{ϵ_N} and $f(y_{k_p}) < f(x^\star) + \epsilon_N$. Consider the intersection point x_{ints} of U_{ϵ_N} and the line segment between $x^\star$ and $y(N_1)$. Here x_{ints} can be expressed as a convex combination of $x^\star$ and $y(N_1)$, that is, there exists $0 < \alpha < 1$ such that $x_{ints} = \alpha x^\star + (1-\alpha) y(N_1)$. Also we have $f(x_{ints}) = f(y(N)) > f(y(N_1))$ and $f(x_{ints}) > f(x^\star)$. Then

$$f(x_{ints}) = f(\alpha x^\star + (1-\alpha) y(N_1)) > \alpha f(x^\star) + (1-\alpha) f(y(N_1)),$$

which contradicts the convexity of $f(x)$. So $\{y(k_p)\}$ has a bounded subsequence, and thus a convergent subsequence denoted as $\{y(k_q)\}$. From the continuity of f, $\{y(k_q)\}$ converges to some point in $X^\star$. ■

Without loss of generality, we suppose that the convergent subsequence $\{y(k_q)\}$ is $\{y(k_p)\}$ itself.

Also, we have the following lemma on the level curve of f:

Lemma 14 *Define $U_\delta = \{y : f(y) - f(x^\star) = \delta\}$ as the level curve of the global objective function, where $f(\cdot)$ is defined in (3.1). Let*

$$d(\delta) = \max_{y \in U_\delta} \min_{p \in X^\star} \|y - p\|_2 \tag{3.13}$$

be the maximum distance from the level curve U_δ to the optimal set $X^\star$. Under Assumptions 4 and 5, $\lim_{\delta \to 0} d(\delta) = 0$.

Proof. We prove by contradiction. Suppose that $\lim_{\delta \to 0} d(\delta) = 0$ does not hold. Then there exists a sequence $\delta_k \to 0$ and a constant $\epsilon > 0$ such that $d(\delta_k) \geq \epsilon$. Let $x^\star_{\delta_k} \in X^\star$ and $y_{\delta_k} \in U_{\delta_k}$ be the points such that $\|y_{\delta_k} - x^\star_{\delta_k}\| = d(\delta_k)$. As U_{δ_k} is bounded from Remark 4, which is obtained from Lemma 2 and Assumptions 4, 5, and 7, the sequence $\{y_{\delta_k}\}$ is also bounded and thus has a convergent subsequence. Without loss of generality, suppose that this convergent subsequence is $\{y_{\delta_k}\}$ itself and its limit point is $y^\star$. Then we have $d(y^\star, X^\star) \geq \epsilon$. By the continuity of the function f, we know $f(y^\star) = \lim_{k \to \infty} f(y_{\delta_k})$. With $f(x^\star) < f(y_{\delta_k}) = f(x^\star) + \delta_k$, we have that $f(y^\star) = f(x^\star)$. Then the level set $\{y : f(y) \leq f(x^\star)\}$ is not connected and thus not convex. This leads to a contradiction with the convexity of f in Assumption 4. It is hence concluded that $\lim_{\delta \to 0} d(\delta) = 0$. ■

Remark 7 *From the proofs of Lemmas 13 and 14, we can see that Assumption 5 on bounded optimal set together with Assumption 7 on bounded subgradients plays an important role in the analysis of the unconstrained case.*

Proof of Theorem 1 for unconstrained case Next, we will show that if k_p is sufficiently large, then for all $k \geq k_p$, $y(k)$ stays either inside U_δ or outside but close to U_δ.

From Assumption 3, there exists $K^{'}_{\alpha} \in \mathbb{N}^+$ such that for all $k > K^{'}_{\alpha}$, $\alpha(k) \leq \frac{\delta}{2G^2}$. From (3.10), there exists $K^{'}_{c} \in \mathbb{N}^+$ such that for all $k > K^{'}_{c}$, $\frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2 \leq \frac{\delta}{8G}$. Then we consider two cases.

1) If $f(y(k)) - f(x^\star) < \delta$ at any iteration k, then $\min_{p \in X^\star} \|y(k) - p\|_2 \leq d(\delta)$ at the

iteration k. It follows that

$$\begin{aligned}\min_{p\in X^\star}\|y(k+1)-p\|_2 &= \min_{p\in X^\star}\|y(k)-\frac{1}{\bar{m}}\alpha(k)\sum_{i=1}^{\bar{m}}g_i(k)-p\|_2\\ &\le \min_{p\in X^\star}\|y(k)-p\|_2+\alpha(k)\|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}g_j(k)\|_2\\ &\le d(\delta)+\alpha(k)\|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}g_j(k)\|_2\le d(\delta)+\alpha(k)G\end{aligned}$$

at the iteration k.

2) If $f(y(k))-f(x^\star)\ge\delta$ at any iteration k, it follows from (3.11) that when the iteration k is greater than $\max\{K_\alpha^{'},K_c^{'}\}$,

$$\min_{p\in X^\star}\|y(k+1)-p\|_2^2\le\min_{p\in X^\star}\|y(k)-p\|_2^2-\alpha(k)\delta\le\min_{p\in X^\star}\|y(k)-p\|_2^2$$

at the iteration k.

Remark 8 *In the proof of Lemma 12 in the appendix, a contradiction is obtained from (3.11) under the assumption that (3.9) holds for all k larger than some $K_0\in\mathbb{N}^+$. But in Case 2 above, we only consider the case when $f(y(k))-f(x^\star)\ge\delta$ holds at one iteration k, which is greater than $\max\{K_\alpha^{'},K_c^{'}\}$. So Case 2 above does not conflict with the proof of Lemma 12 and would not lead to a contradiction.*

From (3.12), for any $\delta>0$, there exists $K_{p,\delta}\in\mathbb{N}^+$ such that for all $k_p>K_{p,\delta}$, $f(y(k_p))-f(x^\star)<\delta$. Then from the two cases considered above, we have that for all $k>\max\{K_\alpha^{'},K_c^{'},K_{p,\delta}\}$,

$$\begin{aligned}\min_{p\in X^\star}\|y(k+1)-p\|_2 &\le \min_{p\in X^\star}\|y(k)-p\|_2+\max_{k>\max\{K_\alpha^{'},K_c^{'}\}}\{\alpha(k)\}G\\ &\le d(\delta)+\frac{\delta}{2G}.\end{aligned}$$

Under Assumptions 4 and 5, it follows from Lemma 14 that $\lim_{\delta\to 0}\min_{p\in X^\star}\|y(k)-p\|_2=0$, which means that $y(k)$ converges to some point in the optimal set $X^\star$. Finally under Assumptions 1, 2, 3, 4, 5, 7, and 8, it follows from Lemma 10 that

$$\lim_{k\to\infty}\min_{p\in X^\star}\sum_{i=1}^{\bar{m}}\|x_i(k)-p\|_2\leq\lim_{k\to\infty}\min_{p\in X^\star}\sum_{i=1}^{\bar{m}}(\|y(k)-p\|_2+\|x_i(k)-y(k)\|_2)=0,$$

which means that the estimates of all agents converge to $X^\star$.

3.2.2 Proof of Convergence for Constrained Case

In this part we will first provide some necessary lemmas and then prove Theorem 1 for the constrained case.

Let $x^\star$ be some point in the optimal set $X^\star$ of the problem (3.1) in the constrained case. Let $v_i(k)$ be defined in (3.7) and

$$y(k)=P_X(\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}x_i(k)). \tag{3.14}$$

Then for the distance between agents' estimates and $x^\star$, we have the following lemma:

Lemma 15 *Let $x_i(k)$ be generated by (3.2) and $y(k)$ be as in (3.14). Under Assumptions 4, 6, and 8, we have that*

$$\begin{aligned}&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|_2^2\\ \leq&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2^2+\alpha(k)^2G^2+\frac{2}{\bar{m}}\alpha(k)\sum_{i=1}^{\bar{m}}\|v_i(k)-y(k)\|_2\\ &-2\alpha(k)[f(y(k))-f(x^\star)],\end{aligned} \tag{3.15}$$

where G denotes the upper bound of the subgradients of f_i in Assumption 6.

Proof. Let $x^\star$ be some point in the optimal set $X^\star$ of the problem (3.1). Let $v_i(k)$ be defined in (3.7). Then we have

$$\begin{aligned}
&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|^2\\
=&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|P_{X_i}(\sum_{j=1}^{\bar{m}}w_{ij}(k)x_j(k)-\alpha(k)g_i(k))-x^\star\|^2\\
\leq&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|\sum_{j=1}^{\bar{m}}w_{ij}(k)x_j(k)-\alpha(k)g_i(k)-x^\star\|^2\\
=&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|\sum_{j=1}^{\bar{m}}w_{ij}(k)x_j(k)-x^\star\|^2+\alpha(k)^2\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|g_i(k)\|^2-2\alpha(k)\sum_{j=1}^{\bar{m}}q_jg_j^T(k)(v_j(k)-x^\star),
\end{aligned}$$

where the inequality is obtained from Lemma 3. As $\|\cdot\|^2$ is convex, we have

$$\begin{aligned}
\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|\sum_{j=1}^{\bar{m}}w_{ij}(k)x_j(k)-x^\star\|^2&\leq\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\sum_{j=1}^{\bar{m}}w_{ij}(k)\|x_j(k)-x^\star\|^2\\
&=\sum_{j=1}^{\bar{m}}(\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}w_{ij}(k))\|x_j(k)-x^\star\|^2\\
&=\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\|x_j(k)-x^\star\|^2,
\end{aligned}$$

where we use Assumption 8 to obtain the first equality. Because $\|g_i(k)\|\leq G$ under Assumption 7, it follows that

$$\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|^2\leq\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_j(k)-x^\star\|^2+\alpha(k)^2G^2-2\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}g_j^T(k)(v_j(k)-x^\star)$$

As $f_j,\ j\in V$, are convex, $f_j(v_j(k))-f(x^\star)\leq g_j^T(k)(v_j(k)-x^\star)$, where $g_j(k)$ is a subgradient of f_j at $v_j(k)$. We thus have

$$\begin{aligned}
&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|^2\\
\leq&\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|^2+\alpha(k)^2G^2-2\alpha(k)\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(f_j(v_j(k))-f_j(x^\star)).
\end{aligned}\tag{3.16}$$

Also, we have that

$$\begin{aligned} f_j(v_j(k)) - f_j(x^\star) &= f_j(v_j(k)) - f_j(y(k)) + f_j(y(k)) - f_j(x^\star) \\ &\geq -G\|v_j(k) - y(k)\| + f_j(y(k)) - f_j(x^\star). \end{aligned} \tag{3.17}$$

Combining (3.16) and (3.17), we obtain that

$$\begin{aligned} \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1) - x^\star\|^2 \leq & \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_j(k) - x^\star\|^2 + \alpha(k)^2G^2 + \frac{2}{\bar{m}}\alpha(k)\sum_{i=1}^{\bar{m}}\|v_i(k) - y(k)\|_2 \\ & - 2\alpha(k)[f(y(k)) - f(x^\star)]. \end{aligned}$$

■

Then similar to Lemma 12, we have the following lemma:

Lemma 16 *Under Assumptions 1, 2, 3, 4, 6, and 8, there exists a subsequence $\{y(k_p)\}$ of $\{y(k)\}$, such that $\lim_{k_p\to\infty} f(y(k_p)) - f(x^\star) = 0$, where $y(k)$ is defined in (3.14).*

Proof. We first prove that $\liminf_{k\to\infty} f(y(k)) - f(x^\star) \leq 0$ by contradiction. If not, there exist $\epsilon > 0$ and $K_\epsilon \in \mathbb{N}^+$, such that $\forall k > K_\epsilon$, $f(y(k)) - f(x^\star) > \epsilon$. Then we have

$$\begin{aligned} \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1) - x^\star\|^2 &\leq \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k) - x^\star\|^2 + \alpha(k)^2G^2 - 2\alpha(k)\epsilon \\ &= \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k) - x^\star\|^2 - \alpha(k)\epsilon + (\alpha(k)^2G^2 - \alpha(k)\epsilon). \end{aligned}$$

As $\lim_{k\to\infty}\alpha(k) = 0$, there exists $K_\alpha \in \mathbb{N}^+$, such that for all$k > K_\alpha$, $0 < \alpha(k) < \frac{\epsilon}{G^2}$, which implies that $\alpha(k)^2G^2 - \alpha(k)\epsilon < 0$. Hence for all $k > K = \max(K_\epsilon, K_\alpha)$, we have

$$\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1) - x^\star\|^2 \leq \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k) - x^\star\|^2 - \alpha(k)\epsilon. \tag{3.18}$$

Because $\sum_{k=1}^{\infty}\alpha(k) = \infty$, it follows that when k_0 is sufficiently large

$$\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(K+1+k_0) - x^\star\|^2 \leq \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(K+1) - x^\star\|^2 - \sum_{t=K+1}^{K+k_0}\alpha(t)\epsilon < 0.$$

This contradicts with $\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|^2 \geq 0$. It can thus be concluded that $\liminf_{k\to\infty} f(y(k))-f(x^\star) \leq 0$.

We can also know that f is continuous form Assumption 7 (see Remark 3). Therefore, there exists a subsequence $\{y(k_p)\}$ of $\{y(k)\}$, such that $\lim_{k_p\to\infty} f(y(k_p)) - f(x^\star) = 0$.

■

Proof of Theorem 1 for constrained case As $\{y(k_p)\} \in X$ is uniformly bounded from Assumption 6, $\{y(k_p)\}$ has a convergent subsequence. Without loss of generality, suppose that the convergent subsequence is $\{y(k_p)\}$ itself, with y_∞ being its limit point. We also know that $y_\infty \in X^\star$ from (16) and Remark 3. Without loss of generality, let $x^\star = y_\infty$. Then we get from Lemma 10 that

$$\lim_{k_p\to\infty}\|x_i(k_p)-x^\star\|_2 \leq \lim_{k_p\to\infty}(\|x_i(k_p)-y(k_p)\|_2 + \|y(k_p)-x^\star\|_2) = 0. \tag{3.19}$$

So the subsequence $\{x_i(k_p)\}$, $\forall i \in V$, converge to the optimal set $X^\star$.

Remark 9 *In the constrained case, the convergence of $\{y(k_p)\}$ results directly from Assumption 6. But in the unconstrained case, we have to impose the bounded optimal set in Assumption 5 and bounded subgradients in Assumption 7 to prove the convergence of $\{y(k_p)\}$ in Lemma 13.*

We then prove the convergence of the estimates $\{x_i(k)\}, \forall i \in V$, to $X^\star$. Define $U_\delta = \{y \in X : f(y) - f(x^\star) \leq \delta\}$ and $d(\delta) = \max_{y\in U_\delta}\min_{x^\star\in X^\star}\|y-x^\star\|_2$. Then we have that $\lim_{\delta\to 0} d(\delta) = 0$ with a similar proof to that of Lemma 14 under Assumptions 4 and 6. There exists $K_\alpha^{'} \in \mathbb{N}^+$ and $K_c^{'} \in \mathbb{N}^+$, such that for all $k > K_\alpha^{'}$, $\alpha(k) \leq \frac{\delta}{2G^2}$ under Assumption 3,

and for all $k > K_c^{'}$, $\frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2 \leq \frac{\delta}{4G}$ from Lemma 10. For the difference between $\|y(k) - x^\star\|$ and $\|x_i(k) - x^\star\|$, we have that

$$\begin{aligned}
&|\|y(k) - x^\star\|_2^2 - \|x_i(k) - x^\star\|_2^2| \\
=&|\|y(k) - x^\star\|_2 - \|x_i(k) - x^\star\|_2|(\|y(k) - x^\star\|_2 + \|x_i(k) - x^\star\|_2) \\
\leq&\|y(k) - x_i(k)\|_2(\|y(k) - x^\star\|_2 + \|x_i(k) - x^\star\|_2) \\
\leq&\|y(k) - x_i(k)\|_2(\max_{p,q\in X} \|p - q\|_2 + \max_{u,v\in X_i} \|u - v\|_2),
\end{aligned}$$

where the first inequality is obtained from the triangle inequality. Notice that the terms after the last inequality are irrespective of $x^\star$ and bounded under Assumption 6, there exists $K_c^{''} \in \mathbb{N}^+$ such that for all $k > K_c^{''}$,

$$|\|y(k) - x^\star\|_2^2 - \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} \|x_i(k) - x^\star\|_2^2| \leq \frac{\delta^2}{4}, \ \forall x^\star \in X^\star \tag{3.20}$$

from Lemma 10. Then we consider two cases.

1) If $f(y(k)) < f(x^\star) + \delta$, then

$$\begin{aligned}
\frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} \|x_i(k+1) - x^\star\|_2^2 &= \frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} \| \sum_{j=1}^{\bar{m}} w_{ij}(k) x_j(k) - \alpha(k) g_i(k) - x^\star\|_2^2 \\
&\leq 2(\frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} \| \sum_{j=1}^{\bar{m}} w_{ij}(k) x_j(k) - x^\star\|_2^2 + \alpha(k)^2 G^2) \\
&\leq 2(\frac{1}{\bar{m}} \sum_{i=1}^{\bar{m}} \|x_i(k) - x^\star\|_2^2 + \alpha(k)^2 G^2),
\end{aligned}$$

where the first inequality is obtained from the fact that $\|a + b\|_2^2 \leq 2(\|a\|_2^2 + \|b\|_2^2)$ and the

second inequality from the convexity of $\|\cdot\|_2^2$ and Assumption 8. So when $k > K_c^{''}$,

$$\begin{aligned}\min_{x^\star\in X^\star}\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|_2^2 &\le \min_{x^\star\in X^\star} 2(\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2^2+\alpha(k)^2G^2)\\ &\le 2\min_{x^\star\in X^\star}(\|y(k)-x^\star\|_2^2+\frac{\delta^2}{4}+\alpha(k)^2G^2)\\ &\le 2(d(\delta)^2+\frac{\delta^2}{4}+\alpha(k)^2G^2),\end{aligned}$$

where (3.20) is used to obtain the second inequality.

Remark 10 *In Case 1) above, we make use of (3.20), which results from Assumption 6. As a result, the analysis of Case 1) above cannot to applied to the unconstrained case where* $X_i = \mathbb{R}^n$.

2) If $f(y(k)) \ge f(x^\star)+\delta$, it follows from (3.18) that when $k > \max\{K_\alpha^{'}, K_c^{'}\}$,

$$\begin{aligned}\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|_2^2 &\le \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2^2-\alpha(k)\delta\\ &< \frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2^2.\end{aligned}$$

Then we have that $\min_{x^\star\in X^\star}\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|_2^2 \le \min_{x^\star\in X^\star}\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2^2$.
From (16), for arbitrary $\delta > 0$, there exists $K_{p,\delta} \in \mathbb{N}^+$ such that for all $k_p > K_{p,\delta}$, $f(x(k_p)) - f(x^\star) < \delta$. Then taking into consideration of the two cases above, when $k > \max\{K_\alpha^{'}, K_c^{'}, K_c^{''}, K_{p,\delta}\}$, we have that

$$\begin{aligned}\min_{x^\star\in X^\star}\frac{1}{\bar{m}}\sum_{i=1}^{\bar{m}}\|x_i(k+1)-x^\star\|_2^2 &\le 2(d(\delta)^2+\frac{\delta^2}{4}+\max_{k>\max\{K_\alpha^{'},K_c^{'},K_c^{''}\}}\alpha(k)^2G^2)\\ &\le 2(d(\delta)^2+\frac{\delta^2}{4}+\frac{\delta^2}{4G^2}).\end{aligned}$$

Notice that $\lim_{\delta\to 0} d(\delta) = 0$. Then we can conclude that $\lim_{k\to\infty}\min_{x^\star\in X^\star}\sum_{i=1}^{\bar{m}}\|x_i(k)-x^\star\|_2 = 0,\ \forall i \in V$. So all agents' estimates $\{x_i(k)\},\ \forall i \in V$, converge to the optimal set $X^\star$.

3.2.3 Analysis of Convergence Rate when $\alpha(k) = \frac{D}{\sqrt{k+1}}$

From (3.8), we have that

$$\begin{aligned}
&2\alpha(k)[f(y(k)) - f(x^\star)] \\
\leq &\|y(k) - x^\star\|_2^2 - \|y(k+1) - x^\star\|_2^2 + \alpha(k)^2 G^2 + 4\frac{1}{\bar{m}} G\alpha(k) \sum_{j=1}^{\bar{m}} \|y(k) - v_j(k)\|_2.
\end{aligned}$$

From the proof of Lemma 8(a) in [34] and by replacing $\phi^i(k)$ in (3.4) with 0, we have that

$$\|v_i(k) - y(k)\|_2 \leq C_1\beta^{k-1} + C_2 \sum_{t=0}^{k-2} \beta^{k-t}\alpha(t) + 2\alpha(k)G,$$

where $0 < \beta < 1$, C_1 and C_2 are some constants. So

$$\begin{aligned}
2\alpha(k)[f(y(k)) - f(x^\star)] \leq &\|y(k) - x^\star\|_2^2 - \|y(k+1) - x^\star\|_2^2 + \alpha(k)^2 G^2 \\
&+ 4\alpha(k)G(C_1\beta^{k-1} + C_2 \sum_{t=0}^{k-2} \beta^{k-t}\alpha(t) + 2\alpha(k)G).
\end{aligned}$$

Let e_N be as in (3.5). Then we have that

$$\begin{aligned}
2\sum_{k=M}^{N} \alpha(k) e_N \leq &\|y(M) - x^\star\|_2^2 - \|y(N+1) - x^\star\|_2^2 + \sum_{k=M}^{N} G^2\alpha(k)^2 \\
&+ 4\alpha(k)G \sum_{k=M}^{N} (C_1\beta^{k-1} + C_2 \sum_{t=0}^{k-2} \beta^{k-t}\alpha(t) + 2\alpha(k)G) \\
\leq &\|y(M) - x^\star\|_2^2 + 9G^2 \sum_{k=M}^{N} \alpha(k)^2 + 4GC_1 \sum_{k=M}^{N} \alpha(k)\beta^{k-1} \\
&+ 4GC_2 \sum_{k=M}^{N} \alpha(k) \sum_{t=0}^{k-2} \beta^{k-t}\alpha(t).
\end{aligned}$$

As a result,

$$e_N \leq \frac{\|y(M)-x^\star\|_2^2 + 9G^2\sum_{k=M}^{N}\alpha(k)^2}{2\sum_{k=M}^{N}\alpha(k)} + \frac{4GC_1\sum_{k=M}^{N}\beta^{k-1}\alpha(k) + 4GC_2\sum_{k=M}^{N}\alpha(k)\sum_{t=0}^{k-2}\beta^{k-t}\alpha(t)}{2\sum_{k=M}^{N}\alpha(k)} \tag{3.21}$$

For the denominator of (3.21), as $\alpha(k) = \frac{D}{\sqrt{k+1}}$, we have that

$$\sum_{k=M}^{N}\alpha(k) = \sum_{k=M}^{N}\frac{D}{\sqrt{k+1}} = O(\sqrt{N}),$$

which can be verified by calculating $\int_M^N \frac{1}{\sqrt{t+1}}\mathrm{dt}$. For the numerator of (3.21), $\|y(M)-x^\star\|_2^2$ is $O(1)$ because it is bounded. For the second term, we have that

$$\sum_{k=M}^{N}\alpha(k)^2 = \sum_{k=M}^{N}\frac{D^2}{k+1} = O(1), \tag{3.22}$$

which can be verified by calculating $\int_M^N \frac{1}{t+1}\mathrm{dt}$. For the third term, as $\beta^{k-1}\alpha(k) \leq \beta^{k-1}\frac{D}{\sqrt{M+1}}$, we have that

$$\sum_{k=M}^{N}\beta^{k-1}\alpha(k) \leq \frac{1}{1-\beta}\frac{D}{\sqrt{M+1}} = O(\frac{1}{\sqrt{N}}), \tag{3.23}$$

because $M = \lfloor \frac{N}{2} \rfloor$.

For the last term in the numerator of (3.21), we substitute $\alpha(k)$ and obtain that $\sum_{k=M}^{N}\alpha(k)\sum_{t=0}^{k-2}\beta^{k-t}\alpha(t) = D^2\sum_{k=M}^{N}\frac{1}{\sqrt{k+1}}\sum_{t=0}^{k-2}\beta^{k-t}\frac{1}{\sqrt{t+1}}$. First, we consider the factor $\sum_{t=0}^{k-2}\beta^{k-t}\frac{1}{\sqrt{t+1}}$ and obtain that

$$\begin{aligned}\sum_{t=0}^{k-2}\beta^{k-t}\frac{1}{\sqrt{t+1}} &= \beta^2\sum_{t=0}^{k-2}\beta^{k-2-t}\frac{1}{\sqrt{t+1}} = \beta^2\sum_{t=0}^{p}\beta^{p-t}\frac{1}{\sqrt{t+1}} \\ &= \beta^2[\sum_{t=0}^{\lfloor p/2\rfloor}\beta^{p-t}\frac{1}{\sqrt{t+1}} + \sum_{t=\lfloor p/2\rfloor+1}^{p}\beta^{p-t}\frac{1}{\sqrt{t+1}}],\end{aligned}$$

where $p = k-2$. For the term $\sum_{t=0}^{\lfloor p/2 \rfloor} \beta^{p-t}\frac{1}{\sqrt{t+1}}$, we have that

$$\sum_{t=0}^{\lfloor p/2 \rfloor} \beta^{p-t}\frac{1}{\sqrt{t+1}} \leq \beta^{p-\lfloor p/2 \rfloor} \sum_{t=0}^{\lfloor p/2 \rfloor} \frac{1}{\sqrt{t+1}} \leq \beta^{p/2} \int_0^{\lfloor p/2 \rfloor} \frac{1}{\sqrt{t+1}} dt = 2\beta^{p/2}(\sqrt{\lfloor p/2 \rfloor} - 1).$$

It is easy to verify that for any $1 > \beta_1 > \sqrt{\beta}$, $\lim_{p\to\infty} \frac{\beta^{p/2}(\sqrt{\lfloor p/2 \rfloor}-1)}{\beta_1^p} = 0$. We thus have that $\sum_{t=0}^{\lfloor p/2 \rfloor} \beta^{p-t}\frac{1}{\sqrt{t+1}} \leq 2C_{\beta_1}\beta_1^p$, where C_{β_1} is a constant. Next, we consider the term $\sum_{t=\lfloor p/2 \rfloor+1}^{p} \beta^{p-t}\frac{1}{\sqrt{t+1}}$. We have that

$$\sum_{t=\lfloor p/2 \rfloor+1}^{p} \beta^{p-t}\frac{1}{\sqrt{t+1}} \leq \frac{1}{\sqrt{\lfloor p/2 \rfloor + 2}} \sum_{t=\lfloor p/2 \rfloor+1}^{p} \beta^{p-t} = \frac{1}{\sqrt{\lfloor p/2 \rfloor + 2}} \beta^{2\lfloor p/2 \rfloor - p + 2} \frac{1-\beta^{p-\lfloor p/2 \rfloor}}{1-\beta}.$$

As $1 \leq 2\lfloor p/2 \rfloor - p + 2 \leq 2$ and $\frac{1-\beta^{p-\lfloor p/2 \rfloor}}{1-\beta} \leq \frac{1}{1-\beta}$, there exists a constant C_β such that $\sum_{t=\lfloor p/2 \rfloor+1}^{p} \beta^{p-t}\frac{1}{\sqrt{t+1}} \leq C_\beta \frac{1}{\sqrt{\lfloor p/2 \rfloor+2}}$. Noticing $p = k-2$, we have that

$$\begin{aligned}\sum_{k=M}^{N} \alpha(k) \sum_{t=0}^{k-2} \beta^{k-t}\alpha(t) \leq & 2C_{\beta_1} \sum_{k=M}^{N} \frac{1}{\sqrt{k+1}} \beta_1^{(k-2)} + C_\beta \sum_{k=M}^{N} \frac{1}{\sqrt{k+1}} \frac{1}{\sqrt{(k+3)/2}} \\ \leq & 2C_{\beta_1} \sum_{k=M}^{N} \frac{1}{\sqrt{k+1}} \beta_1^{(k-2)} + \sqrt{2}C_\beta \sum_{k=M}^{N} \frac{1}{k+1},\end{aligned}$$

where the first term is $O(\frac{1}{\sqrt{N}})$ as analyzed in (3.22) and the second $O(1)$ as analyzed in (3.23).

Taking into consideration all the terms in numerator and denominator, we hence have that $e_N = \frac{O(1)+O(1)+O(\frac{1}{\sqrt{N}})+O(\frac{1}{\sqrt{N}})+O(1)}{O(\sqrt{N})} = O(\frac{1}{\sqrt{N}})$.

3.3 Simulations

In this section, we illustrate the effectiveness of the positive, vanishing and non-summable step sizes via a simulation example.

The multi-agent system is composed of seven agents, and the topology of the network is an undirected ring. The weights in the weight matrix are selected as $w_{ij} = \frac{1}{3}$ if there is an edge (j, i) and $w_{ij} = 0$ otherwise. We choose $f_1(x) = \|x - 3\|_1, f_2(x) = \|x-2\|_1, f_3(x) = \|x-1\|_1, f_4(x) = \|x\|_1, f_5(x) = \|x+1\|_1, f_6(x) = \|x+2\|_1, f_7(x) = \|x+3\|_1$. Here, $f_i,\ i = 1, 2, \cdots, 7$ are selected as non-differentiable. The step sizes are selected as $\alpha(k) = \frac{1}{\sqrt{k+1}},\ k = 0, 1, 2, \cdots$, which is positive, vanishing, non-summable but not square summable, i.e., $\sum_{k=1}^{\infty} \alpha(k)^2 = \infty$. The initial estimates of the agents are generated randomly.

3.3.1 Convergence Result

In this part, we show the convergence result of the simulation example in the constrained case. We select the constraint sets as $X_1 = [-1, 5], X_2 = [-0.5, 3], X_3 = [-4, 2], X_4 = [-3.7, 4], X_5 = [-7, 0.5], X_6 = [-5, 1.7], X_7 = [-8, 0.9]$, which satisfy Assumption 6.

The simulation result is shown in Fig. 3.1. From Fig. 3.1, we can see that the norms of the errors, i.e., the distances between the agents' estimates and the optimizer, diminish to zero for all agents. This simulation example shows that when the step sizes are not square summable but positive, vanishing and non-summable, the distributed subgradient algorithm (3.2) can make the agents' estimates converge to a common minimizer.

3.3.2 Convergence Rate Comparison

In this part, we compare the convergence rate of e_N defined in (3.5) for the unconstrained case between $\alpha(k) = \frac{1}{\sqrt{k+1}}$ and $\alpha(k) = \frac{1}{k+1}$, which is square summable as required

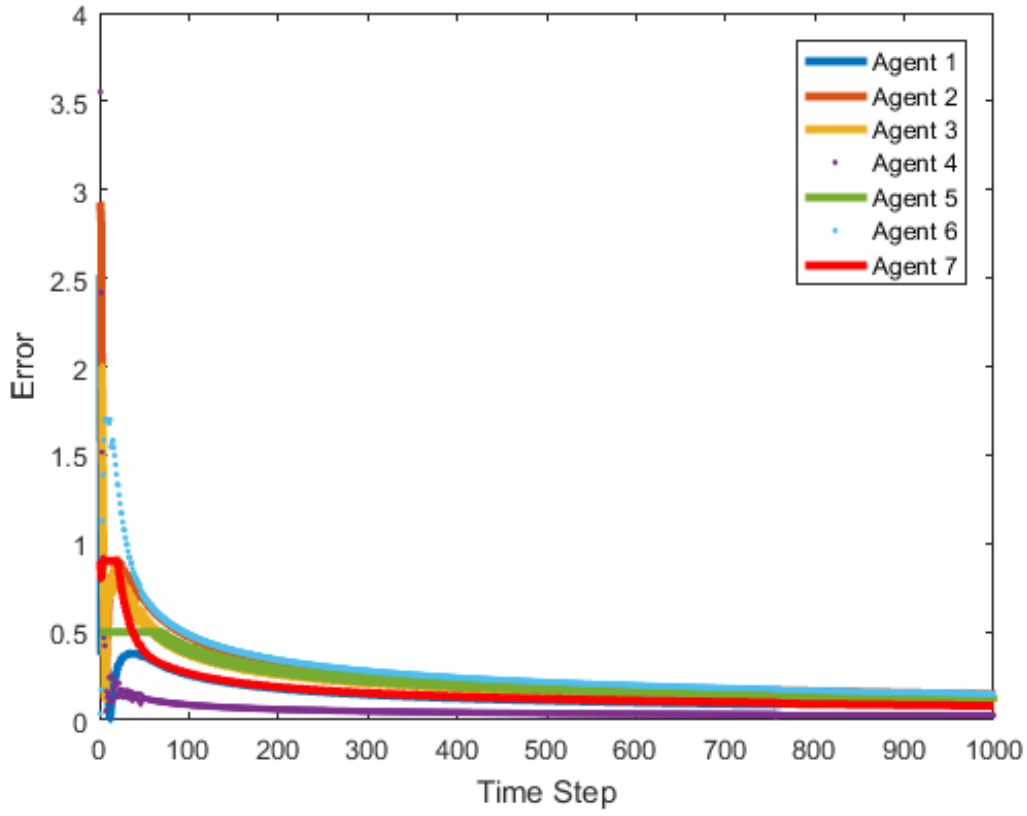

Figure 3.1: Norm of estimate errors

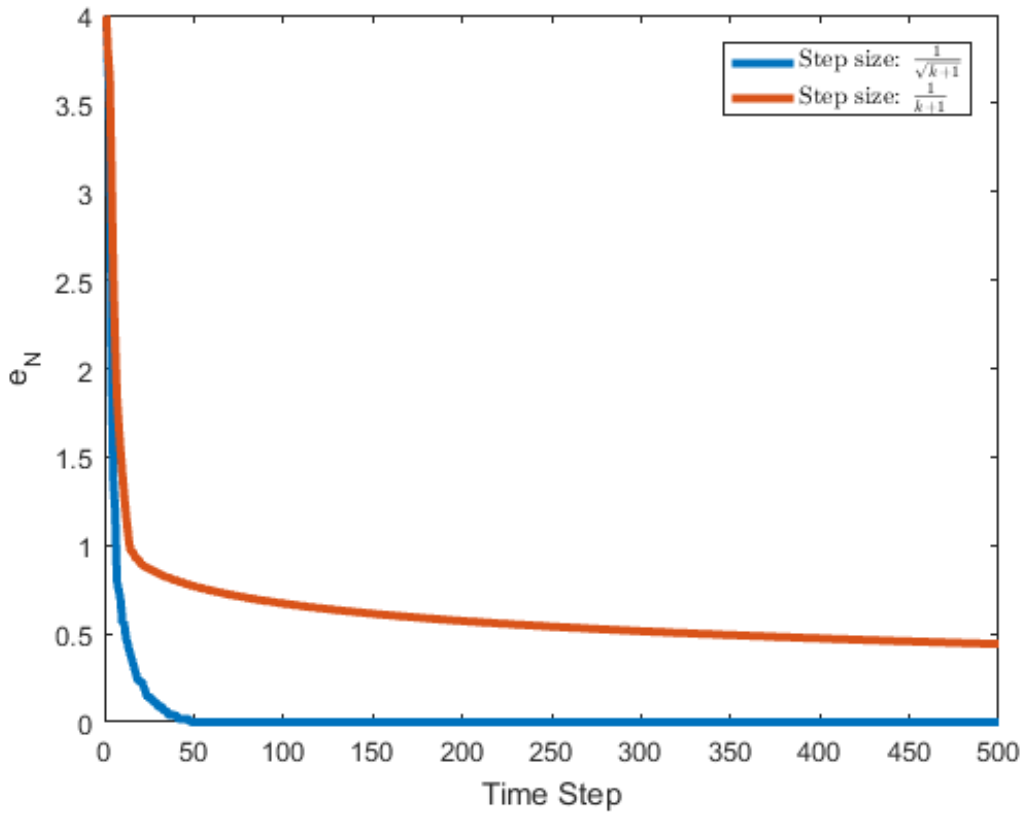

Figure 3.2: Comparison of convergence rates with different selections of step sizes

in the literature. In this case, we select the constraints $X_i = \mathbb{R}$ in accord with the analysis in Section 3.2.3. The comparison is shown in Fig. 3.2. From Fig. 3.2, we can see that e_N converges much faster when $\alpha(k) = \frac{1}{\sqrt{k+1}}$ than when $\alpha(k) = \frac{1}{k+1}$.

Chapter 4

Distributed Minimum Weighted Norm Solution to Linear Equations Associated with Weighted Inner Product

In this chapter, we will design distributed algorithms to find the solution of a system of linear equations $Ax = b$ with the minimum norms associated with inner product. The algorithm make the agents' estimates converge to the desired solution with appropriate approximation. What's more, the algorithm is robust to initialization errors, i.e., the final convergence error is upper bounded by initialization errors.

4.1 Problem Formulation

In this section, we find the minimum weighted norm solution of a group of linear equations in a distributed way, where the weighted norm is associated with a weighted inner product, and can be expressed as $\|x\|_M = \sqrt{x^T M x}$ for some symmetric and positive definite matrix M.

Suppose that we have a group of $\bar{m}$ agents, and agent i knows a group of linear equations $A_i x = b_i$, where $A_i \in \mathbb{R}^{m_i \times n}$, $x \in \mathbb{R}^n$, and $b_i \in \mathbb{R}^{m_i}$. Then the global linear equations that the agents plan to solve cooperatively are $Ax = b$, where $A = \begin{pmatrix} A_1 \\ \vdots \\ A_{\bar{m}} \end{pmatrix} \in \mathbb{R}^{m \times n}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_{\bar{m}} \end{pmatrix} \in \mathbb{R}^m$, and $m = \sum_{i=1}^{\bar{m}} m_i$. For the global linear equations $Ax = b$, we have the following assumption:

Assumption 9 *The system of linear equations $Ax = b$ has at least one solution, but 0 is not a solution of $Ax = b$, i.e., $b \neq 0$.*

The algorithm to find the minimum weighted norm solution of $Ax = b$ is proposed as

$$x_i(k+1) = x_i(k) - \frac{1}{|N_i(k)|} P_i^M [\sum_{j \in N_i(k)} (x_i(k) - x_j(k))],\ i \in V, \tag{4.1}$$

where x_i is agent i's estimate of the minimum weighted norm solution of $Ax = b$, $N_i(k)$ is the neighbor set of agent i at time k, $P_i^M = K_i(K_i^T M K_i)^{-1} K_i^T M$, and the columns of K_i form a basis of $\ker(A_i)$.

Remark 11 *When $M = I$, (4.1) is the same as the distributed algorithm to solve linear equations in [29]. So the algorithm in [29] is a special case of the algorithm here in (4.1). Also in this case, we denote P_i^I as P_i for simplicity, and notice that P_i is in fact the orthogonal projection matrix onto $ker(A_i)$, i.e., $span(K_i)$.*

Remark 12 *Let K be a matrix whose columns form a basis of $ker(A)$. Then we have that $span(K) \subseteq span(K_i)$ as $A_i K = 0$.*

4.2 Main Results

In this section, we will prove that with special initializations, the algorithm (4.1) can find the minimum weighted norm solution of the global linear equations $Ax = b$ in a distributed way. Moreover, if there are bounded errors on the initializations, we will prove that the final convergence point of (4.1) is also bounded away from the global minimum weighted norm solution.

Let's first define the local minimum weighted norm solution of $A_i x = b_i$ for agent i, $i \in V$ as

$$x_i^\star = \arg \min_{x: A_i x = b_i} \|x\|_M^2, \tag{4.2}$$

and the global minimum weighted norm solution of $Ax = b$ as

$$x^\star = \arg \min_{x: Ax = b} \|x\|_M^2. \tag{4.3}$$

4.2.1 Minimum two-norm case

In this part, we deal with the special case where $M = I$, i.e., the norm is two-norm.

Lemma 17 *When $M = I$, $x^\star$ in (4.3) is the unique solution of*

$$\begin{pmatrix} A \\ K^T \end{pmatrix} x = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{4.4}$$

where the columns of the matrix K form a basis of $ker(A)$.

Proof. The Lagrangian for the optimization problem in (4.3) with $W = I$ is

$$L = \frac{1}{2} x^T x + \lambda^T (Ax - b).$$

The minimum $x^\star$ satisfies

$$\frac{\partial L(x^\star, \lambda)}{\partial x} = x^\star + A^T \lambda = 0.$$

Then we have

$$K^T x^\star = -K^T A^T \lambda = -(AK)^T \lambda = 0.$$

Also as the columns of K form a basis of $ker(A)$, the columns of K and the rows of A are linearly independent. We have rank$\left(\begin{pmatrix} A \\ K^T \end{pmatrix}\right) = n$. It follows that (4.4) has a unique solution. As a result, $x^\star$ is the unique solution of (4.4). ■

Next we show the main result when $M = I$.

Theorem 3 *When $M = I$, if the agents start from $x_i^\star$, $i \in V$ in (4.2), i.e., the minimum two-norm solution of their local linear equations, then under Assumptions 1 and 9, they will converge to $x^\star$ in (4.3), i.e., the minimum two-norm solution of the global linear equations.*

Proof. From Lemma 4, we know that there exists a plane p_s in the affine space in which the points are the solutions of the global linear equations $Ax = b$, i.e.,

$$p_s = \{x | Ax = b\}. \tag{4.5}$$

Note that $\dim(p_s) = n - \text{rank}(A)$. We denote $p_i = \{x | A_i x = b_i\},\ i \in V$ the solution plane of local linear equations $A_i x = b_i$. Then we have $\dim(p_i) = n - \text{rank}(A_i)$. Define the plane

$$p_\perp = \{x | K^T x = 0\}, \tag{4.6}$$

where the columns of the matrix K form a basis of $\ker(A)$. Then we know that $\dim(p_\perp) = n - \dim(p_s) = \text{rank}(A)$.

We denote the line passing through the origin and $x^\star$ as $l_\star$. From Assumption 9 that $x^\star \neq 0$, $l_\star$ is unique. Also as the origin does not lie in p_s under Assumption 9, we know from Lemma 5 that the intersection of the line $l_\star$ and p_s is the single point $x^\star$; otherwise $l_\star$ would be in p_s.

From Lemma 17, we know $x^\star \in p_\perp$. And also $0 \in p_\perp$. Then with Lemma 5, we know that $l_\star \in p_\perp$. Because $p_\perp$ is perpendicular to $\text{span}(K)$, i.e., $\ker(A)$, $p_\perp$ is perpendicular to the space $p_{s0} = \{x | Ax = 0\}$. Also as p_{s0} is the associated vector space of p_s, then it follows from Definition 7, $p_\perp$ is perpendicular to p_s.

Let $p_{\perp i} = p_\perp \bigcap p_i$. Then we have that $x^\star \in p_{\perp i}$, that $p_{\perp i}$ is the solution plane of $\begin{pmatrix} A_i \\ K^T \end{pmatrix} x = \begin{pmatrix} b_i \\ 0 \end{pmatrix}$, and that $\dim(p_{\perp i}) = n - (n - \text{rank}(A) + \text{rank}(A_i)) = \dim(p_i) - \dim(p_s)$.

Let $l_{\perp i}$ be the line passing through the origin, intersected with and perpendicular to $p_{\perp i}$, and denote the intersection point of $l_{\perp i}$ and $p_{\perp i}$ as x_{s_i}, i.e. $l_{\perp i} \bigcap p_{\perp i} = x_{s_i}$ and $l_{\perp i} \perp p_{\perp i}$. As $0 \in p_\perp$ and $x_{s_i} \in p_{\perp i} \subset p_\perp$, we have that $l_{\perp i} \in p_\perp$ from Lemma 5. Then it follows that $l_{\perp i} \perp p_s$ as $p_\perp \perp p_s$. So $l_{\perp i}$ is perpendicular to the $\dim(p_s)$ basis of p_s. Since $l_{\perp i} \perp p_{\perp i}$, $l_{\perp i}$ is perpendicular to the $\dim(p_i) - \dim(p_s)$ basis of $p_{\perp i}$. Also the basis of p_s and that of $p_{\perp i}$ are linearly independent from the fact that they are orthogonal. Then it follows that $l_{\perp i}$ is perpendicular to $\dim(p_i)$ linearly independent vectors of p_i. So $l_{\perp i}$ is

perpendicular to p_i. Then from Lemma 6 we know that x_{s_i} is indeed $x_i^\star$. Also as $l_{\perp i} \in p_\perp$, we have that $x_i^\star \in p_\perp$. Thus we have known that $x^\star$ and $x_i^\star,\ i \in V$ lie in the plane $p_\perp$.

From the algorithm (4.1), we know that when $M = I$, $P_i = K_i(K_i^T K_i)^{-1}K_i^T,\ i \in V$ is an orthogonal projection matrix onto $\ker(A_i)$, i.e., $\text{span}(K_i)$. We denote $p_{v,i} = \{x | A_i x = 0\}$, which is the associated vector subspace of the plane p_i. Let $p_{v,\perp i} = p_{v,i} \bigcap p_\perp$. With a similar process to prove $l_{\perp i} \perp p_i$ in the above paragraph, we can show that for all $x \in p_\perp$, the line passing through x, intersected with and perpendicular to $p_{v,\perp i}$ is orthogonal to $p_{v,i}$. Then it follows that $P_i x \in p_{v,\perp i}$.

Note that $p_{v,\perp i}$ is indeed the associated vector space of $p_{\perp i}$. So if for all $j \in V$, $x_j(k)$ lie in $P_\perp$, it follows that $x_i(k+1)$ would be in the plane $x_i(k) + [p_{v,\perp i}]$. If $x_i(k)$ happens to be in $p_{\perp i}$, then the plane $x_i(k) + [p_{v,\perp i}]$ would be $p_{\perp i}$ itself. It follows that $x_i(k+1)$ is also in $p_{\perp i}$. As agent i initializes itself at $x_i^\star \in p_{\perp i} \subset p_\perp$, then from induction we have that $x_i(k),\ k = 0, 1, 2, \cdots$ lies in $p_{\perp i}$, and thus in $p_\perp$. Then it follows that the sequences generated with the algorithm (4.1) when $M = I$ will be all in $p_\perp$.

We also know that when $M = I$, the algorithm (4.1) reduces to the algorithm proposed in [29], and thus the agents reach a consensus at some solution of $Ax = b$. Also the solution has to be in $p_\perp$, which is $\{x | K^T x = 0\}$. Thus, it satisfies both $K^T x = 0$ and $Ax = b$. Then from Lemma 17, we know that the solution is $x^\star$. ■

Next, we will show that if the agents are initialized bounded away from $x_i^\star$, they will converge to some solution of $Ax = b$, and the solution is also bounded away from $x^\star$. First, we have the following lemma on the distance between two parallel planes with the same associated vector space.

Lemma 18 *For two parallel planes $p_1 = \{x|Ax = b\}$ and $p_2 = \{x|Ax = b + \delta b\}$, where A has full row rank, the distance from p_1 to p_2 is given by $d(p_1, p_2) = \|\delta b\|_{(AA^T)^{-1}}$.*

Proof. The problem in Lemma 18 is equivalent to

$$\begin{aligned} \min \quad & \frac{1}{2}\|x - y\|_2^2, \\ \text{subject to} \quad & Ax = b, \\ & Ay = b + \delta b. \end{aligned}$$

Then its Lagrangian is

$$L = \frac{1}{2}\|x - y\|_2^2 + \lambda_1^T(Ax - b) + \lambda_2^T(Ay - (b + \delta b)).$$

Then we have

$$\begin{aligned} \frac{\partial L}{\partial x} &= x - y + A^T\lambda_1 = 0, \\ \frac{\partial L}{\partial y} &= y - x + A^T\lambda_2 = 0. \end{aligned}$$

Together with $Ax = b$ and $Ay = b + \delta b$, we have $\lambda_1 = -(AA^T)^{-1}\delta b$ and $x - y = A^T(AA^T)^{-1}\delta b$. It follows that $\|x - y\|_2^2 = \delta b^T(AA^T)^{-1}\delta b$ and

$$d(p_1, p_2) = \|x - y\|_2 = \|\delta b\|_{(AA^T)^{-1}}.$$

■

Then we have the following result for the case with inexact initializations:

Theorem 4 *If the agents have their initial estimates at $\tilde{x}_i^\star,\ i \in V$, which are locally feasible, i.e., $A\tilde{x}_i^\star = b_i$, but inexact, i.e. $\tilde{x}_i^\star \neq x_i^\star$, then under Assumptions 1 and 9, they will converge to some solution $\tilde{x}^\star$ of $Ax = b$ located in a neighborhood of $x^\star$, and $d(\tilde{x}^\star, x^\star) \leq \max_{i \in V} d(\tilde{x}_i^\star, x_i^\star)$.*

Proof. Let's first prove that $K^T P_i = K^T,\ i \in V$. As $P_i = K_i(K_i^T K_i)^{-1} K_i^T$ is the orthogonal projection matrix onto span(K_i), i.e., for all $x \in \mathbb{R}^n$, $x - P_i x$ is perpendicular to span(K_i). As span(K_i) contains span(K) from Remark 12, $x - P_i x$ is also perpendicular to span(K). Then we have $K^T(x - P_i x) = 0$. It follows that for all $x \in \mathbb{R}^n$, we have $K^T x = K^T P_i x$, and thus

$$K^T = K^T P_i.$$

When $M = I$ in (4.1), it follows from [29] that $x_i(k),\ i \in V$ reach a consensus at some solution $\tilde{x}^\star$ of $Ax = b$. As $K^T = K^T P_i$, we have from (4.1) that

$$\begin{aligned} K^T x_i(k+1) &= K^T x_i(k) - \frac{1}{|N_i(k)|} K^T P_i [\sum_{j \in N_i(k)} (x_i(k) - x_j(k))] \\ &= K^T x_i(k) - \frac{1}{|N_i(k)|} \sum_{j \in N_i(k)} (K^T x_i(k) - K^T x_j(k)), \end{aligned}$$

which is a consensus algorithm for $K^T x$. So $K^T \tilde{x}^\star$ is in the convex hull of $K^T \tilde{x}_i^\star$. From the convexity of norms, we have that $\|K^T \tilde{x}^\star\|_{(K^T K)^{-1}} \leq \max_{i \in V} \|K^T \tilde{x}_i^\star\|_{(K^T K)^{-1}}$. As p_s in (4.5) and $p_\perp$ in (4.6) are orthogonal and $x^\star$ and $\tilde{x}^\star$ lie in p_s, we have that $\tilde{x}^\star - x^\star$ is perpendicular to $p_\perp$ and thus $d(\tilde{x}^\star, p_\perp) = d(\tilde{x}^\star, x^\star)$. As $d(\tilde{x}_i^\star, x_i^\star) \geq d(\tilde{x}_i^\star, p_\perp)$ and $d(\tilde{x}_i^\star, p_\perp) = \|K^T \tilde{x}_i^\star\|_{(K^T K)^{-1}}$ from Lemma 18, we have

$$d(\tilde{x}^\star, x^\star) = d(\tilde{x}^\star, p_\perp) = \|K^T \tilde{x}^\star\|_{(K^T K)^{-1}} \leq \max_{i \in V} \|K^T \tilde{x}_i^\star\|_{(K^T K)^{-1}} \leq \max_{i \in V} d(\tilde{x}_i^\star, x_i^\star).$$

■

Remark 13 *When Assumption 9 does not hold and* 0 *is among the solutions of* $Ax = b$*, both* $x_i^\star$ *in (4.2) and* $x^\star$ *in (4.3) are* 0*. So the results in Theorem 3 and Theorem 4 still hold.*

4.2.2 Minimum weighted norm case

In this part, we will consider the case when M is a general symmetric and positive definite matrix M.

Theorem 5 *When W is symmetric and positive definite, if the agents choose $x_i^\star$, $i \in V$ in (4.2) as their initial values, then the sequences they generated with the algorithm in (4.1) will finally converge to $x^\star$ in (4.3) under Assumptions 1 and 9.*

Proof. As W is symmetric and positive definite, there exists an invertible matrix C, such that $C^TC = W$. Let

$$y = Cx. \tag{4.7}$$

The problems in (4.2) and (4.3) become, respectively

$$y_i^\star = \arg \min_{x:A_{i,W}y=b_i} \|y\|_2^2,$$

and

$$y^\star = \arg \min_{x:A_My=b} \|y\|_2^2,$$

where $A_{i,W} = A_iC^{-1}$ and $A_M = AC^{-1}$. We denote $K_{i,W} = CK_i$ and $K_M = CK$ as the matrices whose columns form bases of $\ker(A_{i,W})$ and $\ker(A_M)$, respectively. We have that $x_i^\star = C^{-1}y_i^\star$ and $x^\star = C^{-1}y^\star$.

Then it follows that the algorithm (4.1) for x becomes the following algorithm for y:

$$C^{-1}y_i(k+1) = C^{-1}y_i(k) - \frac{1}{|N_i(k)|}P_i^M[\sum_{j\in N_i(k)} C^{-1}(y_i(k) - y_j(k))].$$

Then

$$\begin{aligned}
&y_i(k+1)\\
=&y_i(k)-\frac{1}{|N_i(k)|}CP_i^M[\sum_{j\in N_i(k)}C^{-1}(y_i(k)-y_j(k))]\\
=&y_i(k)-\frac{1}{|N_i(k)|}CK_i(K_i^TMK_i)^{-1}K_i^TM[\sum_{j\in N_i(k)}C^{-1}(y_i(k)-y_j(k))]\\
=&y_i(k)-\frac{1}{|N_i(k)|}K_{i,W}(K_{i,W}^TK_{i,W})^{-1}K_{i,W}^T[\sum_{j\in N_i(k)}(y_i(k)-y_j(k))],
\end{aligned} \tag{4.8}$$

which is the algorithm (4.1) for y in the case of $M=I$. Then from Theorem 3, we know that if the agents begin with $y_i^\star$ and travel along the algorithm (4.8), they will finally reach $y^\star$. Accordingly, if they start from $x_i^\star$ and travel along (4.1), they will arrive at $x^\star$. ■

Remark 14 *When $f(x)$ is an increasing function of x, let*

$$x_f^\star=\arg\min_{x:Ax=b}f(\|x\|_M),$$

and

$$x_{f,i}^\star=\arg\min_{x:A_ix=b_i}f(\|x\|_M).$$

Then the algorithm in (4.1) can also be used to find $x_f^\star$ when the agents are initialized at $x_{f,i}^\star$, because $x_f^\star=x^\star$ and $x_{f,i}^\star=x_i^\star$.

When the agents have locally feasible but inexact initial values, we have the following result for a general symmetric and positive definite M:

Theorem 6 *If the agents have initial estimates $\tilde{x}_i^\star, i\in V$, which are locally feasible, i.e., $A\tilde{x}_i^\star=b_i$, but inexact, i.e. $\tilde{x}_i^\star\neq x_i^\star$, then under Assumptions 1 and 9, they will converge to some solution $\tilde{x}^\star$ of $Ax=b$ in a neighborhood of $x^\star$ and $\|\tilde{x}^\star-x^\star\|_M\leq\max_{i\in V}\|\tilde{x}_i^\star-x_i^\star\|_M$.*

Proof. From Theorems 4 and 5, we know that $y_i, i \in V$ in (4.8) finally converge to some solution $\tilde{y}^\star$ of $A_M y = b$ and $d(\tilde{y}^\star, y^\star) \leq \max_{i \in V} d(\tilde{y}_i^\star, y_i^\star)$. Then It follows that accordingly $x_i, i \in V$ will converge to some solution $\tilde{x}^\star = C^{-1}\tilde{y}^\star$ of $Ax = b$ and

$$\|\tilde{x}^\star - x^\star\|_M \leq \max_{i \in V} \|\tilde{x}_i^\star - x_i^\star\|_M.$$

■

4.3 Simulations

In this section, we provide two simulation examples to show the effectiveness of the proposed algorithm (4.1) in the cases of both the exact initializations and inexact initializations.

In both examples, to obtain the positive definite matrix M, we first generate a random orthogonal matrix U and a random diagonal matrix W_d with positive diagonal elements, and then make $W = U^T W U$. We consider an agent network consisting of 15 agents under the topology of a directed ring, and each agent knows one row of the augmented matrix $[A\ b]$, where $A \in \mathbb{R}^{15\times 20}$ and $b \in \mathbb{R}^{15}$ are also generated randomly. The topology of the directed ring.

4.3.1 Exact initialization

In this part, we show with a simulation example that the algorithm in (4.1) makes all agents converge to the global minimum weighted norm solution if they are initialized at their local minimum weighted norm solutions. The parameters of the example are generated in a random way as stated above.

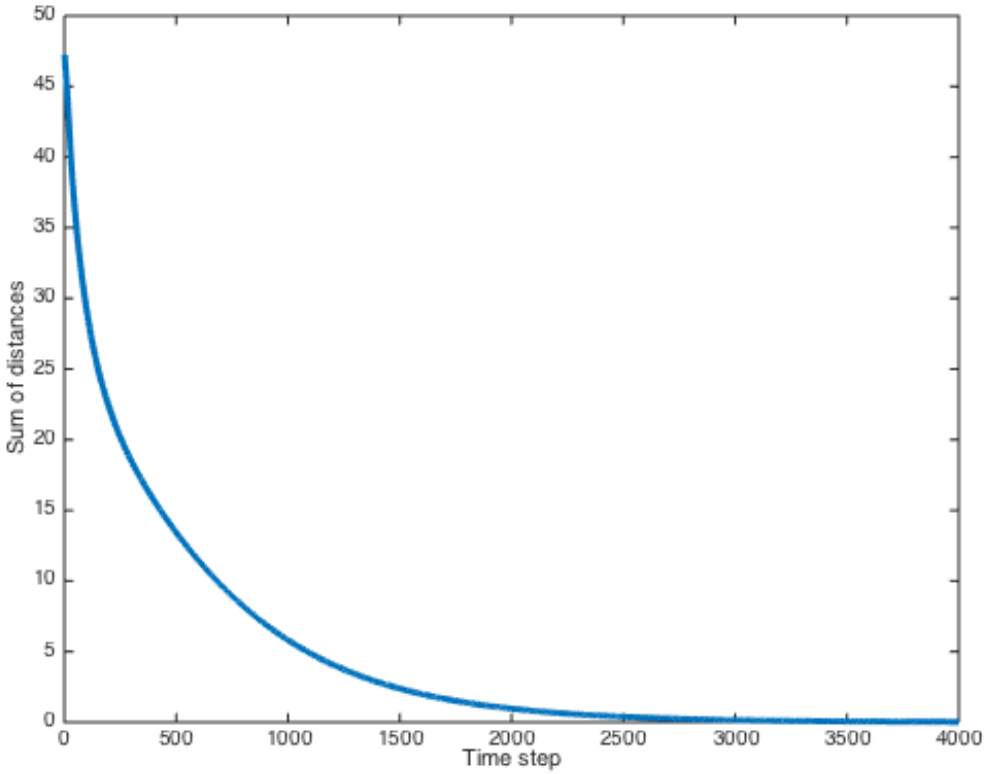

Figure 4.1: Change of distance from agents estimates and the global minimum weighted norm solution

The simulation result is shown in Fig. 4.1, from which we can see that the estimation errors of all agents on the minimum weighted norm solution of the global linear equations gradually converge to 0. So we can tell that all agents arrive at the minimum weighted norm solution of $Ax = b$ at last.

4.3.2 Inexact initialization

In this part, we use a simulation example to illustrate the effectiveness of the algorithm (4.1) in the case of inexact initializations. The parameters W, A and b are

obtained as mentioned in the second paragraph in this section, and the initialization errors are first generated randomly and then multiplied by K_i to make sure the initial values are locally feasible. As the parameters are obtained in a random way, they might be different from those in the previous case.

The simulation results are presented in Figs. 4.2, 4.3 and 4.4. Fig. 4.2 shows the change of the maximum distance from the other agents to agent 1, from which we can see that as the maximum distance vanishes, the agents finally come to the same point. Fig. 4.3 shows that difference $Ax_1 - b$, which indicates that the estimate of agent 1 finally becomes a solution of the global linear equations $Ax = b$ because the difference tends to 0. Fig. 4.4 shows the change of the average of the norm of all agents estimation errors associated with W. From it we can know that the agents estimates does not converge to the minimum weighted norm solution of the global linear equations because of the initial errors, but they are bounded away from it. We can also know that the final error is bounded by the maximum of the initial errors. So from all the three figures, we have that with inexact initializations, different agents will finally converge to the same solution of $Ax - b$, and the solution is not exactly the global minimum weighted norm solution but the final error is bounded by the maximum of the initial errors.

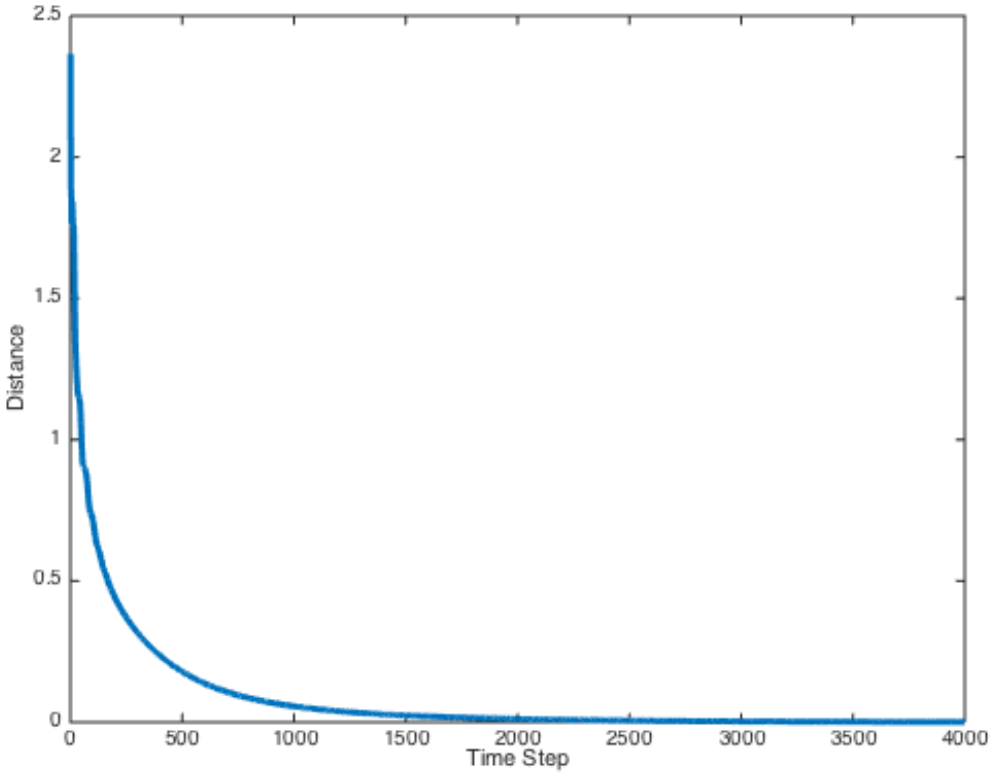

Figure 4.2: Change of maximum distance from other agents to agent 1

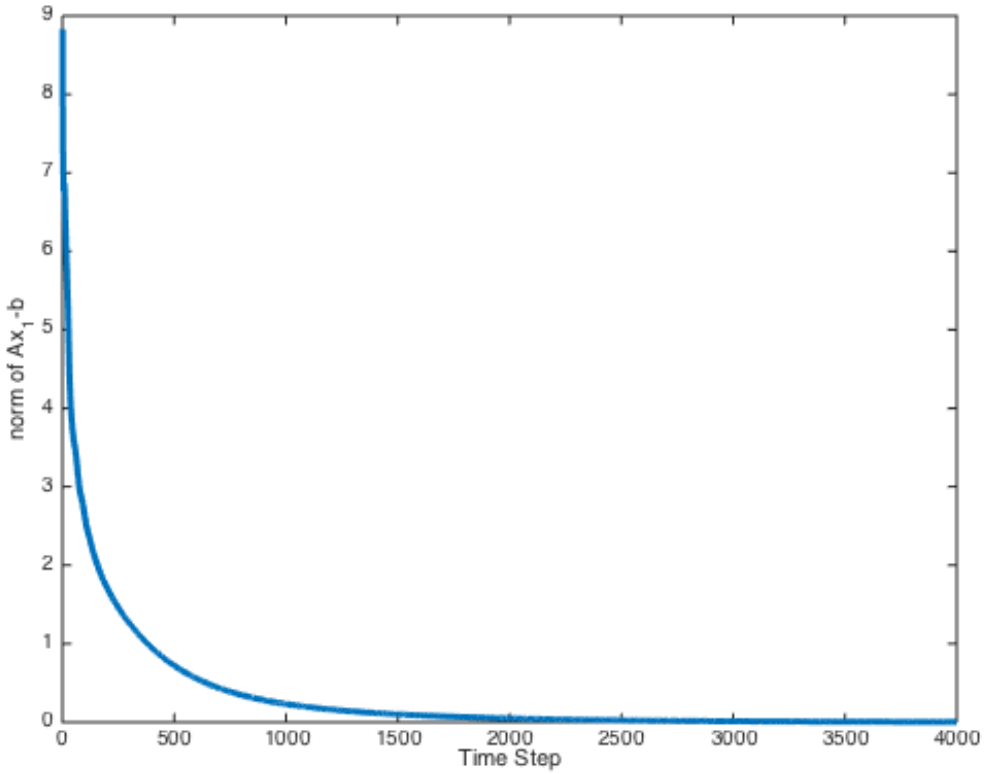

Figure 4.3: Change of norm of $Ax_1 - b$

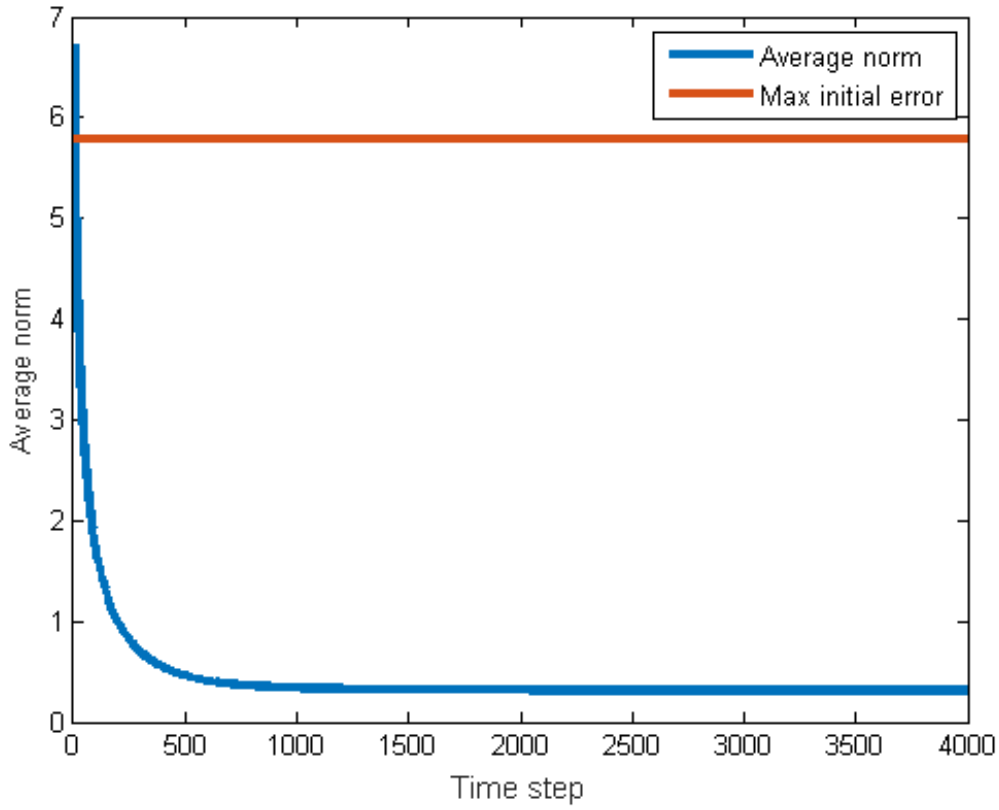

Figure 4.4: Change of the average of norms of difference agents estimation errors

Chapter 5

Distributed Algorithm to Solve a System of Linear Equations with Unique or Multiple Solutions from Arbitrary Initializations

In this chapter, we will design a distributed algorithm to solve a system of linear equations $Ax = b$ that allows arbitrary initialization and at the same time is effective when $Ax = b$ has multiple solutions.

5.1 Problem Formulation

Suppose that we have a group of $\bar{m}$ agents, and each agent knows a local system of linear equations, $A_i x = b_i,\ i = 1, \cdots, \bar{m}$, where $A_i \in \mathbb{R}^{m_i \times n}, x \in \mathbb{R}^n, b_i \in \mathbb{R}^{m_i}$, and $\sum_{i=1}^{\bar{m}} m_i = m$. The goal of all the agents is to find a common solution of all their local systems of linear equations, or in another way, to solve in a cooperative way the global system of linear equations $Ax = b$, where $A = \begin{pmatrix} A_1^T \\ \vdots \\ A_{\bar{m}}^T \end{pmatrix}^T \in \mathbb{R}^{m \times n}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_{\bar{m}} \end{pmatrix}^T \in \mathbb{R}^m$. The agents communicate with their neighbors under the topologies in Assumption 1.

We will design a distributed algorithm for the agents to achieve their goal. The designed algorithm will be proved to find a solution of $Ax = b$ with either unique or multiple solutions from arbitrary initializations at a geometric rate.

5.2 Main Results

In this section, we design a discrete-time distributed algorithm to solve a system of linear equations $Ax = b$. The algorithm allows arbitrary initializations and can converge at a geometric rate when $Ax = b$ has either unique or multiple solutions. When $Ax = b$ has a unique solution, the agents' estimates in the designed algorithm reach a consensus at the solution of $Ax = b$. When the system of linear equations $Ax = b$ has multiple solutions, the agents' estimates are proved to converge to the solution of $Ax = b$ determined by the initializations, the communication topologies, and the minimum 2-norm solution of $Ax = b$.

5.2.1 Distributed Algorithm to Solve Linear Equations

In this part, we describe the proposed distributed algorithm to solve a system of linear equations $Ax = b$. The algorithm we propose is as follows:

$$\begin{aligned} v_i(k)&=\sum_{j=1}^{\bar{m}} w_{ij}(k)x_j(k), \\ x_i(k+1)&=\begin{cases} v_i(k)-\frac{\gamma_i\|A_iv_i(k)-b_i\|^2A_i^T(A_iv_i(k)-b_i)}{\|A_i^T(A_iv_i(k)-b_i)\|^2}, A_iv_i(k)\neq b_i, \\ v_i(k), \qquad\qquad\qquad\qquad \text{otherwise}, \end{cases} \end{aligned} \tag{5.1}$$

where $x_i(k)$ is agent i's estimate of the solution of $Ax = b$ at time k, and γ_i is a constant in the open interval $(0, 1)$. The algorithm consists of two parts. The first part is a consensus process driving all agents to a common point, while the second part is trying to solve the local system of linear equations $A_ix = b_i$.

Remark 15 *We can recast solving $Ax = b$ as a distributed optimization problem as $\min_x \sum_{i=1}^{\bar{m}} \|A_ix-b_i\|_2^2$ and use generic distributed optimization algorithms to solve it. But the generic distributed optimization algorithms are usually of sub-linear convergence. At the same time, there are some generic distributed optimization algorithms that can converge at a linear rate, but they may have special requirements. For example, [45] requires a fixed and symmetric weight matrix associated with an undirected communication graph while [41] and [35] require more communication efforts, e.g., communicating not only the estimates but also the gradients. What's more, when $Ax = b$ has multiple solutions and thus $\sum_{i=1}^{\bar{m}} \|A_ix - b_i\|^2$ would not be strongly convex, the generic distributed algorithms fail to converge at a linear rate. In summary, we may recast the problem as a distributed optimization problem, but the generic distributed optimization algorithms may not work as well as the algorithm in (5.1).*

Remark 16 *There might be some similarity in structure in terms of a combination of consensus and local gradient direction to some generic distributed optimization algorithms, e.g. [17]. But we make use of the local function values, i.e.,* $\|A_i v_i(k) - b_i\|_2^2$*, and the square of norms of the gradient, i.e.,* $\|A_i^T(A_i v_i(k)) - b_i\|_2^2$ *in the local gradient direction in the algorithm (5.1), which is different from the generic algorithms in the literature including [17].*

Remark 17 *Although the algorithm in [49] may be shown to work for the case when* $Ax = b$ *has multiple solutions with additional efforts, the selection of the parameter matrices* G_i *in [49] is not direct and may require additional computation efforts, while the choice of the parameters* γ_i *in (5.1) is direct. Also, the analysis on the convergence of (5.1) indeed provides a simpler way to analyze the algorithm in [49]. Finally, we show where the convergence point is for (5.1) when* $Ax = b$ *has multiple solutions in Proposition 2, which may be useful to find a solution of* $Ax = b$ *with special properties, e.g., the solution with minimum 2-norm.*

The convergence of (5.1) is stated in the following theorem:

Theorem 7 *The agents' estimates* $x_i(k)$, $\forall i \in V$ *in (5.1) converge to a common solution of* $Ax = b$ *at a geometric rate from arbitrary initializations under Assumptions 1 and 2. When* $Ax = b$ *has multiple solutions, the limit point of* $x_i(k)$, $\forall i \in V$ *can be expressed as* $x_{MN} + x_{NA}$*, where* x_{MN} *is the solution of* $Ax = b$ *with the minimum* 2*-norm and* x_{NA} *is the consensus point of the projections of the agents' initial conditions under the communication topologies* $\mathcal{G}_k, k = 0, 1, 2, \cdots$*.*

Remark 18 *Theorem 7 ensures the geometric convergence rate of (5.1), but it does not give any quantitative upper bound on the convergence rate. It is generally difficult to obtain such a bound, but in Sections 5.3.1 and 5.3.2, we provide an analysis on such bounds for two special cases when A is orthogonal or the communication topology is complete.*

We next prove Theorem 7. We prove the case when $Ax = b$ has a unique solution in Section 5.2.2, and that when $Ax = b$ has multiple solutions in Section 5.2.3.

But first, we show the continuity and the M-Fejer property used in (5.1), which paves the way for the proof of Theorem 7. Define

$$T_i x = \begin{cases} x - \gamma_i [\frac{\|A_i x - b_i\|_2^2}{\|A_i^T (A_i x - b_i)\|_2^2} A_i^T (A_i x - b_i)], & A_i x \neq b_i, \\ x, & \text{otherwise} \end{cases},$$

where γ_i is a constant in the open interval $(0, 1)$. Then we have that $x_i(k+1) = T_i v_i(k)$ in (5.1).

We then show that T_i is continuous. Let $x_i^\star$ be a solution of the local system of linear equations $A_i x_i^\star = b_i$ and $e = x - x_i^\star$. Then when $A_i x \neq b_i$, i.e. $A_i e \neq 0$,

$$T_i x = x - \gamma_i \|A_i e\|_2^2 A_i^T A_i e / (\|A_i^T A_i e\|_2^2).$$

As $\sigma_{\min}(A_i)\|A_i e\|_2 \leq \|A_i^T A_i e\|_2 \leq \sigma_{\max}(A_i)\|A_i e\|_2$ when $A_i e \neq 0$, we have that $\lim_{A_i e \to 0, A_i e \neq 0} T_i x = x$. So we get that T_i is continuous. In the rest of this chapter, we use

$$T_i x = x - \gamma_i [\frac{\|A_i x - b_i\|_2^2}{\|A_i^T (A_i x - b_i)\|_2^2} A_i^T (A_i x - b_i)] \tag{5.2}$$

to represent (5.1) for simplicity and regard the case in which $A_i x = b_i$ as the limit of (5.2).

We then show the M-Fejer property of (5.2) with the following lemma:

Lemma 19 *[46] Let f be a convex function, and $g(x)$ be its subgradient at the point x. Also, let $f^\star = \inf_x f(x)$, $c \geq f^\star$, and $M(c) = \{x : f(x) \leq c\}$. Define*

$$T(x) = x - \frac{\gamma_f[f(x) - c]}{\|g(x)\|_2^2} g(x), 0 < \gamma_f < 2. \tag{5.3}$$

Then T in (5.3) is an $M(c)$-Fejer mapping.

If we let $f(x) = \frac{1}{2}\|A_i x - b_i\|_2^2$, $c = 0$, and note that $\gamma_i = \frac{\gamma_f}{2}$, we can see that (5.2) is the special case of (5.3). So T_i in (5.2) is a continuous M-Fejer mapping, with M being the set of solutions of $A_i x = b_i$.

The above discussions are summarized in the following lemma:

Lemma 20 *T_i in (5.2) is a continuous M-Fejer mapping with $M = \{x : A_i x = b_i\}$.*

Remark 19 *The distributed algorithm in (5.1) is closely related to a subgradient algorithm in [46], which is a Fejer-type algorithm. But the subgradient algorithm of Fejer-type in [46] cannot in general be applied to distributed optimization problems because of the lack of the knowledge of the optimal value of the global objective function.*

5.2.2 Unique Solution Case

In this part, we prove the convergence at a geometric rate of (5.1) when $Ax = b$ has a unique solution.

Denote the solution of $Ax = b$ as $x^\star$. Let $e_i(k) = x_i(k) - x^\star$ and $e_{v_i}(k) = v_i(k) - x^\star$. The convergence of $x_i(k)$ to $x^\star$ is then equivalent to the convergence of $e_i(k)$ to 0. So we focus on proving that $e_i(k)$ vanishes at a geometric rate in this part.

Notice that $A_i x^\star = b_i$. Then we can get from (5.1) that

$$\begin{aligned} e_{v_i}(k) &= \sum_{j=1}^{\bar{m}} w_{ij}(k) e_j(k), \\ e_i(k+1) &= e_{v_i}(k) - \gamma_i \frac{\|A_i e_{v_i(k)}\|_2^2}{\|A_i^T A_i e_{v_i}(k)\|_2^2} A_i^T A_i e_{v_i}(k), \end{aligned} \tag{5.4}$$

Notice that (5.4) is indeed the algorithm (5.1) when $b = 0$, and in this case,

$$\begin{aligned} T_i x &= x - \gamma_i \frac{\|A_i x\|_2^2 A_i^T A_i x}{\|A_i^T A_i x\|_2^2} \\ &= (I - \gamma_i \frac{\|A_i x\|_2^2 A_i^T A_i}{\|A_i^T A_i x\|_2^2}) x, \end{aligned} \tag{5.5}$$

which is also an M_i-Fejer mapping, with $M_i = \{x : A_i x = 0\}$.

We have that T_i in (5.5) is homogeneous, i.e., $T_i(\lambda x) = \lambda T_i x,\ \forall \lambda \in \mathbb{R}$, because

$$\begin{aligned} T_i(\lambda x) &= (I - \gamma_i \frac{\|A_i \lambda x\|_2^2 A_i^T A_i}{\|A_i^T A_i \lambda x\|_2^2}) \lambda x \\ &= \lambda (I - \gamma_i \frac{\|A_i x\|_2^2 A_i^T A_i}{\|A_i^T A_i x\|_2^2}) x \\ &= \lambda T_i x \end{aligned}$$

when $\lambda \neq 0$, and when $\lambda = 0$, $T_i(\lambda x) = \lambda T_i x = 0$. Also, we have that if a mapping T is homogeneous, $T0 = 0$. With the homogeneity, we can define the induced norm for T_i:

Definition 10 *For a homogeneous mapping T, its induced norm is defined as* $\|T\| = \sup_{\|x\| \neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{\|x\|=1} \|Tx\|$.

We have that $\|Tx\| \leq \|T\|\|x\|$ and $\|T_2 T_1\| \leq \|T_2\|\|T_1\|$. We then define the mixed norm of a vector of such mappings. Let $\boldsymbol{T}$ as in (2.4), and $\boldsymbol{T}$ is homogeneous if $T_i,\ i = 1, 2, \cdots, \bar{m}$

are homogeneous. We then define the mixed norm of $\boldsymbol{T}$ as

$$\|\boldsymbol{T}\|_{2,\infty} = \| \begin{pmatrix} \|T_1\|_2 \\ \vdots \\ \|T_{\bar{m}}\|_2 \end{pmatrix} \|_\infty = \max_{i=1,2,\cdots,\bar{m}} \|T_i\|_2.$$

We find that the mixed norms are induced norms defined in Definition 10, as shown in the following lemma.

Lemma 21 *Let $\boldsymbol{T}$ be defined as in (2.4), and $T_i,\ i = 1, 2, \cdots, \bar{m}$ are homogeneous mappings. Then $\|\boldsymbol{T}\|_{2,\infty} = \sup_{\|x\|_{2,\infty}=1} \|\boldsymbol{T}x\|_{2,\infty}$.*

Proof. We have that

$$\sup_{\|x\|_{2,\infty}=1} \|\boldsymbol{T}x\|_{2,\infty} = \sup_{\|x\|_{2,\infty}=1} \max_{i=1,2,\cdots,\bar{m}} \|T_i x_i\|_2$$

$$= \max_{i=1,2,\cdots,\bar{m}} \sup_{\|x\|_{2,\infty}=1} \|T_i x_i\|_2 = \max_{i=1,2,\cdots,\bar{m}} \sup_{\|x_i\|_2=1} \|T_i x_i\|_2$$

$$= \max_{i=1,2,\cdots,\bar{m}} \|T_i\|_2 = \|\boldsymbol{T}\|_{2,\infty}.$$

So $\|\boldsymbol{T}\|_{2,\infty} = \sup_{\|x\|_{2,\infty}=1} \|\boldsymbol{T}x\|_{2,\infty}$. ■

Remark 20 *From Lemma 21, we can see that $\|\cdot\|_{2,\infty}$ is sub-multiplicative, i.e., $\|\boldsymbol{T}x\|_{2,\infty} \leq \|\boldsymbol{T}\|_{2,\infty}\|x\|_{2,\infty}$ and $\|\boldsymbol{T}_1\boldsymbol{T}_2\|_{2,\infty} \leq \|\boldsymbol{T}_1\|_{2,\infty}\|\boldsymbol{T}_2\|_{2,\infty}$.*

With the above discussions, we find that a homogeneous and continuous M-Fejer mapping is contractive when M has a special structure, as stated in the following lemma.

Lemma 22 *If T is a homogeneous and continuous M-Fejer mapping and its fixed point set M is a linear subspace, then there exists a constant $c < 1$ such that $\|Tx\| \leq c\|x\|,\ \forall x \in M^\perp$.*

Proof. As M is a linear subspace, $0 \in M$. Then we have that $\|Tx\| \leq \|x\|$. So $c \leq 1$.

Next, we prove $c < 1$.

When $x = 0$, it is obvious that $\|Tx\| \leq c\|x\|, \forall c < 1$ when T is homogeneous.

Next we prove by contradiction that $c < 1$ when $x \neq 0$. Suppose not. There exist sequences $\{c_k\}$ and $\{x_k\}$, such that $c_k \leq 1, c_k \to 1$, $x_k \in M^{\perp}, x_k \neq 0$, $k = 1, 2, 3, \cdots$ and $\|Tx_k\| \geq c_k\|x_k\|$. Let $y_k = \frac{x_k}{\|x_k\|}$. Then $\|y_k\| = 1$. From the homogeneity of T, we have that $\|Ty_k\| \geq c_k$. As the sequence $\{y_k\}$ is bounded, it has a convergent subsequence. Without loss of generality, we suppose $\{y_k\}$ itself is convergent to y_∞. Then we have that $\|y_\infty\| = 1$. As $M^{\perp}$ is closed, $y_\infty \in M^{\perp}$. From the continuity of T and norms, we have that $\|Ty_\infty\| \geq 1$. As $\|Ty_\infty\| \leq \|y_\infty\|$ from $0 \in M$, $\|Ty_\infty\| = 1 = \|y_\infty\|$. From Lemma 7, we have that $Ty_\infty = y_\infty$. This contradicts with the fact that $y_\infty \in M^{\perp}$ and $\|y_\infty\| = 1$. So $c < 1$.

As a result, there exists $c < 1$ such that $\|Tx\| \leq c\|x\|$. ■

From (5.4), we get that $e_{v_i}(k+1) = \sum_{j=1}^{\bar{m}} w_{ij}(k+1) T_j e_{v_j}(k)$. Denote $\boldsymbol{e}_v(k) = \begin{pmatrix} e_{v_1}(k) \\ \cdots \\ e_{v_{\bar{m}}}(k) \end{pmatrix}$ and $\boldsymbol{T}$ as in (2.4), then we have that $\boldsymbol{e}_v(k+1) = (W(k+1) \otimes I)\boldsymbol{T}\boldsymbol{e}_v(k)$, and that

$$\boldsymbol{e}_v(t) = (W(t) \otimes I)\boldsymbol{T}(W(t-1) \otimes I)\boldsymbol{T} \cdots (W(s+1) \otimes I)\boldsymbol{T}\boldsymbol{e}_v(s),$$

for $t > s$. Define

$$\Phi_T(t:s) = (W(t) \otimes I)\boldsymbol{T}(W(t-1) \otimes I)\boldsymbol{T} \cdots (W(s) \otimes I)\boldsymbol{T}. \tag{5.6}$$

Then we have that $\boldsymbol{e}_v(t) = \Phi_T(t : s+1)\boldsymbol{e}_v(s)$.

Remark 21 *$\Phi_T(t:s)$ can be written as* $\begin{pmatrix} \Phi_{T,1}(t:s) \\ \Phi_{T,2}(t:s) \\ \cdots \\ \Phi_{T,\bar{m}}(t:s) \end{pmatrix}$, *where* $\Phi_{T,i}(t:s) = \sum_{j=1}^{\bar{m}} w_{ij}(t)T_j(\Phi_{T,j}(t-1:s))$ *and* $\Phi_{T,i}(s:s) = \sum_{j=1}^{\bar{m}} w_{ij}(s)T_j$. *We can then prove by mathematical induction that $\Phi_{T,i}(t:s)$ is homogeneous if each T_i is homogeneous, and thus that $\Phi_T(t:s)$ is also homogeneous. Hence Lemmas 21 and 22 apply to $\Phi_T(t:s)$.*

With the above results, we next prove Theorem 7 when $Ax = b$ has a unique solution, which is stated in the following proposition:

Proposition 1 *Under Assumptions 1 and 2, the agents' estimates $\{x_i(k)\}, i = 1, 2, \cdots, \bar{m}$ in the distributed algorithm (5.1) converge to the solution of $Ax = b$ at a geometric rate when $Ax = b$ has a unique solution.*

Proof. Under Assumptions 1 and 2, the product of the weight matrices associated with the communication topologies $\Phi(k+(\bar{m}+2)B:k)$ is positive from Lemma 1. From Lemma 8, $\Phi_T(k+(\bar{m}+2)B:k)$ is an M-Fejer mapping with $M = \{\mathbf{1}_{\bar{m}} \otimes x : x \in \bigcap_{i=1}^{\bar{m}} \{x : A_i x = 0\}\}$. As $Ax = b$ has a unique solution, $Ax = 0$ has the unique solution 0, and thus $M = \{0\}$. Then from Remark 21 and Lemma 22, $\|\Phi_T(k + (\bar{m} + 2)B : k)\|_{2,\infty} \leq c_W(k) < 1$. Here $c_W(k)$ is the upper bound defined in Lemma 22. $c_W(k)$ is indeed determined by the entries in $W(k), W(k+1), \cdots, W(k + (\bar{m} + 2)B)$, so $c_W(k)$ can be regarded as a function of $w_{ij}(t), t = k, k+1, \cdots, k+(\bar{m}+2)B$. As $T_i, i = 1, 2, \cdots, \bar{m}$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ are continuous, $c_W(k)$ is also continuous on $w_{ij}(t), t = k, k+1, \cdots, k+(\bar{m}+2)B$. As $w_{ij}(t), t = k, k+1, \cdots, k+(\bar{m}+2)B$ form a closed bounded set in $\mathbb{R}^{(\bar{m}+2)Bn^2}$ and $c_W(k) < 1$ for all $w_{ij}(t)$,

we have that there exists a constant $c < 1$ such that $c_W(k) \leq c < 1$. That is to say $\|\Phi_T(k + (\bar{m}+2)B : k)\|_{2,\infty} \leq c < 1, k = 1, 2, 3, \cdots$.

Then we have from Remarks 20 and 21 that

$$\begin{aligned}\|\boldsymbol{e}_v(t(\bar{m}+2)B+1)\|_{2,\infty} \leq & \|\Phi_T(t(\bar{m}+2)B+1:2)\|_{2,\infty}\|\boldsymbol{e}_v(1)\|_{2,\infty} \\ \leq & c^t\|\boldsymbol{e}_v(1)\|_{2,\infty},\end{aligned}$$

and $\|\boldsymbol{e}_v(k)\|_{2,\infty} \leq c^{\lfloor \frac{k-1}{(\bar{m}+2)B} \rfloor} \max\limits_{1\leq t\leq(\bar{m}+2)B} \|\boldsymbol{e}_v(t)\|_{2,\infty}$.

We have that $\|\boldsymbol{e}(t)\|_{2,\infty} = \|\boldsymbol{T}\boldsymbol{e}_v(t-1)\|_{2,\infty} \leq \|\boldsymbol{T}\|_{2,\infty}\|\boldsymbol{e}_v(t-1)\|_{2,\infty}$ from Remark 20. As T_i is a homogeneous and continuous M_i-Fejer mapping and 0 is one of its fixed points, $\|T_i\|_2 \leq 1$, and thus $\|\boldsymbol{T}\|_{2,\infty} \leq 1$. So we have that $\|\boldsymbol{e}(t)\|_{2,\infty} \leq \|\boldsymbol{e}_v(t-1)\|_{2,\infty}$. Then we can conclude that $\boldsymbol{e}(t)$ vanishes at a geometric rate, and as a result, $x_i(t)$ converges to $x^\star$, the solution of $Ax = b$, at the same geometric rate. ■

5.2.3 Multiple Solution Case

In this part, we show the geometric convergence rate of (5.1) when $Ax = b$ has more than one solutions, and that the final convergence point is determined by the initializations, communication topologies, and the minimum 2-norm solution of $Ax = b$. We do this by an orthogonal decomposition of the agents' estimates $x_i(k)$ onto the row space and null space of A.

First, we investigate the orthogonal projection onto $\ker(A)$. Let K be the matrix whose columns are a basis of $\ker(A)$. The projection matrix onto $\ker(A)$ is $P_{\ker(A)} =$

$K(K^TK)^{-1}K^T$. Then

$$\begin{aligned}
&P_{\ker(A)}x_i(k+1)\\
=&P_{\ker(A)}v_i(k)-\gamma_i\frac{\|A_iv_i(k)-b_i\|^2P_{\ker(A)}A_i^T(A_iv_i(k)-b_i)}{\|A_i^T(A_iv_i(k)-b_i)\|_2^2}]\\
=&\sum_{j=1}^{\bar{m}}w_{ij}(k)P_{\ker(A)}x_j(k),
\end{aligned}$$

where we have used the facts that $P_{\ker(A)}A_i^T=0$ and that $P_{\ker(A)}$ is linear. So we have that the dynamics of the orthogonal projections of the agents' estimates onto $\ker(A)$ are consensus processes. As a result, the orthogonal projections of the agents' estimates onto $\ker(A)$, $P_{\ker(A)}x_i(k)$, $i=1,2,\cdots,\bar{m}$, converge to the same point, denoted as $x^\star_{NA}$, at a geometric rate determined by the communication topologies under Assumption 1.

Then we investigate the projection onto $\text{span}(A^T)$, the row space of A. The projection matrix onto $\text{span}(A^T)$ is $P_{\text{span}(A^T)}=A_B^T(A_BA_B^T)^{-1}A_B$, where A_B has full row rank with $\text{span}(A_B^T)=\text{span}(A^T)$. Let $x^\star$ be a solution of $Ax=b$, $e_i(k)=x_i(k)-x^\star$ and $e_{v_i}(k)=v_i(k)-x^\star$. Also, notice that $A_ix^\star=b_i$. We have that

$$\begin{aligned}
P_{\text{span}(A^T)}e_{v_i}(k)&=P_{\text{span}(A^T)}(\sum_{j=1}^{\bar{m}}w_{ij}(k)e_j(k))\\
&=\sum_{j=1}^{\bar{m}}w_{ij}(k)P_{\text{span}(A^T)}e_j(k),\qquad(5.7)\\
P_{\text{span}(A^T)}e_i(k+1)&=P_{\text{span}(A^T)}e_{v_i}(k)-\frac{\gamma_i\|A_iP_{\text{span}(A^T)}e_{v_i}(k)\|_2^2A_i^TA_iP_{\text{span}(A^T)}e_{v_i}(k)}{\|A_i^TA_iP_{\text{span}(A^T)}e_{v_i}(k)\|_2^2},
\end{aligned}$$

where we use the fact that $P_{\text{span}(A^T)}A_i^T=A_i^T$ because all columns of A_i^T are in $\text{span}(A^T)$. Eq. (5.7) forms a special case of (5.1) when $b=0$. Noticing that $P_{\text{span}(A^T)}e_{v_i}(k)\in\text{span}(A^T)$, all the limit points of (5.7) also lie in $\text{span}(A^T)$. As $Ax=0$ has a unique solution $x=0$ in $\text{span}(A^T)$ and $\text{span}(A^T)=(\bigcap_{i=1}^{\bar{m}}\ker(A_i))^\perp$, Lemma 22 applies here. Then repeating the

proof process of Proposition 1, we can prove that $P_{\text{span}(A^T)}e_i(k)$ converges to zero at a geometric rate. Thus, we have that $\lim_{k\to\infty} P_{\text{span}(A^T)}(x_i(k) - x^\star) = 0$.

As $\text{span}(A^T) = (\ker(A))^\perp$, we have that $P_{\text{span}(A^T)} + P_{\ker(A)} = I$, and that

$$x_i(k) - x^\star = P_{\text{span}(A^T)}(x_i(k) - x^\star) + P_{\ker(A)}(x_i(k) - x^\star).$$

As $P_{\ker(A)}x_i(k)$ reaches a consensus at some point, $x^\star_{NA}$, $P_{\ker(A)}(x_i(k) - x^\star)$ also reaches a consensus at $x^\star_{NA} - P_{\ker(A)}x^\star$. As $AP_{\ker(A)} = AK(K^TK)^{-1}K^T = 0$, we have that $P_{\ker(A)}(x_i(k) - x^\star)$ is a solution of $Ax = 0$. Then the consensus point $x^\star_{NA} - P_{\ker(A)}x^\star$ is also a solution of $Ax = 0$. Combined with $\lim_{k\to\infty} P_{\text{span}(A^T)}(x_i(k) - x^\star) = 0$, $x_i(k)$, $i = 1, 2, \cdots, \bar{m}$, converge to $(x^\star_{NA} - P_{\ker(A)}x^\star) + x^\star$, which is a solution of $Ax = b$ because $A(x^\star_{NA} - P_{\ker(A)}x^\star) = 0$ and $Ax^\star = b$. As both $P_{\text{span}(A^T)}(x_i(k) - x^\star)$ and $P_{\ker(A)}(x_i(k) - x^\star)$ converge at geometric rates, $x_i(k)$ also converges at a geometric rate.

Next, we specify the limit point $(x^\star_{NA} - P_{\ker(A)}x^\star) + x^\star$. From $P_{\text{span}(A^T)} + P_{\ker(A)} = I$, we have that

$$(x^\star_{NA} - P_{\ker(A)}x^\star) + x^\star = x^\star_{NA} + (I - P_{\ker(A)})x^\star = x^\star_{NA} + P_{\text{span}(A^T)}x^\star.$$

The first part, $x^\star_{NA}$, is the consensus point of $P_{\ker(A)}x_i(k)$, $i = 1, 2, \cdots, \bar{m}$, which is determined by the initializations, the matrix A and communication topologies. We next show that the second part, $P_{\text{span}(A^T)}x^\star$, is indeed the solution of $Ax = b$ with minimum 2-norm. As $P_{\text{span}(A^T)}x^\star \in \text{span}(A^T)$, we have that $K^TP_{\text{span}(A^T)}x^\star = 0$ because $\text{span}(K) = \ker(A)$. We also have $P_{\text{span}(A^T)}A^T = A^T$ because every row of A is in $\text{span}(A^T)$, and thus $AP_{\text{span}(A^T)}x^\star = Ax^\star = b$. From Lemma 17, we obtain that $P_{\text{span}(A^T)}x^\star$ is the minimum 2-norm solution of $Ax = b$, which is determined by $Ax = b$ itself.

The above discussions are summarized in the following proposition:

Proposition 2 *Under Assumptions 1 and 2, the agents' estimates in the distributed algorithm (5.1) converge to the same solution of $Ax = b$ at a geometric rate when $Ax = b$ has multiple solutions. The limit point of $x_i(k),\ k = 1, 2, 3, \cdots,\ i = 1, 2, \cdots, \bar{m}$ can be expressed as $x_{MN} + x_{NA}$, where x_{MN} is the solution of $Ax = b$ with minimum 2-norm and x_{NA} is the consensus point of the projections of the agents' initializations under the communication topologies $\mathcal{G}_t, t = 0, 1, 2, \cdots$.*

Remark 22 *The analysis in this part may also be used to analyze the algorithm in [49] when $Ax = b$ has multiple solutions.*

5.2.4 Alternative Proof of Convergence

If we are only interested in the convergence instead of convergence at a geometric rate, the convergence of the algorithm (5.1) can be proved in a simpler way with an extension of the results in [9] when the edge weights are selected as the reciprocals of the number of the agents' neighbors. In this part, we extend the results in [9] to jointly strongly connected topologies and then apply it to proving the convergence of (5.1).

Suppose we have $\bar{m}$ agents, and each agent owns a private continuous M_i-Fejer mapping $T_i,\ i = 1, \cdots, \bar{m}$. Also, suppose that they share at least one common fixed point, i.e., $\bigcap_{i=1}^{\bar{m}} M_i \neq \varnothing$. We have the following assumption on T_i.

Assumption 10 *$T_i,\ i = 1, \cdots, \bar{m}$ are continuous M_i-Fejer mappings, and $\bigcap_{i=1}^{\bar{m}} M_i \neq \varnothing$.*

To find a common fixed point of T_i in a distributed way, the following algorithm is proposed

in [9]:

$$x_i(k+1) = T_i(\frac{1}{|N_i(k)|}\sum_{j\in N_i(k)} x_j(k)), \tag{5.8}$$

where $x_i(k),\ i = 1, \cdots, \bar{m}$ are the estimates of the common fixed point by agent i, and $N_i(k)$ is the neighbor set of agent i at time k. It is proved in [9] that the algorithm (5.8) converges to a common fixed point of $T_i,\ i = 1, \cdots, \bar{m}$ when the communication topologies of the agent network are strongly connected at each time step. But we find that the algorithm (5.8) can also converge under jointly strongly connected communication topologies.

But first, we need the following lemma on the convergence of the sequence generated by the mappings.

Lemma 23 *[7]* [1] *Let $T_j, j = 1, \cdots, p$ be p M-Fejer mappings with respect to some norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $\{j_i\}_{i=1}^{\infty}$ be an admissible sequence, i.e. for any $1 \le r \le p$ there are infinitely many integers i such that $j_i = r$, and $x(1) \in \mathbb{R}^n$ is given. Then the sequence $\{x(k)\}$ generated by $x(k) = T_{j_k}(x(k-1))$ converges if and only if $T_j, j = 1, \cdots, p$, have a common fixed point. Moreover, in this case the limit of $\{x(k)\}$ is one of such common fixed points.*

Next, we show that the agents' estimates in the algorithm in (5.8) converges to a common fixed point of Fejer type mappings of agents.

Lemma 24 *The algorithm (5.8) converges to a common fixed point of $T_i,\ i = 1, \cdots, \bar{m}$ under Assumptions 1 and 10.*

Proof. Similar to the approach adopted in [9], let $v_i(k) = \frac{1}{|N_i(k)|}\sum_{j\in N_i} x_j(k)$. Then we have

[1] Continuous M-Fejer mappings are also called paracontracting operators in [7] or paracontractions in [9], but in this dissertation we use the term M-Fejer mapping to emphasize the fixed point set M.

that $v_i(k+1) = \frac{1}{|N_i(k+1)|} \sum_{j \in N_i} T_j v_j(k)$. Let $\boldsymbol{v}(k) = \begin{pmatrix} v_1(k) \\ \dots \\ v_{\bar{m}}(k) \end{pmatrix}$, and $W(k)$ be the weight matrix associated with the communication topology at time k with $w_{ij}(k) = \frac{1}{|N_i(k)|}$ if $(j,i) \in E$ and $w_{ij}(k) = 0$ else for algorithm (5.8). Then we have that $\boldsymbol{v}(k+q) = \Phi_T(k+q:k+1)\boldsymbol{v}(k)$, where Φ_T is defined in (5.6). From Lemma 1, we have that when $q \geq (\bar{m}-1)B$, all entries of $\Phi(k+q:k+1)$ are positive. Then applying Lemma 9, we know that $\Phi_T(k+q:k+1)$ is a continuous M-Fejer mapping, with $M = \{\mathbf{1}_{\bar{m}} \otimes y : y \in \bigcap_{i=1}^{\bar{m}} M_i\}$. As there are only a finite choice of $W(k)$, there are also only finite choices of $\Phi_T(k+q:k+1),\ k = 1,2,3,\cdots$. Then applying Lemma 23, we have that $\boldsymbol{v}(t(\bar{m}-1)B+1), t = 1,2,3,\cdots$ converges to some point y_M in M. From Lemma 8, we know that $\|\boldsymbol{v}(k+1) - y_M\|_{p,\infty} \leq \|\boldsymbol{T}\boldsymbol{v}(k) - y_M\|_{p,\infty} \leq \|\boldsymbol{v}(k) - y_M\|_{p,\infty}$. So we can conclude that the sequence $\boldsymbol{v}(k), k = 1,2,3,\cdots$ converges to $y_M \in M$. Taking into consideration that $M = \{\mathbf{1}_{\bar{m}} \otimes y : y \in \bigcap_{i=1}^{\bar{m}} M_i\}$, we can conclude that $v_i(k),\ i = 1,\cdots,\bar{m}$ reach a consensus at a common fixed point of $T_i,\ i = 1,2,\cdots,\bar{m}$. From (5.8) and the continuity of $T_i,\ i = 1,2,\cdots,\bar{m},\ x_i(k+1),\ i = 1,2,\cdots,\bar{m}$ also converge to the same common fixed point. ■

Remark 23 *Similar results to Lemma 24 are shown in [8, 9], where the communication topologies are time varying but strongly connected at each time instance. But in Lemma 24, we suppose that the communication topologies are jointly strongly connected, which is more general than that in [8, 9].*

We have shown that T_i in (5.2) is a continuous M_i-Fejer mapping with $M_i = \{x : A_i x = b_i\}$ in Section 5.2.1. Then from Lemma 24, we obtain the convergence of (5.1) when

$w_{ij}(k) = \frac{1}{|N_i(k)|}$.

The results of the analysis in this part are summarized in the following proposition:

Proposition 3 *Under Assumptions 1 and 10, the agents' estimates in (5.1) converge to a solution of $Ax = b$ when $w_{ij}(k) = \frac{1}{|N_i(k)|}$.*

Remark 24 *The alternative proof provided in this part, i.e., the proof of Proposition 3 is simpler than those in Sections 5.2.2 and 5.2.3, but it only guarantees convergence rather than convergence at a geometric rate, and it does not provide any information on which point the algorithm converges to when $Ax = b$ has multiple solutions. Also, the weights of the edges in Proposition 3 can only be selected as the reciprocal of the numbers of the neighbors of the agents, while in Theorem 7, the edge weights can be chosen arbitrarily under Assumption 2.*

5.3 Convergence Rate Upper Bounds for Special Cases

In this section, we will give an analysis on the upper bound of the convergence rate with respect to the parameter γ_i and the condition number of A. As a meaningful upper bound for the general cases are too complex to be obtained, we consider two special cases when A is an orthogonal matrix and when the communication graph is complete with a uniform edge weight. For simplicity, we suppose in this section that $b = 0$.[2] Then the

[2]The analysis in this part also applies to general b, because the convergence of the agents' estimates to a solution of $Ax = b$ is equivalent to the convergence of estimation errors to zero. See Sections 5.2.2 and 5.2.3 for details.

algorithm (5.1) becomes

$$v_i(k) = \sum_{j=1}^{\bar{m}} w_{ij}(k) x_j(k),$$

$$x_i(k+1) = v_i(k) - \frac{\gamma_i \|A_i v_i(k)\|_2^2 A_i^T A_i v_i(k)}{\|A_i^T A_i v_i(k)\|_2^2}$$

and $T_i x = x - \frac{\gamma_i \|A_i x\|_2^2 A_i^T A_i}{\|A_i^T A_i x\|_2^2} x$.

5.3.1 Orthogonal Matrix

When A is orthogonal, we have that $A_i A_i^T$ is an m_ith order identity matrix. Then we have that $\frac{\|A_i v_i(k)\|_2^2}{\|A_i^T A_i v_i(k)\|_2^2} = 1$. It follows that

$$\boldsymbol{v}(k+1) = (W(k+1) \otimes I)\boldsymbol{T}\boldsymbol{v}(k),$$

where $\boldsymbol{T}$ is defined in (2.4), and $T_i = I - \gamma_i A_i^T A_i$, which is linear. It follows that $\boldsymbol{v}(k+s) = \Phi_T(k+s:k+1)\boldsymbol{v}(k)$, where Φ_T is defined in (5.6), which is also linear in this case.

Remark 25 *In general, the algorithm in (5.1) is nonlinear while in this part it is linear as A is orthogonal. So the analysis in this part, e.g., Lemmas 26, 27, and 29, does not apply to the general case in Thoerem 7.*

To make the analysis in this part clearer, we will use the concept of graph compositions in [4] which is more closely related to the product of weight matrix than graph unions. The composition of two graphs $\mathcal{G}_2$ and $\mathcal{G}_1$, written as $\mathcal{G}_2 \circ \mathcal{G}_1$, with the same vertex set V is a graph with the vertex set V, and the edge set E such that $(j,i) \in E$, if there exists a vertex k such that (j,k) is an edge of $\mathcal{G}_1$ and (k,i) is an edge of $\mathcal{G}_2$ [4]. Here $(j,k),(k,i)$ is a route from vertex j to vertex i in $\mathcal{G}_2 \circ \mathcal{G}_1$. The definition of the composition and the route

can be extended to any finite sequence of graphs $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_k$. We have that $W(2)W(1)$ is the weight matrix of $\mathcal{G}_2 \circ \mathcal{G}_1$ [4]. For the communication topologies of the agent network, we make an assumption that is slightly different from Assumption 1.

Assumption 11 . *There exists a positive integer B such taht the compositions of $\mathcal{G}_{(k+1)B}$, $\mathcal{G}_{(k+1)B-2}$, $\cdots$, $\mathcal{G}_{kB+1}$ are strongly connected for $k = 0, 1, 2, \cdots$.*

We will use the following lemma on the composition of a sequence of strongly connected graphs:

Lemma 25 *[4] Let $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_q$ be a finite sequence of strongly connected $\bar{m}$th order graphs. If $q \geq \bar{m} - 1$, then $\mathcal{G}_q \circ \cdots \mathcal{G}_1$ is complete.*

It is easy to verify through matrix product that the entries of Φ_T are polynomials of T_i consisting of many monomials of T_i. Next, we will obtain some results on such monomials. First, we have the following lemma relating the monomials to the communication topologies.

Lemma 26 *Let $W(1), \cdots, W(q)$ be a sequence of weight matrices associated with graphs $\mathcal{G}_1, \cdots, \mathcal{G}_q$. If $j = i_1, \cdots, i_{q+1} = i$ is a route over $\mathcal{G}_1, \cdots, \mathcal{G}_q$, then $T_{i_q} T_{i_{q-1}} \cdots T_{i_1}$ is a component of the ijth entry of $\Phi_T(q:1)$ with coefficient $W_{ii_q}(q) W_{i_q i_{q-1}}(q-1) \cdots W_{i_2 j}(1)$ when A is orthogonal.*

Proof. We prove by induction.

When $q = 1$, the ijth entry of $\Phi_T(q:1)$ is $w_{ij} T_j$ if there is an edge from j to i. Suppose that it holds for $q-1$. Then $\Phi_T(q:1) = (W_q \otimes I)\boldsymbol{T}\Phi_T(q-1:1)$. As there is a route $j = i_1, \cdots, i_q$ over $\mathcal{G}_1, \mathcal{G}_1, \cdots, \mathcal{G}_{q-1}$, the $i_q j$th entry of $\Phi_T(q-1:1)$ has a component

with $T_{i_{q-1}} \cdots T_{i_1}$. Then by the matrix product, the ijth entry of $\Phi_T(q:1)$ has a component with $T_{i_q} T_{i_{q-1}} \cdots T_{i_1}$.

It is easy to verify from matrix product that the corresponding coefficient is $W_{ii_q} W_{i_q i_{q-1}} \cdots W_{i_2 j}$. ■

Then after a long enough time, we can see from the following lemma that every entry of Φ_T contains a monomial that is composed of all possible T_i.

Lemma 27 *Under Assumption 11, each entry of $\Phi_T(q:1)$ in (5.6) has a component composed of the product of all $T_i, i = 1, \cdots, \bar{m}$ when $q \geq (\bar{m}-1)^2 B + 1$ if A is orthogonal.*

Proof. Let $\Phi(t:s)$ be defined as in (2.1). From Assumption 11, we have that the products $\Phi(B,1)$, $\Phi(2B:B+1)$, $\cdots$, $\Phi((\bar{m}-1)B:(\bar{m}-2)B+1)$ are all associated with strongly connected graphs. Denote the corresponding graphs as $\mathcal{G}_{\Phi_{1:B}}, \mathcal{G}_{\Phi_{B+1:2B}}, \cdots, \mathcal{G}_{\Phi_{(\bar{m}-2)B+1:(\bar{m}-1)B}}$, which are composition of graphs $\mathcal{G}_B \circ \mathcal{G}_{B-1} \cdots \mathcal{G}_1$, $\mathcal{G}_{2B} \circ \mathcal{G}_{2B-1} \cdots \mathcal{G}_{B+1}$, $\cdots$, $\mathcal{G}_{(\bar{m}-1)B} \circ \mathcal{G}_{(\bar{m}-1)B-1} \cdots \mathcal{G}_{(\bar{m}-2)B+1}$. Then from Lemma 1, we know that the composition graph of $\mathcal{G}_{\Phi_{1:B}}, \mathcal{G}_{\Phi_{B+1:2B}}, \cdots, \mathcal{G}_{\Phi_{(\bar{m}-2)B+1:(\bar{m}-1)B}}$, denoted as $\mathcal{G}_{1:(\bar{m}-1)B}$, is complete. Then for arbitrary $i_1, i_2 = 1, \cdots, \bar{m}$, there is an edge from i_1 to i_2 in $\mathcal{G}_{1:(\bar{m}-1)B}$. Then we can find a route $j = i_1, i_2, \cdots, i_{\bar{m}} = i$, where $i_1, i_2, \cdots, i_{\bar{m}}$ are distinct vertices, such that there is a route from $j = i_1$ to i_2 over $\mathcal{G}_{\Phi_{1:(\bar{m}-1)B}}$, that from i_2 to i_3 over $\mathcal{G}_{\Phi_{(\bar{m}-1)B+1:2(\bar{m}-1)B}}$, $\cdots$, and that from $i_{\bar{m}-1}$ to $i_{\bar{m}} = i$ over $\mathcal{G}_{\Phi_{(\bar{m}-2)(\bar{m}-1)B+1:(\bar{m}-1)^2 B}}$. And also we can find a route from i to i from $\mathcal{G}_{(\bar{m}-1)^2 B}$ to $\mathcal{G}_{(\bar{m}-1)^2 B+1}$ because of self edges. We can conclude from Lemma 26 that the ijth entry of $\Phi_T(q:1)$ has a component with all possible T_i, $i = 1, 2, \cdots, \bar{m}$ when $q \geq (\bar{m}-1)^2 B + 1$. ■

Next, we will give an estimation on the upper bound of $\Phi_T(q:1)$. To do that, we need the following lemma to show the Fejer-type property of the component mentioned in Lemma 26.

Lemma 28 *[48] Let $T_j, j = 1, \cdots, \bar{m}$ be M_j-Fejer type mappings, and $M = \bigcap_{j=1}^{\bar{m}} M_j \neq \varnothing$, then $T = T_{\bar{m}} T_{\bar{m}-1} \cdots T_1$ is an M-Fejer type mapping.*

With the above results, we then estimate the upper bound of the convergence rate in the following lemma:

Lemma 29 *If A is orthogonal, $\|\Phi_T(q:1)\|_{2,\infty} \leq 1 - (1-c)\eta^q$ under Assumptions 2 and 11, where $0 < c < 1$.*

Proof. From Lemma 27, every entry in $\Phi_T(q:1)$ has a component consisting of all possible T_i, denoted by $c_i^{(a)} P_{T,i}^{(a)}$, where the coefficient $c_i^{(a)}$ is the product of the edge weights in the corresponding routes, and $P_{T,i}^{(a)}$ is the product of all possible T_i. As $Ax = 0$ admits a unique solution due to the orthogonality of A, we have that the intersection of $M_i = \text{Fix}(T_i) = \{x : A_i x = 0\},\ i = 1, 2, \cdots, \bar{m}$ is $\{0\}$, and its orthogonal complement is $\mathbb{R}^n$. From Lemma 28, $P_{T,i}^{(a)}$ is an M-Fejer mapping with $M = \{0\}$. Also, as $T_i, i = 1, 2, \cdots, \bar{m}$ are homogeneous, $P_{T,i}^{(a)}$ is also homogeneous. Then from Lemma 22, we know that there exists a positive constant $c < 1$ such that $\|P_{T,i}^{(a)}\|_2 \leq c$.

We denote the other monomials in the ith entry of $\Phi_T(q:1)$ as $c_i^{(na,1)} P_{T,i}^{(na,1)}$, $c_i^{(na,2)} P_{T,i}^{(na,2)}$, $\cdots$, $c_i^{(na,f)} P_{T,i}^{(na,f)}$, where f is the number of these monomials, $P_{T,i}^{(na,k)}, k = 1, 2, \cdots, f$ are products of some of $T_j, j = 1, 2, \cdots, \bar{m}$, and $c_i^{(na,k)}$ are the coefficients. As at least one $T_j, j = 1, 2, \cdots, \bar{m}$ is missing in $P_{T,i}^{(na,k)}, k = 1, 2, \cdots, f$, the fixed point

sets of $P_{T,i}^{(na,k)}, k = 1, 2, \cdots, f$ are linear subspaces containing nonzero element. Hence, $\|P_{T,i}^{(na,k)}\|_2 = 1, k = 1, 2, \cdots, f$.

We can see that $\|\cdot\|_2$ of the ith entry of $\Phi_T(q:1)$ is no greater than $c_i^{(a)}\|P_{T,i}^{(a)}\|_2 + \sum_{k=1}^{f} c_i^{(na,k)}\|P_{T,i}^{(na,k)}\|_2$, and thus no greater than $c_i^{(a)}c + \sum_{k=1}^{f} c_i^{(na,k)} = (c_i^{(a)} + \sum_{k=1}^{f} c_i^{(na,k)}) - (1-c)c_i^{(a)}$. Note that $c_i^{(a)} + \sum_{k=1}^{f} c_i^{(na,k)}$ is in fact the summation of all entries in th ith row of $\Phi(q:1)$ and thus is one. Hence $\|\cdot\|_2$ of the ith entry of $\Phi_T(q:1)$ is no greater than $1-(1-c)c_i^{(a)}$. Thus $\|\Phi_T(q:1)\|_{2,\infty} \leq 1-(1-c)\min_i c_i^{(a)}$. Note that $c_i^{(a)}$ is the product of q edge weights. From Assumption 2, there exists $\eta > 0$ such that $w_{ij}(t) \geq \eta$ if $w_{ij}(t) > 0$, we have that $c_i^{(a)} \geq \eta^q$ and $\min_i c_i^{(a)} \geq \eta^q$, so $\|\Phi_T(q:1)\|_{2,\infty} \leq 1-(1-c)\eta^q$. ■

Then, we will use the above lemmas to obtain an upper bound of the convergence rate of the algorithm (5.1) for an orthogonal matrix A. From Lemma 27, we know that every entry of $\Phi_T(k(\bar{m}-1)^2)B+1:k+1)$ has a monomial in the form of $\lambda\prod_j(I-\gamma_j A_j^T A_j)$, where it contains all possible $j = 1, 2, \cdots, \bar{m}$ and may have repetitive items. As A is orthogonal, it follows that $A_i A_j^T = 0, i \neq j$. Then we have that for $i \neq j$,

$$\begin{aligned}(I-\gamma_i A_i^T A_i)(I-\gamma_j A_j^T A_j) =& I-\gamma_i A_i^T A_i - \gamma_j A_j^T A_j + \frac{\gamma_i\gamma_j}{4}A_i^T A_i A_j^T A_j \\ =& I-\gamma_i A_i^T A_i - \gamma_j A_j^T A_j \\ =& (I-\gamma_j A_j^T A_j)(I-\gamma_i A_i^T A_i).\end{aligned}$$

Then by re-ordering the items in the monomial, we have that $\prod_j(I-\gamma_j A_j^T A_j) = \prod_{i=1}^{\bar{m}}(I-\gamma_i A_i^T A_i)^{q_i}$, where q_i is the number of occurrence of $I-\gamma_i A_i^T A_i$. Note that $A_i^T A_i$ is idempotent because

$$(A_i^T A_i)^p = A_i^T A_i A_i^T A_i \cdots A_i^T A_i = A_i^T A_i,$$

where we have used the fact that A is orthogonal and $A_i A_i^T = I$. Then we have that

$$\begin{aligned}(I - \gamma_i A_i^T A_i)^{k_i} &= \sum_{k=0}^{q_i} \binom{q_i}{k} (-\gamma_i A_i^T A_i)^k \\ &= I - \gamma_i A_i^T A_i + (\sum_{k\geq 2} \binom{q_i}{k} (-\gamma_i)^k) A_i^T A_i,\end{aligned}$$

where we denote $(-\gamma_i A_i^T A_i)^0 = I$ for simplicity. Denote $R_i = (\sum\limits_{k\geq 2} \binom{q_i}{k})(-\gamma_i)^k) A_i^T A_i$. Because A is orthogonal, $A_i A_j^T = 0$. It follows that $R_i A_j^T = 0$ and $R_i R_j = 0$. Then we have that

$$\begin{aligned}\prod_{i=1}^{\bar{m}} (I - \gamma_i A_i^T A_i)^{q_i} &= \prod_{i}^{\bar{m}} (I - \gamma_i A_i^T A_i + R_i) \\ &= I - (\sum_{i=1}^{\bar{m}} \gamma_i A_i^T A_i) + \sum_{i=1}^{\bar{m}} R_i.\end{aligned}$$

As $\sum\limits_{k\geq 2} \binom{q_i}{k} (-\gamma_i)^k = (1-\gamma_i)^{q_i} - (1-\gamma_i) < 0$ when $\gamma_i \in (0,1)$, R_i is negative semi-definite. It follows that

$$\begin{aligned}\prod_{i=1}^{\bar{m}} (I - \gamma_i A_i^T A_i)^{q_i} &\leq I - (\sum_{i=1}^{\bar{m}} \gamma_i A_i^T A_i) \\ &= I - A^T \mathrm{Diag}(\gamma_i) A \\ &\leq I - \min_{i\in V}\{\gamma_i\} A^T A \\ &= I - \min_{i\in V}\{\gamma_i\} I,\end{aligned}$$

where $\mathrm{Diag}(\gamma_i)$ is a diagonal matrix of which the diagonal entries are γ_i and the off-diagonal ones are zero. Then it follows from Lemma 29 that

$$\|\Phi_T(k(\bar{m}-1)^2)B + 1 : k+1)\|_{2,\infty} < 1 - \min_{i\in V}\{\gamma_i\} \eta^{(\bar{m}-1)^2 B + 1}.$$

The above discussion is summarized in the following proposition:

Proposition 4 *Suppose A is orthogonal. Then under Assumptions 2 and 11, $\|\Phi_T(k(\bar{m}-1)^2)B+1:k+1)\|_{2,\infty} < 1-\min_{i\in V}\{\gamma_i\}\eta^{(\bar{m}-1)^2B+1}$, where $\bar{m}$ is the number of agents, γ_i is a parameter of the algorithm in (5.1), B is defined in Assumption 11, and η is defined in Assumption 2.*

Remark 26 *From Theorem 4, we can see that a larger $\min_{i\in V}\{\gamma_i\}$ will help to accelerate the convergence of the algorithm (5.1) when A is orthogonal. A smaller B and a bigger η, which are used to characterize the connectivity of the agent network, also contribute to faster convergence.*

5.3.2 Complete Graph with Uniform Edge Weight

In the previous subsection, we obtain an upper bound unrelated to A because A is assumed to be orthogonal. In this subsection, we will consider another special case and show that the matrix A indeed has an influence on the convergence rate.

In this part, we assume that the communication topology of the agent network is complete and the weight of the edges is $\frac{1}{\bar{m}}$. Then the weight matrix $W = \frac{1}{\bar{m}}\mathbf{1}_{\bar{m}}\mathbf{1}_{\bar{m}}^T$, and $\boldsymbol{v}(k+1) = \frac{1}{\bar{m}}\mathbf{1}_{\bar{m}}\mathbf{1}_{\bar{m}}^T\boldsymbol{T}\boldsymbol{v}(k)$. Then we have that

$$\begin{aligned} v_i(k+1) &= \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}(I-\frac{\gamma_j\|A_jv_j(k)\|_2^2A_j^TA_j}{\|A_j^TA_jv_j(k)\|_2^2})v_j(k) \\ &= \frac{1}{\bar{m}}[\sum_{j=1}^{\bar{m}}(I-\frac{\gamma_j\|A_jv_j(k)\|_2^2A_j^TA_j}{\|A_j^TA_jv_j(k)\|_2^2})]v_i(k), \end{aligned}$$

where we have used the fact that $v_i(k) = v_j(k)$ in the second equality. As $\|A_j^TA_jv_j(k)\|_2^2 \le \sigma_{\max}^2(A_i)\|A_jv_j(k)\|_2^2$, we have that

$$I-\frac{\gamma_j\|A_jv_j(k)\|_2^2A_j^TA_j}{\|A_j^TA_jv_j(k)\|_2^2} \le I-\frac{\gamma_j}{\sigma_{\max}^2(A_j)}A_j^TA_j.$$

As $A_j^T A_j \leq \sum_{i=1}^{\bar{m}} A_i^T A_i = A^T A$, $\sigma_{\max}^2(A_j) \leq \sigma_{\max}^2(A)$. It follows that

$$I - \frac{\gamma_j \|A_j v_j(k)\|_2^2 A_j^T A_j}{\|A_j^T A_j v_j(k)\|_2^2} \leq I - \frac{\gamma_j}{\sigma_{\max}^2(A)} A_j^T A_j,$$

and

$$\begin{aligned}
\frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} (I - \frac{\gamma_j \|A_j v_j(k)\|_2^2 A_j^T A_j}{\|A_j^T A_j v_j(k)\|_2^2}) \leq & \frac{1}{\bar{m}} \sum_{j=1}^{\bar{m}} (I - \frac{\gamma_j}{\sigma_{\max}^2(A)} A_j^T A_j) \\
\leq & I - \frac{\min_{i \in V}\{\gamma_i\}}{\bar{m}\sigma_{\max}^2(A)} \sum_{j=1}^{\bar{m}} A_j^T A_j \\
= & I - \frac{\min_{i \in V}\{\gamma_i\} A^T A}{\bar{m}\sigma_{\max}^2(A)} \\
\leq & (1 - \frac{\min_{i \in V}\{\gamma_i\} \sigma_{\min}^2(A)}{\bar{m}\sigma_{\max}^2(A)}) I.
\end{aligned}$$

The above discussion in this part is summarized in the following proposition:

Proposition 5 *When the communication topology of the agent network is complete and the edge weight are chosen as the reciprocal of the number of agents, the convergence rate of (5.1) is upper bounded by* $(1 - \frac{\min_{i \in V}\{\gamma_i\} \sigma_{\min}^2(A)}{\bar{m}\sigma_{\max}^2(A)})$, *where* $\bar{m}$ *is the number of agents,* $\sigma_{\min}(A)$ *and* $\sigma_{\max}(A)$ *are the smallest and largest singular values of* A, *respectively.*

Remark 27 *From Theorem 5, we can see that the condition number of* A, *which is* $\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$, *plays a role in the convergence rate of the algorithm (5.1). The larger the condition number is, the slower the convergence rate is in this case. A larger* $\min\{\gamma_i\}$ *also helps to accelerate the convergence, which is the same as in Remark 26.*

5.4 Simulations

In this section, we provide some simulation examples to show the effectiveness of the algorithm proposed in this chapter. First we show the convergence of the algorithm (5.1) for $Ax = b$ with general A and b. Then we illustrate the influence of γ_i on the convergence rate when A is orthogonal with a simulation example. For simplicity, we make all $\gamma_i,\ i \in V$ equal, i.e., $\gamma_i = \gamma$ and thus $\min_i \gamma_i = \gamma$. Finally, we do simulations on different matrices A with different condition numbers.

5.4.1 General Matrix A

We consider an agent network consisting of 50 agents under the topology of a directed ring. Here each agent knows one row of the augmented matrix $\begin{bmatrix} A & b \end{bmatrix}$, where $A \in \mathbb{R}^{50\times 70}$ and $b \in \mathbb{R}^{15}$ are both generated randomly. From Fig. 5.1, for all three cases of $\gamma = 0.25, 0.5, 0.75$, the estimation errors tend to zero. Also, we can see from the two figures that γ has an influence on the performance of the algorithm in general cases. Though the results in the example seem to support a larger γ for faster convergence in general cases, but the rigorous theoretic analysis is left blank.

5.4.2 Orthogonal Matrix A

We consider an agent network consisting of 70 agents under the topology of a directed ring, and each agent knows one row of the augmented matrix $\begin{bmatrix} A & b \end{bmatrix}$, where $A \in \mathbb{R}^{70\times 70}$ is an orthogonal matrix and $b \in \mathbb{R}^{15}$ is generated randomly.

We select γ to be 0.25, 0.5, and 0.75 and show the simulation results in Fig. 5.2.

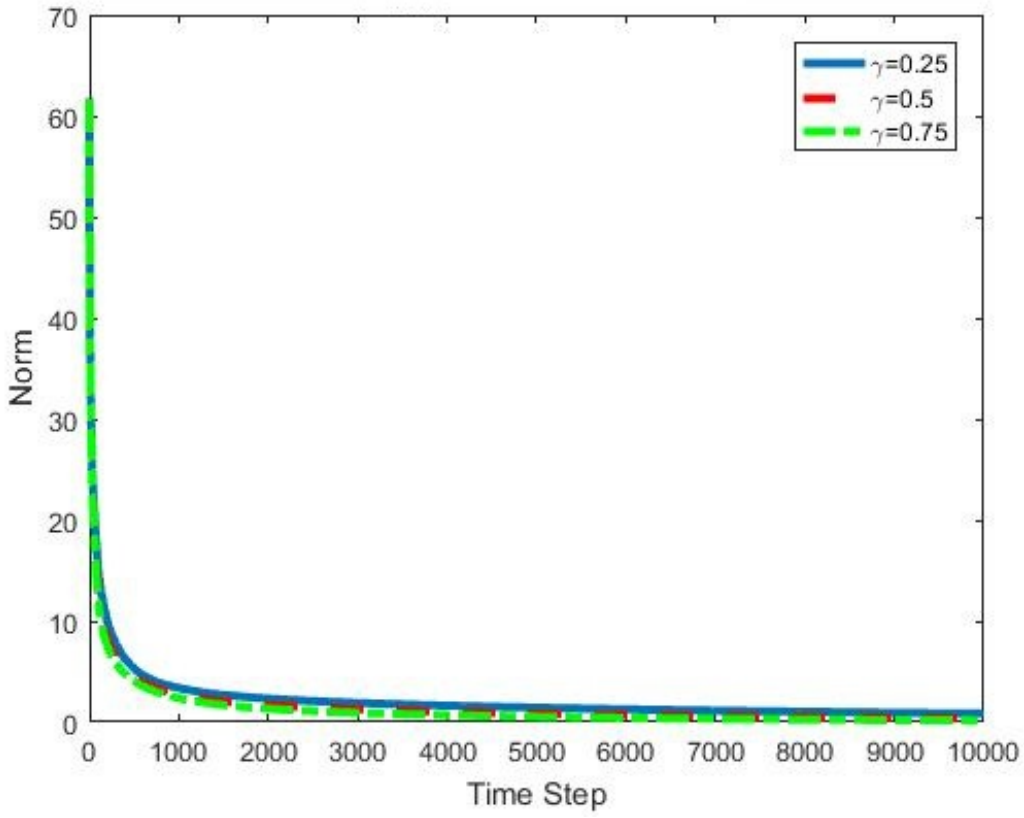

Figure 5.1: $\|\cdot\|_{2,\infty}$ of estimation errors

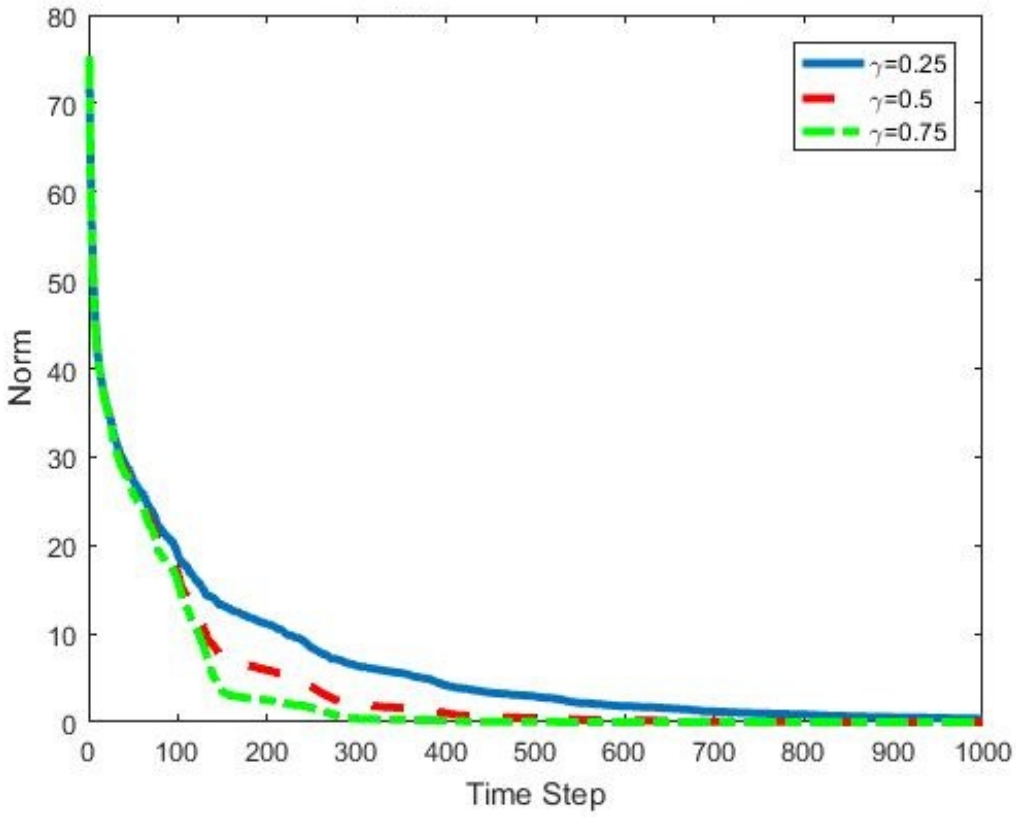

Figure 5.2: $\|\cdot\|_{2,\infty}$ of estimation errors

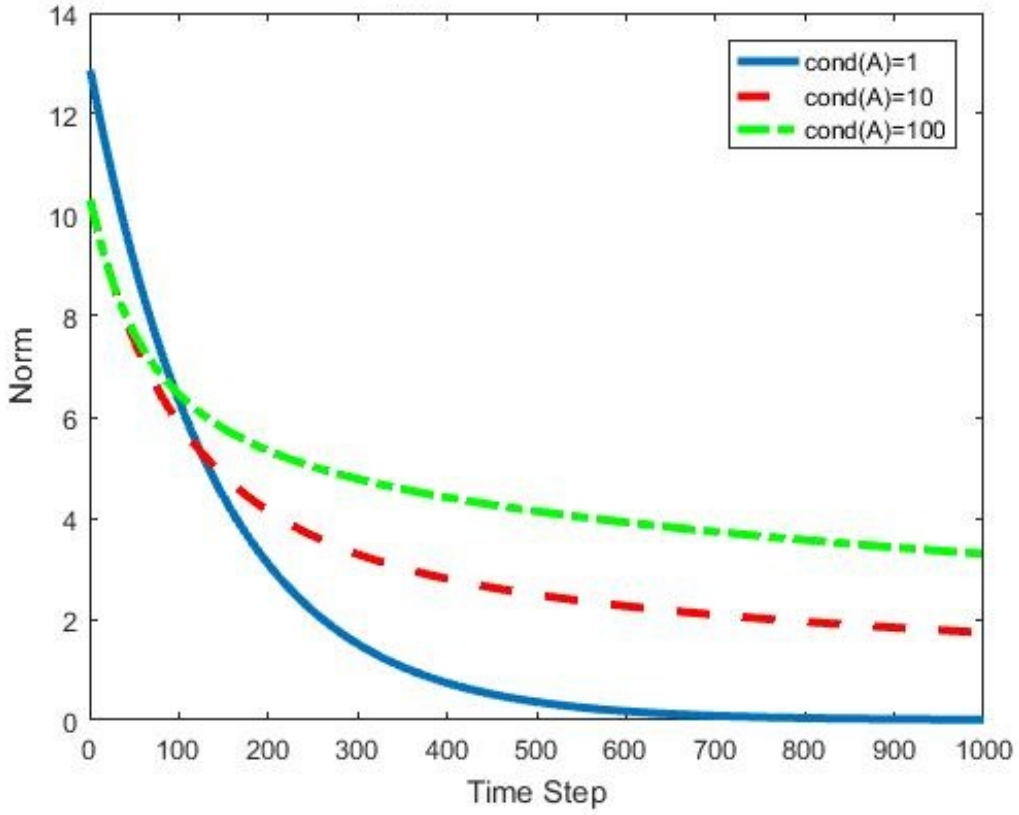

Figure 5.3: Change of $\|\cdot\|_{2,\infty}$ of Estimation Errors

From Fig. 5.2, we can see that a larger γ makes the algorithm converge faster, which coincides the analysis in Section 5.3.1.

5.4.3 Complete Graph

We consider an agent network consisting of 70 agents under the topology of a complete graph and each agent knows one row of the augmented matrix $\begin{bmatrix} A & b \end{bmatrix}$, where $b \in \mathbb{R}^{15}$ is generated randomly and A is generated three times with condition number being 1, 10, 100, respectively. Here γ is set to be 0.5.

The simulation results are shown in Fig. 5.3. From it, we can find that the larger the condition number of A is, the slower the agents converge, which is consistent with the result in Section 5.3.2.

Chapter 6

Communication-Efficient Distributed Algorithm to Solve $Ax = b$ with Sparse A

In the previous section, we proposed a distributed algorithm to solve a system of linear equations $Ax = b$. The algorithm requires agents to transmit the states of all agents, i.e., the whole vector x, to all its neighbors. But in many applications, especially when A is sparse with many zero entries or blocks, it is unnecessary and perhaps impossible for the agents to know the whole vector. For example, in power systems, the Ybus is usually sparse because each bus is physically connected to only a few buses. The buses might not be able to obtain the states of all buses in the system due to various restrictions, e.g.,limited storage and limited communication. So it is important to design distributed algorithms requires less communication for sparse matrix. In this chapter, we propose a distributed

algorithm to solve $Ax = b$ for sparse A with special structure. The proposed algorithm reduces communication burden among agents, because it allows agents to share through a communication link their own states and the states of one of their neighbors instead of the states of all agents.

6.1 Preliminary

In this part, we consider that the matrix A has the follow sparsity structure:

Definition 11 (Laplacian sparsity) *A matrix A has the Laplacian sparsity structure of a graph $\mathcal{G}$ if $a_{ij} \neq 0$ only if i and j are connected in $\mathcal{G}$, where a_{ij} is the (i,j)th entry of matrix A. A block matrix A has the Laplacian sparsity structure of a graph $\mathcal{G}$ if $A_{ij} \neq 0$ only if i and j are connected in $\mathcal{G}$, where A_{ij} is the (i,j)th block of matrix A.*

Example 1 *The Ybus of a power system has Laplacian sparsity structure if the communication topology is the same as the physical one.*

For the graph $\mathcal{G}$, we further assume that

Assumption 12 *The communication topology of the agent network is fixed, undirected, and connected.*

6.2 Problem Formulation

We can regard solving $Ax = b$ as an optimization problem to minimize $\frac{1}{2}\|Ax-b\|^2$, and equivalently,

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i\in V}\|A_i x_i - b_i\|^2 \\ \text{subject to} \quad & x_i = x_j,\ \forall i,j \in V \end{aligned} \tag{6.1}$$

with Assumption 12, where A_i is the rows of matrix A owned by agent i.

Next, we will make use of the Laplacian sparsity structure of A. Let $A_i^{(i)}$ be a row of block matrix whose blocks are those nonzero ones in A_i and $x^{(i)}$ the corresponding entries in x. For example,

Example 2 *We consider a network composed of four nodes and the Laplacian of the network is* $L = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$. *Then* $A = \begin{pmatrix} 1 & 3.4 & 0 & 0 \\ 0.8 & 5 & -9 & 0 \\ 0 & -6.23 & -3 & 6 \\ 0 & 0 & -5 & -0.96 \end{pmatrix}$ *is a matrix with Laplacian sparsity structure and each block of A is a scalar. If each agent owns one row of A, then* $A_1^{(1)} = \begin{pmatrix} 1 & 3.4 \end{pmatrix}$, $A_2^{(2)} = \begin{pmatrix} 0.8 & 5 & -9 \end{pmatrix}$, $A_3^{(3)} = \begin{pmatrix} -6.23 & -3 & 6 \end{pmatrix}$, *and* $A_4^{(4)} = \begin{pmatrix} -5 & -0.96 \end{pmatrix}$. *Correspondingly,* $x^{(1)} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $x^{(2)} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, $x^{(3)} = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix}$, *and* $x^{(4)} = \begin{pmatrix} x_3 \\ x_4 \end{pmatrix}$.

To simplify analysis, we further assume that

Assumption 13 *Each agent has local indices for its neighbors.*

Assumption 13 is not necessary for implementation of the distributed algorithm to be designed, but it simplifies the analysis.

Example 3 *In Example 2, if agent 4 owns the fourth row of A, then $k_3^{(4)} = 1$, $k_4^{(4)} = 2$, $e_{k_3}^{(4)} = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$, and $e_{k_4}^{(4)} = \begin{pmatrix} 0 & 1 \end{pmatrix}^T$.*

Then we can transform (6.1) to an optimization problem dependent on $A_i^{(i)}$ and $x_i^{(i)}$ as stated below:

Lemma 30 *Under Assumption 12 and the Laplacian sparsity of A, solving $Ax = b$ is equivalent to solving*

$$\begin{aligned} \min \quad & f = \frac{1}{2} \sum_{i \in V} \|A_i^{(i)} x^{(i)} - b_i\|^2 \\ \textit{subject to} \quad & x_i^{(i)} = x_i^{(j)},\ x_j^{(j)} = x_j^{(i)},\ (i,j) \in E, \end{aligned} \tag{6.2}$$

where $x_j^{(i)}$ is the local estimate on x_j by agent i.

Proof. If $x^\star$ is a solution of $Ax = b$, it is obvious that $x^\star$ is also an optimal solution to (6.2) with $x^{(i)} = \begin{pmatrix} x_j^\star \end{pmatrix}_{j \in N_i}$, which is a collector of the entries in $x^\star$ corresponds the neighbors of agent i.

If $x^{(1)}, x^{(2)}, \cdots, x^{(\bar{m})}$ form a solution of (6.2), denote $x^\star = \begin{pmatrix} x_1^{(1)} \\ x_2^{(2)} \\ \vdots \\ x_{\bar{m}}^{(\bar{m})} \end{pmatrix}$. For any agent i, we have $x_j^\star = x_j^{(j)} = x_j^{(i)}, \forall j \in N_i$. Notice that A has Laplacian sparsity structure, so $x^{(i)} = \begin{pmatrix} x_j^\star \end{pmatrix}_{j \in N_i}$ and thus $A_i x^\star = A_i^{(i)} x^{(i)} = b_i$. ■

Lemma 31 *Let*

$$\begin{aligned} f_p &= \frac{1}{2}[\sum_{i\in V}\|A_i^{(i)}x^{(i)}-b_i\|^2+\sum_{(i,j)\in E}(\|x_i^{(i)}-x_i^{(j)}\|^2+\|x_j^{(j)}-x_j^{(i)}\|^2)] \\ &= \frac{1}{2}\sum_{i\in V}[\|A_i^{(i)}x^{(i)}-b_i\|^2+\frac{1}{2}\sum_{j\in N_i}(\|x_i^{(i)}-x_i^{(j)}\|^2+\|x_j^{(j)}-x_j^{(i)}\|^2)]. \end{aligned} \tag{6.3}$$

If $Ax=b$ has solutions, (6.2) is equivalent to

$$\min\ f_p. \tag{6.4}$$

Proof. If $Ax=b$ has solutions, we can see that the solutions of (6.2) and (6.4) are those satisfying that $A_i^{(i)}x^{(i)}=b_i,\ \forall i\in V$ and $x_i^{(i)}=x_i^{(j)},\ x_j^{(j)}=x_j^{(i)},\ (i,j)\in E$. ■

Remark 28 *We can also put some weights in front of the consensus terms in (6.3) and obtain that*

$$f_p=\frac{1}{2}\sum_{i\in V}[\|A_i^{(i)}x^{(i)}-b_i\|^2+\frac{q_i}{2}\sum_{j\in N_i}(\|x_i^{(i)}-x_i^{(j)}\|^2+\|x_j^{(j)}-x_j^{(i)}\|^2)],$$

where $q_i>0$ are the weights. This might be helpful to accelerate the convergence, but it does not give rise to any more difficulty in analysis. So in the rest part, we still stick to (6.3).

6.3 Main Results

In this part, we propose the communication-efficient distributed algorithm to solve $Ax=b$ based on gradient descent method with constant step size.

We apply gradient descent method with constant step size α to (6.4). Let $\boldsymbol{x}=$

$$\begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(\bar{m})} \end{pmatrix} \text{ and } \nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_p}{\partial x^{(1)}} \\ \frac{\partial f_p}{\partial x^{(2)}} \\ \vdots \\ \frac{\partial f_p}{\partial x^{(\bar{m})}} \end{pmatrix}, \text{ then we have that}$$

$$\boldsymbol{x}(k+1) = \boldsymbol{x}(k) - \alpha \nabla f(\boldsymbol{x}).$$

For the gradient of f, we have that

$$\frac{\partial f_p}{\partial x^{(i)}} = A_i^{(i)T}(A_i^{(i)}x^{(i)} - b_i) + \sum_{j \in N_i} [(x_i^{(i)} - x_i^{(j)}) \otimes e^{(i)}_{k_i^{(i)}} + (x_j^{(i)} - x_j^{(j)}) \otimes e^{(i)}_{k_j^{(i)}}], \tag{6.5}$$

where $e^{(i)}_{k_j^{(i)}} = \begin{cases} 1, & k_j^{(i)} \text{ is the local index of agent } j \text{ by agent } i, \\ 0, & \text{else} \end{cases}$, and $\dim(e^{(i)}_{k_p^{(i)}}) = \sum_{j \in N_i} \dim(x_j^{(i)})$.

We obtain that for agent i,

$$\begin{aligned} x^{(i)}(k+1) =& x^{(i)}(k) - \alpha A_i^{(i)T}(A_i^{(i)}x^{(i)}(k) - b_i) \\ & - \alpha\{\sum_{j \in N_i} [(x_i^{(i)}(k) - x_i^{(j)}(k)) \otimes e^{(i)}_{k_i^{(i)}} + (x_j^{(i)}(k) - x_j^{(j)}(k)) \otimes e^{(i)}_{k_j^{(i)}}]\}. \end{aligned} \tag{6.6}$$

Remark 29 *The communications from agent j to agent i are only the estimate of states of agent i, e.g. $x_i^{(j)}$, and agent j, $x_j^{(j)}$, by agent j rather than the states of all agents, so (6.6) reduces a lot of communications compared to the algorithm in Chapter 5. Also, compared with [28], (6.6) does not require sharing estimates of the states of their common neighbors. As a result, (6.6) needs less communication between agents than [28].*

The convergence of (6.6) is shown below.

Theorem 8 *If $Ax = b$ has solutions, (6.6) converges to a solution of $Ax = b$ in finite time or at a linear rate provided that $0 < \alpha < \frac{2}{\lambda_{\max}(\nabla^2 f_p)}$, where f_p is defined in (6.4) and $\nabla^2 f_p$ is the Hessian of f_p.*

Before proving Theorem 8 when $Ax = b$ has multiple solutions, we need the following lemmas as preparations.

Lemma 32 *Let M positive semi-definite and thus symmetric. Let $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ be the orthogonal matrix and $\Lambda = \begin{pmatrix} \Lambda_1 & \\ & 0 \end{pmatrix}$ be the diagonal matrix such that $M = U\Lambda U^T = U_1\Lambda_1 U_1^T$, where Λ_1 is a positive definite diagonal matrix. Let $y = U^T x$, then the following statements are equivalent:*

1) $Mx = 0$,

2) $U_1\Lambda_1 U_1^T x = 0$,

3) $\Lambda y = 0$.

Proof. As $M = U_1\Lambda_1 U_1^T$, *1)* and *2)* are equivalent.

$Mx = U\Lambda U^T x = U\Lambda y$. As U is orthogonal and thus invertible, $Mx = 0$ is equivalent to $\Lambda y = 0$ and thus *1)* and *3)* are equivalent.

As a result, *1)*, *2)*, and *3)* are equivalent. ■

Lemma 33 *Let v be the projection of x_0 onto the hyper-plane $\{x|Cx = d\}$ and C be row full rank. Then $v = x_0 - C^T(CC^T)^{-1}(Cx_0 - d)$, and $\|v - x_0\|_2^2 = (Cx_0 - d)^T(CC^T)^{-1}(Cx_0 - d)$.*

Proof. It is easy to see that $v = \arg\min_{\{x:Cx=d\}} \frac{1}{2}\|x - x_0\|^2$. Then we consider its Lagrangian: $L = \frac{1}{2}\|x - x_0\|^2 + \lambda^T(Cx - d)$. Then we have that v is the solution of

$$\frac{\partial L}{\partial x} = x - x_0 + C^T\lambda = 0,$$
$$\frac{\partial L}{\partial x} = Cx - d = 0.$$

We have that $x = x_0 - C^T(CC^T)^{-1}(Cx_0 - d), \lambda = (CC^T)^{-1}(Cx_0 - d)$.

As a result,

$$v = x = x_0 - C^T(CC^T)^{-1}(Cx_0 - d)$$

and

$$\begin{aligned}\|v - x_0\|_2^2 &= (v - x_0)^T(v - x_0) = (Cx_0 - d)^T(CC^T)^{-1}(CC^T)(CC^T)^{-1}(Cx_0 - d) \\ &= (Cx_0 - d)^T(CC^T)^{-1}(Cx_0 - d).\end{aligned}$$

■

Then we have the follows lemma.

Lemma 34 *Let M be a nonzero positive semi-definite matrix and $f(x) = \frac{1}{2}x^TMx$. Let $X^\star = \arg\min f(x) = \{x^\star : Mx^\star = 0\}$ be the optimal set of $f(x)$. Then the gradient descent method $x(k+1) = x(k) - \alpha\nabla f(x(k)) = x(k) - \alpha Mx(k)$ converges to $X^\star$ in finite time or at a linear rate provided that $0 < \alpha < \frac{2}{\lambda_{\max}(M)}$, where $\lambda_{\max}(M)$ is the maximum eigenvalue of M.*

Proof. Denote $x^\star(k) = \arg\min_{x^\star \in X^\star} \|x(k) - x^\star\|_2$ as the closest optimal point to $x(k)$. As $X^\star = \{x^\star : Mx^\star = 0\} = \{x^\star : U_1\Lambda_1U_1^Tx^\star = 0\}$, where U_1 is defined in Lemma 32, we can see that from Lemmas 32 and 33

$$\begin{aligned}\|x(k) - x^\star(k)\|_2^2 &= (x(k) - x^\star(k))^TU_1\Lambda_1U_1^T(U_1\Lambda_1^2U_1^T)^{-1}U_1\Lambda_1U_1^T(x(k) - x^\star(k)) \\ &= (x(k) - x^\star(k))^TU_1U_1^T(x(k) - x^\star(k)).\end{aligned} \tag{6.7}$$

Then we have that

$$\begin{aligned}\|x(k+1) - x^\star(k+1)\|_2^2 &\le \|x(k+1) - x^\star(k)\|_2^2 \\ &= \|x(k) - \alpha Mx(k) - x^\star(k)\|_2^2 \\ &= \|x(k) - x^\star(k)\|_2^2 + \alpha^2x(k)^TM^TMx(k) - 2(x(k) - x^\star(k))^TMx(k).\end{aligned}$$

As $Mx^\star(k) = 0$,

$$x(k)^T M^T M x(k) = (x(k) - x^\star(k))^T M^T M (x(k) - x^\star(k))$$

and

$$(x(k) - x^\star(k))^T M x(k) = (x(k) - x^\star(k))^T M (x(k) - x^\star(k)).$$

So we have that

$$\begin{aligned}
&\|x(k+1) - x^\star(k+1)\|_2^2 \\
\leq &\|x(k) - x^\star(k)\|_2^2 + \alpha^2 (x(k) - x^\star(k))^T M^T M (x(k) - x^\star(k)) - 2(x(k) - x^\star(k))^T M (x(k) - x^\star(k)).
\end{aligned}$$

Let y and Λ be defined as in Lemma 32. Then we have that

$$\begin{aligned}
&\|x(k+1) - x^\star(k+1)\|_2^2 \\
\leq &\|x(k) - x^\star(k)\|_2^2 + \alpha^2 (y(k) - y^\star(k))^T \Lambda^2 (y(k) - y^\star(k)) - 2\alpha (y(k) - y^\star(k))^T \Lambda (y(k) - y^\star(k)) \\
= &\|x(k) - x^\star(k)\|_2^2 + (y(k) - y^\star(k))^T (\alpha^2 \Lambda^2 - 2\alpha\Lambda)(y(k) - y^\star(k)).
\end{aligned}$$

Next, we consider two cases.

1) If $\Lambda(y(k) - y^\star(k)) = 0$, then $M(x(k) - x^\star(k)) = 0$ from Lemma 32. We can then obtain that $Mx(k) = 0$ from the fact that $Mx^\star = 0$. In this case, $x(k) \in X^\star$ and (6.6) converges in finite time.

2) If $\Lambda(y(k) - y^\star(k)) \neq 0$, then $M(x(k) - x^\star(k)) \neq 0$ from Lemma 32. As

$$y(k) - y^\star(k) = U^T (x(k) - x^\star(k)) = \begin{pmatrix} U_1^T (x(k) - x^\star(k)) \\ \\ U_2^T (x(k) - x^\star(k)) \end{pmatrix},$$

we have that

$$(y(k) - y^\star(k))^T (\alpha^2 \Lambda^2 - 2\alpha\Lambda)(y(k) - y^\star(k)) = [U_1^T (x(k) - x^\star(k))]^T (\alpha^2 \Lambda_1^2 - 2\alpha\Lambda_1)[U_1^T (x(k) - x^\star(k))].$$

As $0 < \alpha < \frac{2}{\lambda_{\max}(M)} = \frac{2}{\lambda_{\max}(\Lambda_1)}$ and $\lambda_{\max}(\Lambda_1) > 0$, we have that $0 < \alpha\lambda_{\max}(\Lambda_1) < 2$ and thus $\alpha\lambda_{\max}(\Lambda_1))^2 - 2(\alpha\lambda_{\max}(\Lambda_1) < 0$. Similarly, we have that $\alpha\lambda_{\min}(\Lambda_1))^2 - 2(\alpha\lambda_{\min}(\Lambda_1) < 0$. Let $\lambda(\Lambda_1)$ be any eigenvalue of Λ_1. We can obtain that

$$\begin{aligned}\alpha^2\lambda(\Lambda_1)^2 - 2\alpha\lambda(\Lambda_1) &= (\alpha\lambda(\Lambda_1))^2 - 2(\alpha\lambda(\Lambda_1)) \\ &\leq \max\{(\alpha\lambda_{\min}(\Lambda_1))^2 - 2(\alpha\lambda_{\min}(\Lambda_1)), (\alpha\lambda_{\max}(\Lambda_1))^2 - 2(\alpha\lambda_{\max}(\Lambda_1))\} \\ &< 0.\end{aligned}$$

Denote

$$\bar{\lambda} = \max\{(\alpha\lambda_{\min}(\Lambda_1))^2 - 2(\alpha\lambda_{\min}(\Lambda_1)), (\alpha\lambda_{\max}(\Lambda_1))^2 - 2(\alpha\lambda_{\max}(\Lambda_1))\}.$$

Then we have that

$$\begin{aligned}(y(k) - y^\star(k))^T(\alpha^2\Lambda^2 - 2\alpha\Lambda)(y(k) - y^\star(k)) &\leq \bar{\lambda}[U_1^T(x(k) - x^\star(k))]^T[U_1^T(x(k) - x^\star(k))] \\ &= \bar{\lambda}\|x(k) - x^\star(k)\|_2^2,\end{aligned}$$

where the last equality results from (6.7).

Thus,

$$\|x(k+1) - x^\star(k+1)\|_2^2 \leq |x(k) - x^\star(k)\|_2^2 + \bar{\lambda}\|x(k) - x^\star(k)\|_2^2 \leq [1 + \bar{\lambda}]\|x(k) - x^\star(k)\|_2^2.$$

Notice that $\bar{\lambda} < 0$, $\{x(k)\}$ converges to $X^\star$ at a linear rate.

As a result, (6.6) converges to $X^\star$ in finite time or at a linear rate. ■

Proof of Theorem 8 : Notice that f_p in (6.4) is quadratic. When $Ax = b$ has a unique solution, f_p is strongly convex and the convergence at a linear rate of (6.6) is a direct result of Theorem 2.1.14 in [37]. When $Ax = b$ has multiple solutions, Theorem 8 is a direct result of Lemma 34.

Remark 30 (Distributed Selections of Step Sizes) *In (6.6), different agents use a common step size α. But in many applications, consensus on the step size might be difficult for the agent network. So in this part, we will provide a way to select the step size in a distributed way.*

Note that under the assumptions of Lemma 30, (6.2) is also equivalent to

$$\min f_{p,\beta} = \frac{1}{2}\sum_{i\in V}\beta_i[\|A_i^{(i)}x^{(i)} - b_i\|_2^2 + \frac{1}{2}\sum_{j\in N_i}(\|x_i^{(i)} - x_i^{(j)}\|_2^2 + \|x_j^{(j)} - x_j^{(i)}\|_2^2)], \tag{6.8}$$

if $\beta_i > 0,\ \forall i \in V$. If we apply the gradient method with constant step size α to (6.8), we can see that by choosing their own β_i, agents can select $\alpha\beta_i$ in a distributed way.

6.4 Simulations

In this part, we provide a simulation example to show the effectiveness of the distributed algorithm in (6.6).

The example is from the Newton-Raphson method to solve the power flow problem for the IEEE 13 Node Test Feeder. The power flow problem is nonlinear itself, but in every step of the Newton-Raphson method, we need to solve a system of linear equations $-J\begin{pmatrix}\Delta V_R\\ \Delta V_I\end{pmatrix} = \begin{pmatrix}\Delta P\\ \Delta Q\end{pmatrix}$, where $J = \begin{pmatrix}J_{11} & J_{12}\\ J_{21} & J_{22}\end{pmatrix}$ is the Jocobian of the power flow equations with respect to V_R and V_I, V_R is the real part of the voltages, V_I is the imaginary part of the voltages, and ΔP and ΔQ are the mismatches between the calculated and specified active and reactive powers.

For this example, we suppose the communication graph shares the topology with physical system. All four blocks $J_{11}, J_{12}, J_{21}, J_{22}$ of the Jacobian have Laplacian sparsity

structure. So the Jacobian J is more complex than defined in Definition 11. However, we can still apply (6.6) to solve it in a distributed way because the four blocks have Laplacian sparsity structure. We only carry out the simulation to solve the system of linear equations appeared in the first iteration in Newton Raphson method.

The simulation result is shown in Fig. 6.1. It shows the distance between $\begin{pmatrix} x_1^{(1)} \\ x_2^{(2)} \\ \vdots \\ x_{12}^{(12)} \end{pmatrix}$ and the exact solution. From Fig. 6.1, we can see that the agents' estimates finally converges to the accurate solution of $-Jx = b$.

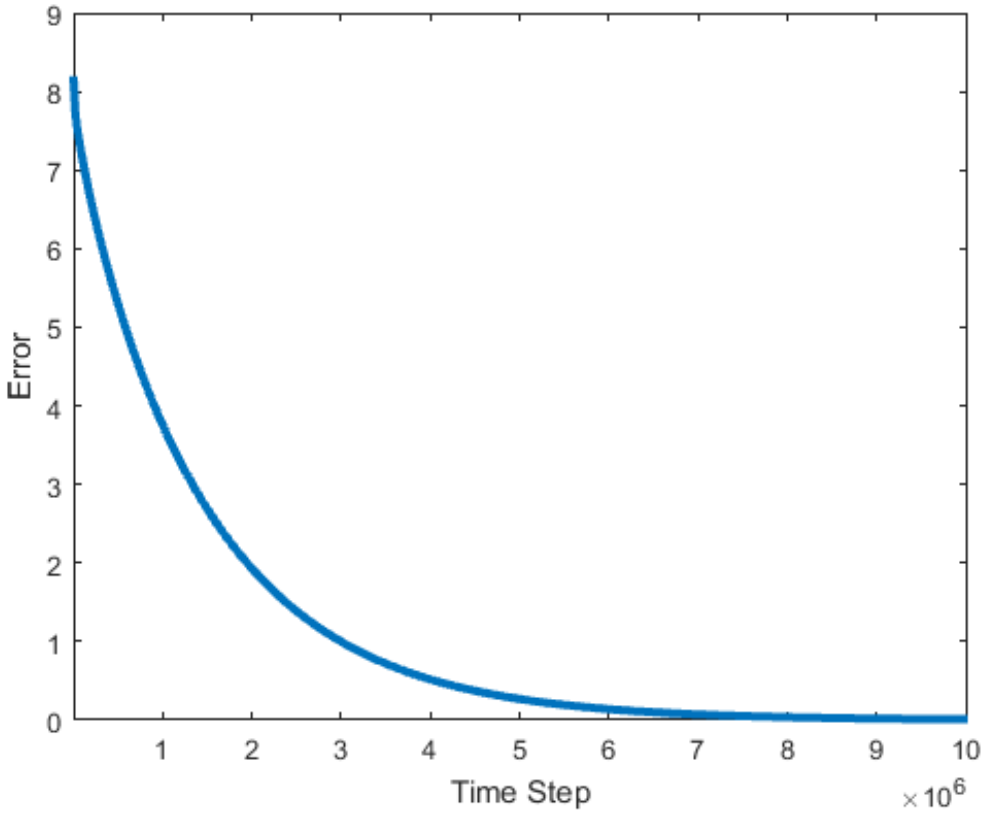

Figure 6.1: Norms of Errors between $x_i^{(i)}(k)$ and Accurate Solutions

Chapter 7

Conclusions and Future Works

7.1 Conclusions

We relaxed the condition of step sizes by removing the square summable requirement for the distributed subgradient algorithm and showed that the positive, vanishing and non-summable step sizes were sufficient for the convergence of the distributed subgradient algorithm to a minimizer of the global objective function when the topologies were balanced. We also showed that the fastest convergence rate is consistent with the centralized case when the step sizes are selected in a special form that is non-square summable in the unconstrained case.

We then proposed a discrete-time distributed algorithm to find the solution with minimum weighted norm associated with the weighted inner product of a system of solvable linear equations $Ax = b$. It was shown that if the agents started from the minimum weighted norm solution of their local linear equations and updated their estimates of the solution by the proposed algorithm, they would finally converge to the minimum weighted norm solution

of the global linear equations. It was also proved that if there were bounded initialization errors, the agents would converge to some solution of $Ax = b$ in a neighborhood of the global minimum weighted norm solution bounded by the initialization errors.

A discrete-time distributed algorithm was also proposed to solve a system of linear equations $Ax = b$. It was proved that the algorithm converged to a solution of $Ax = b$ at a geometric rate from arbitrary initializations when $Ax = b$ has either unique or multiple solutions. The common limit point of different agents was determined by the initializations, communication topologies, and the minimum 2-norm solution of $Ax = b$. The upper bounds of the convergence rate for two special cases were also derived related to the parameter γ_i and condition number of A.

A communication efficient distribution algorithm to solve $Ax = b$ with Laplacian sparsity structure was proposed. Only the states of two connected agents was shared connected by a communication link, which reduces the communication burden of all communication links. The algorithm was proved to converge at a linear rate or in finite time when $Ax = b$ has solutions.

7.2 Future Works

There are some future works we can push for. First, distributed algorithms to find the minimum 1-norm solution of a system of linear equations $Ax = b$ is an interesting topic as such a solution can be found in many applications such as compressed sensing and signal and image processing. We are also interested in accelerating the distributed algorithms to solve $Ax = b$ with dense or sparse matrix especially when A is ill-conditioned,

i.e., with large condition numbers. Then, we may work on distributed solutions to linear programming problems by solving linear equations and linear inequalities in a distributed way. Beyond these topics, the influence of time delay, noise, and asynchronous update is appealing topic worth researching, which may occur in real applications. Last but not the least, the combination of theoretic results and different application domains, e.g., power systems, transportation networks, and manufacture networks, are attractive.

Bibliography

[1] B. D. O. Anderson, S. Mou, A. S. Morse, and U. Helmke. Decentralized gradient algorithm for solution of a linear equation. *ArXiv e-prints*, September 2015.

[2] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[4] M. Cao, A. S. Morse, and B. D. O. Anderson. Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM Journal on Control and Optimization*, 47(2):575–600, 2008.

[5] T.-H. Chang, A. Nedic, and A. Scaglione. Distributed constrained optimization by consensus-based primal-dual perturbation method. *Automatic Control, IEEE Transactions on*, 59(6):1524–1538, June 2014.

[6] A.A. D'Amico, L. Sanguinetti, and D.P. Palomar. Convex separable problems with linear constraints in signal processing and communications. *Signal Processing, IEEE Transactions on*, 62(22):6045–6058, Nov 2014.

[7] L. Elsner, I. Koltracht, and M. Neumann. Convergence of sequential and asynchronous nonlinear paracontractions. *Numerische Mathematik*, 62(1):305–319, 1992.

[8] D. Fullmer, J. Liu, and A. Stephen Morse. An asynchronous distributed algorithm for computing a common fixed point of a family of paracontractions. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 2620–2625, Dec 2016.

[9] D. Fullmer, L. Wang, and A. S. Morse. A distributed algorithm for computing a common fixed point of a family of paracontractions. *IFAC-PapersOnLine*, 49(18):552 – 557, 2016. 10th IFAC Symposium on Nonlinear Control Systems NOLCOS 2016.

[10] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of Theoretical Biology*, 29(3):471 – 481, 1970.

[11] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. Cambridge Books Online.

[12] D. Jakovetic, J. Xavier, and J.M.F. Moura. Fast distributed gradient methods. *Automatic Control, IEEE Transactions on*, 59(5):1131–1146, May 2014.

[13] S. Kim and H. Ahn. Convergence of a generalized subgradient method for nondifferentiable convex optimization. *Math. Program.*, 50(1):75–80, February 1991.

[14] C. K. Koc, A. Guvenc, and B. Bakkalo[Ggrave]Lu. Exact solution of linear equations on distributed-memory multiprocessors. *Parallel Algorithms and Applications*, 3(1-2):135–143, 1994.

[15] A. I. Kostrikin. *Introduction to Algebra. Part II*. Fizmatlit, Moscow, 2000.

[16] J. Kuang. *Applied Inequalities*. Shandong Science and Technology Press, 2004.

[17] S. Lee and A. Nedic. Distributed random projection algorithm for convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):221–229, April 2013.

[18] P. Lin, W. Ren, and Y. Song. Distributed multi-agent optimization subject to nonidentical constraints and communication delays. *Automatica*, 65:120 – 131, 2016.

[19] J. Liu, X. Chen, T. Basar, and A. Nedic. A continuous-time distributed algorithm for solving linear equations. In *2016 American Control Conference (ACC)*, pages 5551–5556, July 2016.

[20] J. Liu, S. Mou, and A. S. Morse. An asynchronous distributed algorithm for solving a linear algebraic equation. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 5409–5414, December 2013.

[21] S. Liu, Z. Qiu, and L. Xie. Continuous-time distributed convex optimization with set constraints. In *IFAC World Congress*, volume 19, pages 9762–9767, 2014.

[22] I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Math. Program.*, 129(2):255–284, October 2011.

[23] J. Lu and C. Y. Tang. Distributed asynchronous algorithms for solving positive definite linear equations over networks-part i: Agent networks. In *Estimation and Control of Networked Systems*, volume 1, pages 252–257, 2009.

[24] J. Lu and C. Y. Tang. Distributed asynchronous algorithms for solving positive definite linear equations over networks-part ii: Wireless networks. In *Estimation and Control of Networked Systems*, volume 1, pages 258–263, 2009.

[25] J. Lu and Choon Y. Tang. Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case. *Automatic Control, IEEE Transactions on*, 57(9):2348–2354, Sept 2012.

[26] A. Margaris, S. Souravlas, and M. Roumeliotis. Parallel Implementations of the Jacobi Linear Algebraic Systems Solve. *ArXiv e-prints*, March 2014.

[27] J.F.C. Mota, J.M.F. Xavier, P.M.Q. Aguiar, and M. Puschel. Distributed optimization with local domains: Applications in mpc and network flows. *Automatic Control, IEEE Transactions on*, 60(7):2004–2009, July 2015.

[28] S. Mou, Z. Lin, L. Wang, D. Fullmer, and A.S. Morse. A distributed algorithm for efficiently solving linear equations and its applications (special issue jcw). *Systems & Control Letters*, 91:21 – 27, 2016.

[29] S. Mou, J. Liu, and A. S. Morse. A distributed algorithm for solving a linear algebraic equation. *IEEE Transactions on Automatic Control*, 60(11):2863–2878, November 2015.

[30] S. Mou, J. Liu, and A. S. Morse. A distributed algorithm for solving a linear algebraic equation. *IEEE Transactions on Automatic Control*, 60(11):2863–2878, Nov 2015.

[31] S. Mou and A. S. Morse. A fixed-neighbor, distributed algorithm for solving a linear algebraic equation. In *Control Conference (ECC), 2013 European*, pages 2269–2273, July 2013.

[32] A. Nedic and A. Olshevsky. Distributed optimization over time-varying directed graphs. *Automatic Control, IEEE Transactions on*, 60(3):601–615, March 2015.

[33] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, 54(1):48–61, Jan 2009.

[34] A. Nedic, A. Ozdaglar, and P.A. Parrilo. Constrained consensus and optimization in multi-agent networks. *Automatic Control, IEEE Transactions on*, 55(4):922–938, April 2010.

[35] A. Nedich, A. Olshevsky, and W. Shi. Achieving Geometric Convergence for Distributed Optimization over Time-Varying Graphs. *ArXiv e-prints*, July 2016.

[36] A. Nemirovski. Information-based complexity of convex programming. `http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf`, 1994.

[37] I.U.E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Mathematics and its applications. Kluwer Academic Publishers, 2004.

[38] F. Pasqualetti, R. Carli, A. Bicchi, and F. Bullo. Distributed estimation and detection under local information. *IFAC Proceedings Volumes*, 43(19):263 – 268, 2010. 2nd IFAC Workshop on Distributed Estimation and Control in Networked Systems.

[39] F. Pasqualetti, R. Carli, and F. Bullo. A distributed method for state estimation and false data detection in power networks. In *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 469–474, Oct 2011.

[40] F. Pasqualetti, R. Carli, and F. Bullo. Distributed estimation via iterative projections with application to power network monitoring. *Automatica*, 48(5):747 – 758, 2012.

[41] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, PP(99):1–1, 2017.

[42] R.T. Rockafellar. *Convex Analysis*. Princeton landmarks in mathematics and physics. Princeton University Press, 1997.

[43] G. Shi, B. D. O. Anderson, and U. Helmke. Network Flows that Solve Linear Equations. *ArXiv e-prints*, October 2015.

[44] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization. *ArXiv e-prints*, April 2014.

[45] W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[46] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayǹski. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag New York, Inc., New York, NY, USA, 1985.

[47] A.R. Tarrida. *Affine Maps, Euclidean Motions and Quadrics*. Springer Undergraduate Mathematics Series. Springer London, 2011.

[48] V. V. Vasin and I. I. Eremin. *Operators and iterative processes of Fejér type: theory and applications*, volume 53. Walter de Gruyter, 2009.

[49] L. Wang, D. Fullmer, and A. S. Morse. A distributed algorithm with an arbitrary initialization for solving a linear algebraic equation. In *2016 American Control Conference (ACC)*, pages 1078–1081, July 2016.

[50] P. Wang, W. Ren, and Z. Duan. Distributed solution to linear equations from arbitrary initialization. In *2017 American Control Conference*, May 2017.

[51] E. Wei, A. Ozdaglar, and A. Jadbabaie. A distributed newton method for network utility maximization i: Algorithm. *Automatic Control, IEEE Transactions on*, 58(9):2162–2175, Sept 2013.

[52] M. Yang and C. Y. Tang. A distributed algorithm for solving general linear equations over networks. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 3580–3585, Dec 2015.

[53] K. You, S. Song, and R. Tempo. A networked parallel algorithm for solving linear algebraic equations. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1727–1732, Dec 2016.

[54] D. Young. Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society*, 76(1):92–111, 1954.

[55] M. Zargham, A. Ribeiro, A. Ozdaglar, and A. Jadbabaie. Accelerated dual descent for network flow optimization. *Automatic Control, IEEE Transactions on*, 59(4):905–920, April 2014.

[56] M. Zhu and S. Martinez. On distributed convex optimization under inequality and equality constraints. *Automatic Control, IEEE Transactions on*, 57(1):151–164, Jan 2012.

[57] M. Zhu and S. Martinez. An approximate dual subgradient algorithm for multi-agent non-convex optimization. *Automatic Control, IEEE Transactions on*, 58(6):1534–1539, June 2013.