Ph.D. Dissertation

Convergence Results for the Decentralized Gradient Descent and the Gradient-push with Event-triggered Communication

Jimyeong Kim

The Graduate School
Sungkyunkwan University
Department of Mathematics

Contents

List of Tables

List of Figures

Notation

The notations in this section are, mostly, standard throughout the thesis, unless specifically modified in the appropriate chapter or obvious from the context.

- I, identity matirx.
- $\mathbf{1} = [1, 1, \cdots, 1]^T$, the column vector all ones whose size depends on the context.
- X_*, the optimal set of a given problem.
- W, the mixing matrix
- β, the spectral norm of the matrix $W - \frac{1}{m}\mathbf{1}\mathbf{1}^T$
- x_*, an optimizer of a given problem, i.e. $x_* \in X_*$.
- For a matrix $A \in \mathbb{R}^{n\times m}$, a_{ij} denotes the (i, j)-th entry of A
- For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ represent the smallest and the largest eigenvalues of A, respectively.
- For a row vector $x \in \mathbb{R}^{1\times d}$, $\mathbf{x} = [x^T; x^T; \cdots; x^T]$ whose size of the column depends on the context.
- $\lfloor x \rfloor$, the largest integer less or equal to x.
- e, the exponential number.
- $\|\cdot\|$, the standard l^2 norm.

Abstract

Convergence Results for the Decentralized Gradient Descent and the Gradient-Push with Event-Triggered Communication

Distributed optimization has received a lot of interest in recent years due to its wide applications in various fields. Decentralized optimization, in particular, stands out for its efficiency and practicality as it operates without a central coordinating entity. This thesis explores convergence results for decentralized gradient descent when a communication pattern is undirected and the Gradient-Push algorithm when a communication pattern is directed.

The decentralized gradient descent, introduced by Nedic and Ozdaglar in [A. Nedic and A. Ozdaglar, IEEE Trans. Automat. Control, 54 (2009), pp. 48–61], has gained considerable prominence owing to its simplicity and effectiveness when a communication pattern is undirect. Much research has focused on analyzing the convergence rate of the decentralized gradient descent with a constant step size, revealing an exponential convergence rate. However, it's important to note that this convergence result indicates that the decentralized gradient descent algorithm converges to a point that is within $O(\alpha)$ of an optimal point, but not necessarily to the optimal point itself. In this thesis, we revisit the convergence analysis of the decentralized gradient descent algorithm. Specifically, we explore the impact of using a diminishing step size and provide exact convergence results for cases where the total cost function f is strongly convex, even when the local cost functions f_i are not necessarily convex.

We extend our convergence results from unconstrained problems to constrained ones, considering the decentralized projected gradient descent with both constant and diminishing step sizes. Specifically, for a constant step size, we demonstrate that decentralized projected gradient descent converges within $O(\sqrt{\alpha})$-neighborhoods of an optimal point. Additionally, we provide one-dimensional and half-space examples that attain $O(\alpha)$-convergence results as in the unconstrained case.

Lastly, we shift our focus to a directed communication scenario and explore the Gradient Push (GP) algorithm, as introduced by Nedic and Olshevsky in [A. Nedic and A. Olshevsky, IEEE Trans. Automat. Control, 60 (2014), pp. 601-615]. Specifically, we investigate the application of event-triggered communication to the GP algorithm, referred to as GP with event-triggered communication, and attain asymptotic convergence results. To the best of our knowledge, this marks the first instance of obtaining convergence results in a directed communication scenario.

Keyword: convex optimization, decentralized optimization, decentralized gradient descent, gradient-push, event-triggered communciation

Chapter 1

Introduction

Distributed optimization involves minimizing a total cost function, which is composed of the sum of individual local cost functions, each associated with a local agent. In recent years, there has been a growing interest in distributed optimization techniques due to their significant role in solving engineering problems across various domains, including distributed control [7, 9], signal processing [5, 24], and machine learning [4, 19]

A crucial aspect that distinguishes distributed optimization is the choice between a centralized or decentralized approach, with roughly two types of distributed optimization to consider In a centralized optimization framework, a central coordinator assumes the responsibility of gathering information from all participating nodes, solving the optimization problem, and distributing the optimal solution back to the nodes. This centralized approach offers the advantage of a global perspective on the optimization problem.

However, the centralized approach also introduces challenges, particularly in terms of communication. With all nodes required to exchange information with the central coordinator, communication overhead can become a bottleneck, especially when dealing with numerous nodes or networks with limited bandwidth or high latency. Moreover, the central coordinator becomes a single point of failure, as the entire system's performance relies on its proper functioning.

To overcome the limitations associated with centralized optimization, decentralized optimization approaches have garnered significant attention. In decentralized optimization, the optimization problem is distributed across multiple nodes, with each node independently solving its own local sub-problem. The nodes then communicate and exchange information with their neighboring nodes to iteratively refine their solutions. This decentralized approach mitigates communication overhead and eliminates the necessity for a central coordinator, thereby enhancing scalability and fault tolerance.

In this thesis, our focus is on decentralized optimization scenarios where each agent possesses its own local cost function and collaboratively aims to find a minimizer of the sum of the local cost functions without relying on a central controller. This problem can be formulated as follows:

$$\min_{x\in\Omega} f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{1.0.1}$$

where each local cost $f_i : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ is a differentiable function only known to agent i. There are various decentralized algorithms containing the decentralized gradient descent [33, 36, 10, 26, 49, 25], the dual averaging method [17], consensus-based dual decomposition [18, 43], and alternating direction method of multipliers (ADMM) based algorithms [28, 41]. We also refer [47, 40, 34] for a variety of the decentralized optimization algorithms for an undirected and [31, 11] for a direct communication graph.

Although the decentralized concept plays an effective role in terms of communication, these distributed algorithms require each agent to communicate with their neighbors at every iteration. This communication overhead can be burdensome in restricted environments. Additionally, the work mentioned in [42] highlights that power consumption due to communication may surpass that of computation for control inputs or optimization algorithms. To reduce the cost of communication, the event-triggering approach has appeared as a promising paradigm to reduce the communication load in distributed systems. In the distributed detection problem over sensor network [2, 1], each sensor censors its local data and sends the updated data to the fusion center only when the data is informative. For distributed control problems, agents send their coordinate information only when a triggering condition is satisfied [29, 15].

In this thesis, our focus is specifically on the decentralized gradient descent [33, 31] and its convergence properties. We aim to determine the exact convergence rate for the decentralized gradient descent and discuss the optimality of the condition for step size and convergence rate using simple examples. Furthermore, we present the first results of the convergence rate for a gradient-push algorithm with event-triggered communication.

Previous Works

Nedić and Ozdaglar [33] showed that for the decentralized gradient descent with the step size $\alpha(t) \equiv \alpha$, the cost value $f(\cdot)$ at an average of the iterations converges to an $O(\alpha)$-neighborhood of an optimal value of f. Ram et al. [36] proved that the decentralized gradient descent, involving a projection to a compact set, converges to an optimal point if the step size satisfies $\sum_{t=1}^{\infty} \alpha(t) = \infty$ and $\sum_{t=1}^{\infty} \alpha(t)^2 < \infty$. In the work of Chen [10], the decentralized projected gradient descent algorithm with step size $\alpha(t) = c/t^p$ with $0 < p < 1$ was considered and the convergence rate was achieved as $O(1/t^p)$ for $0 < p < 1/2$, $O(\log t/\sqrt{t})$ for $p = 1/2$, and $O(1/t^{1-p})$ for $1/2 < p < 1$. The work of Liu et al. [26] obtained the convergence rate $O(1/\sqrt{t})$ when the step size is chosen as $\alpha(t) = c/t$ for a suitable range of $c > 0$. We mention that all the aforementioned works were established with the gradient bound assumption $\|\nabla f_i\|_\infty < \infty$ and convexity assumption on each function f_i.

Recently, Yuan et al. [49] considered the decentralized gradient descent algorithm without the gradient bound assumption. They showed that if each local cost function is convex and the total cost is strongly convex, then the decentralized gradient descent algorithm with constant step size $\alpha(t) \equiv \alpha$ converges exponentially to an $O(\alpha)$-neighborhood of an optimizer x_*. Liu et al. [25] showed that when the uniform boundedness of the gradient assumption is replaced by the L-smoothness, the projected decentralized gradient descent algorithm with constant step size $\alpha(t) \equiv \alpha$ converges exponentially to an $O(\alpha)$-neighborhood of an optimizer x_* The previous results we mentioned are summarized in Table 1.1.

For the event-triggered communication, recent works [21, 27, 30, 45, 50] developed distributed optimization algorithms with event-triggered communication to overcome the communication overhead of distributed systems. Lu-Li [27] designed the distributed gradient descent with event-triggered communication for the distributed optimization on the whole space, and it was further studied in Li-Mu [23] to establish a convergence rate. For the distributed optimization on a bounded domain, Kajiyama et al. [21] designed the projected distributed gradient descent with event-triggered communication. Liu et al. [25] extended the work to the case with constant step-size. Cao-Basar [8] studied the online distributed problem using the distributed event-triggered gradient method. Xiong et al. [48] considered the distributed stochastic mirror descent with event-triggered communication. The distributed estimation problem was studied by He et al. [20] utilizing the event-triggered communication. In these algorithms, each agent sends its state only when the difference between the current state and the latest sent state is larger than a threshold to reduce possible unnecessary network utilization.

	Cost	Smooth	Learning rate	Rate	Proj.
[33]	C	GB	$O\left(\frac{1}{\alpha t}\right)+O(\alpha)$	No	
[36]	C	GB	$\sum \alpha(t)=\infty$ $\sum \alpha(t)^2<\infty$	$o(1)$	Yes
[10]	C	GB	$\alpha(t)=\frac{c}{t^p}$	$O(\frac{1}{t^p})$ if $p \in (0,\frac{1}{2})$ $O(\frac{\log t}{\sqrt{t}})$ if $p=\frac{1}{2}$ $O(\frac{1}{t^{1-p}})$ if $p \in (\frac{1}{2},1)$	Yes
[26]	SC	GB	$\alpha(t)=\frac{c}{t}$	$O(1/\sqrt{t})$	Yes
[49]	C	LS	$\alpha(t)\equiv\alpha$	$O(\frac{1}{\alpha t})+O(\frac{\alpha}{1-\beta})$	No
[49]	SC	LS	$\alpha(t)\equiv\alpha$	$O(e^{-ct})+O(\frac{\alpha}{1-\beta})$	No
[25]	SC	LS	$\alpha(t)=\alpha$	$O(e^{-ct}+\alpha+\frac{D}{\sqrt{n}})$	Yes
Thm 3.5.1 Thm 3.5.2	SC	LS	$\alpha(t)=\frac{a}{(t+w)^p}$	$O(t^{-p})$ if $p \in (0,1]$	No
Thm 4.6.1	SC	LS	$\alpha(t)=\alpha$	$O(e^{-ct}+\sqrt{\alpha})$	Yes
Thm 4.6.2 Thm 4.6.3	SC	LS	$\alpha(t)=\frac{a}{(t+w)^p}$	$O(t^{-p/2})$ if $p \in (0,1]$	Yes
Thm 4.7.2	1-dim example	LS	$\alpha(t)=\alpha$	$O(e^{-ct}+\alpha)$	Yes
Thm 4.7.10	Half-space example	LS	$\alpha(t)=\alpha$	$O(e^{-ct}+\alpha)$	Yes

Table 1.1: This table summarizes the convergence results for DGD. Here C (resp., SC) means that the total cost function is assumed to be convex (resp., strongly convex). Also x_* is an optimizer of (1.0.1) and $\tilde{x}_i(t)=\frac{1}{t}\sum_{s=0}^{t-1} x_i(s)$. We write 'No' in 'Proj.' if $\Omega=\mathbb{R}^d$

Contributions

In this thesis, we investigate the convergence property of the decentralized (projected) gradient descent algorithm for a general class of non-increasing step size $\{\alpha(t)\}_{t\in\mathbb{N}_0}$ given as $\alpha(t)=a/(t+w)^p$ for $a>0$, $w\geqslant 1$ and $0<p\leqslant 1$. Furthermore, we only assume strong convexity on the total cost function f, with cost functions f_i not necessarily being convex. The convergence results of this paper shed light on choosing suitable values of a and w for fast convergence of the decentralized (projected) gradient descent algorithm.

We revisit the decentralized projected gradient algorithm with a constant step size. We establish convergence results, which demonstrate that a sequence generated by the decentralized projected gradient algorithm converges exponentially fast to a neighbor within an $O(\sqrt{\alpha})$ radius of an optimal point. Additionally, we present one-dimensional and half-space examples that show that the sequence generated by decentralized projected gradient converges to an $O(\alpha)$ radius of an optimal point. This highlights the important observation that, even in constrained scenarios, these specific examples exhibit the same $O(\alpha)$ error as in the unconstrained case.

The consensus-based decentralized optimization algorithms with event-triggering communication mentioned earlier have been proposed for use with undirected graphs. However, these methods cannot be applied to a situation where the network of the agents is a directed graph. We remark that when the agents have different ranges of communication, the network of agents is usually given as a directed graph. In this thesis, we introduce the gradient-push algorithm with event-triggering communication, applicable to both directed and time-varying graphs, and provide convergence results for this algorithm.

Outline

- In Chapter 2, we introduce preliminary concepts essential for understanding the content of this thesis. Firstly, we delve into the fundamental concepts and properties of convexity, smoothness, and projection operators. Next, we explore key elements of graph theory, which are necessary for depicting the communication patterns among agents. Finally, we present some sequential estimates that will be frequently employed to establish convergence rates.

- In Chapter 3, we conduct a convergence analysis of decentralized gradient descent within the entire space, denoted as $\Omega = \mathbb{R}^d$. Specifically, we demonstrate that decentralized gradient descent achieves exact convergence to an optimal point when utilizing a diminishing step size without convexity for a local cost function. Furthermore, we present a straightforward example highlighting the criticality of the step size condition for obtaining a uniformly bounded sequence. To complement our findings, we include numerical experiments. This chapter is based on the work Choi-Kim [12].

- In Chapter 4, we study the decentralized projected gradient descent. We begin this chapter by addressing the challenges posed by a projection operator, which make it challenging to apply the methods discussed in Chapter 3 directly. To tackle these difficulties, we introduce a novel approach designed to overcome them and achieve convergence results for both constant and diminishing step sizes. Specifically, in constant cases, the DPG algorithm achieves $O(\sqrt{\alpha})$-convergence. In addition, We construct examples, namely one-dimensional and half-space examples, which achieve $O(\alpha)$-convergence. This chapter is based on the work Choi-Kim [13].

- In Chapter 5, we shift our focus to the gradient-push algorithm, which is used in directed communication patterns. In this chapter, we introduce event-triggered mechanisms and analyze the convergence of this algorithm in the context of event-triggered communication. As a result, we achieve asymptotic convergence under conditions of step size decay and summability, and we also derive an exact convergence rate for $\alpha(t) = 1/\sqrt{t}$. We present numerical experiments to demonstrate the effectiveness and verify the convergence results of the algorithm. This chapter is based on the work Kim-Choi [22].

Chapter 2

Preliminaries

This chapter reviews several important definitions, properties and concepts that will be used throughout the thesis.

2.1 Smoothness and Strong Convexity

We begin with the definition of convex set and convex function.

Definition 2.1.1 (convex set)**.** A set $\Omega \in \mathbb{R}^d$ is convex set if $\lambda x + (1-\lambda)y \in \Omega$ for all $x, y \in \Omega$ and $\lambda \in [0,1]$.

Definition 2.1.2 (convex function)**.** A function $f : \Omega \to \mathbb{R} \cup \{\infty\}$ is convex on Ω if Ω is a convex set and the following inequality holds:

$$f\left(\lambda x + (1-\lambda)y\right) \leqslant \lambda f(x) + (1-\lambda) f(y) \tag{2.1.1}$$

for all $x, y \in \Omega$ and $\lambda \in [0,1]$.

Instead of directly utilizing (2.1.1) in convex optimization, it is common to employ the following lemma as an alternative approach.

Lemma 2.1.3 ([37])**.** *Suppose that f is differentiable on Ω. Let f be a convex function over a given convex set Ω. Then*

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle \tag{2.1.2}$$

for all $x, y \in \Omega$.

Now we review the definition of L-smoothness.

Definition 2.1.4 (L-smoothness)**.** Let $L \geqslant 0$. A function $f : \Omega \to \mathbb{R} \cup \{\infty\}$ is said to be L-smooth over a set Ω if it is differentiable over Ω and satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\|. \tag{2.1.3}$$

The following descent lemma is extremely useful thoroughout the thesis.

Lemma 2.1.5 ([3])**.** *Let $f : \Omega \to \mathbb{R} \cup \{\infty\}$ be an L-smooth function over a given convex set Ω. Then for any $x, y \in \Omega$,*

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2. \tag{2.1.4}$$

In Chapter 3 and 4, we will obtain the convergence rate for the decentralized gradient algorithm under the strong convexity assumption which is defined as follows.

Definition 2.1.6 (strong convexity)**.** A function $f : \Omega \to \mathbb{R} \cup \{\infty\}$ is said to be μ-strongly convex on Ω for a given $\mu > 0$ if Ω is convex set and satisfies

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2}\lambda(1 - \lambda)\|x - y\|^2. \tag{2.1.5}$$

Throughout the thesis, we use the terminology "strongly convex" and "μ-strongly convex" interchangeably. We also introduce the useful lemma for strongly convex function.

Lemma 2.1.7 ([3])**.** *Suppose that f is differentiable on Ω. Let $f : \Omega \to \mathbb{R} \cup \{\infty\}$ be a μ-strongly convex function over a given convex set Ω. Then for any $x, y \in \Omega$,*

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \tag{2.1.6}$$

The following theorem guarantee that a strongly convex function has a unique minimizer.

Theorem 2.1.8 ([3])**.** *Let $f : \Omega \to \mathbb{R} \cup \{\infty\}$ be a μ-strongly convex function. Then f has a unique minimizer.*

Lemma 2.1.9 ([6])**.** *Let $f : \Omega \to \mathbb{R} \cup \{\infty\}$ be a μ- strongly convex and L-smooth function over a given convex set Ω. Then for all $x, y \in \mathbb{R}^n$,*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geqslant \frac{L\mu}{L + \mu}\|x - y\|^2 + \frac{1}{L + \mu}\|\nabla f(x) - \nabla f(y)\|^2 \tag{2.1.7}$$

2.2 Projection Operator

In this section, we study the projection operator $\mathcal{P}_\Omega$ defined by

$$\mathcal{P}_\Omega[x] = \arg\min_{y \in \Omega} \|y - x\|^2. \tag{2.2.1}$$

Definition 2.2.1 (proximal operator)**.** Given a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, the proximal operator of f is the operator given by

$$\operatorname{prox}_f(x) = \arg\min_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}, \tag{2.2.2}$$

for any $x \in \Omega$.

The proximal operator has the following property.

Theorem 2.2.2 ([3])**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function. Then for any $x, y \in \mathbb{R}^d$, we have*

$$\|\operatorname{prox}_f(x) - \operatorname{prox}_f(y)\| \leqslant \|x - y\|. \tag{2.2.3}$$

Consider the indicator function $\delta_\Omega(x)$, which is defined by

$$\delta_\Omega(x) = \begin{cases} 0, & x \in \Omega, \\ \infty, & x \notin \Omega. \end{cases} \tag{2.2.4}$$

Then the projection operator can be viewed as the proximal operator of the indicator function, i.e.

$$\operatorname{prox}_{\delta_\Omega}(x) = \arg\min_{y \in \mathbb{R}^d} \left\{ \delta_\Omega(u) + \frac{1}{2}\|y - x\|^2 \right\} = \arg\min_{y \in \Omega} \|y - x\|^2 = \mathcal{P}_\Omega(x). \tag{2.2.5}$$

Therefore we derive the following properties for the projection operator.

Lemma 2.2.3. *Let $\Omega \subset \mathbb{R}^d$ be convex and closed.*

1. *For any $x, y \in \mathbb{R}^d$, we have*

$$\|\mathcal{P}_\Omega[x] - \mathcal{P}_\Omega[y]\| \leqslant \|x - y\|. \tag{2.2.6}$$

2. *For any $x \in \mathbb{R}^d$ and $y \in \Omega$, we have*

$$\|\mathcal{P}_\Omega[x] - y\| \leqslant \|x - y\|. \tag{2.2.7}$$

Proof. Using (2.2.5) and Theorem 2.2.2, (2.2.6) is obtained directly. And Using $\mathcal{P}_\Omega[y] = y$ for $y \in \Omega$, we have (2.2.7). □

Lastly, we make the following lemma, which will be used in Chapter 4.

Lemma 2.2.4. *Let $\Omega \subset \mathbb{R}^d$ be convex and closed. Then, for any $x_1, \cdots, x_n \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^n \left\| \mathcal{P}_\Omega[x_i] - \frac{1}{n}\sum_{j=1}^n \mathcal{P}_\Omega[x_j] \right\|^2 \leqslant \sum_{i=1}^n \left\| x_i - \frac{1}{n}\sum_{j=1}^n x_j \right\|^2.$$

Proof. For given $\{a_i\}_{i=1}^n \subset \mathbb{R}^d$, consider the function $F : \mathbb{R}^d \to \mathbb{R}$ defined by

$$F(x) = \sum_{i=1}^{n} \|a_i - x\|^2,$$

This function is minimized when $x = \frac{1}{n}\sum_{i=1}^n a_i$, and using this we find

$$\sum_{i=1}^{n} \left\| \mathcal{P}_\Omega[x_i] - \frac{1}{n}\sum_{j=1}^{n} \mathcal{P}_\Omega[x_j] \right\|^2 \leqslant \sum_{i=1}^{n} \left\| \mathcal{P}_\Omega[x_i] - \mathcal{P}_\Omega\left[\frac{1}{n}\sum_{j=1}^{n} x_j\right] \right\|^2.$$

Combining this with (2.2.6), we get the desired inequality. □

2.3 Graph and the Mixing Matrix

In decentralized optimization, a local agent informs its own information to other agents relying on shared communication networks which are characterized by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each edge in $\mathcal{E}$ can be represented by $(i, j) \in \mathcal{E}$ means that i can send messages to j. In this thesis, we only consider graphs with finitely many vertices and connected which are defined as follows.

Definition 2.3.1 (connected graph)**.** A graph $\mathcal{G}$ is connected if any two vertices in the graph have a path between them.

We define in-neighbors and out-neighbors of node i, respectively, as $N_i^{\text{in}} = \{j|(j,i) \in \mathcal{E}\}$ and $N_i^{\text{out}} = \{j|(i,j) \in \mathcal{E}\}$. The adjacency matrix $A(\mathcal{G})$ of a graph $\mathcal{G}$ is 01-matrix, which is defined as $[A(\mathcal{G})]_{ij} = 1$ if $(i,j) \in \mathcal{E}$ and $[A(\mathcal{G})]_{ij} = 0$ otherwise. The $D(\mathcal{G})$ is a diagonal matrix with $[D(\mathcal{G})]_{ii} = |N_i^{\text{out}}|$. In the context of decentralized optimization, the mixing matrix can be associated with the adjacency matrix of a graph. Roughly, there are two types of graphs, undirected and directed. We first study the mixing matrix for an undirected graph.

Undirected Graph

In an undirected graph, each edge in $\mathcal{E}$ can be represented by unordered set $\{i, j\} \in \mathcal{E}$ which means that i can send messages to j and vice versa. W is a decentralized mixing matrix with respect to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $w_{ij} = 0$ when $i \neq j$ and $\{i.j\} \notin \mathcal{E}$. A mixing matrix $W \in \mathbb{R}^{m \times m}$ with respect to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ used in decentralized optimization often satisfies the following assumptions:

Assumption 2.3.2. The mixing matrix $W \in \mathbb{R}^{m\times m}$, with $m = |\mathcal{V}|$, satisfies

(A) $W = W^T$

(B) $\mathrm{Null}(I - W) = \mathrm{span}(1)$

(C) $\rho(W - (1/m)\mathbf{1}\mathbf{1}^T) < 1.$,

where $\rho(\cdot)$ denotes the spectral radius of a matrix.

One example of the mixing matrix for an undirected graph satisfying Assumption 2.3.2 is as follows.

Example 2.3.3 (Laplacian-based mixing matrix [46, 38]). The Laplacian $L(\mathcal{G})$ of a graph $\mathcal{G}$ is defined as $L(\mathcal{G}) = D(\mathcal{G}) - A(\mathcal{G})$. The Laplacian-based mixing matrix is defined as

$$W = I - \frac{1}{\tau} L(\mathcal{G}), \tag{2.3.1}$$

where τ is a constant satisfying $\tau > \frac{1}{2}\lambda_{\max}(L)$.

Decentralized optimization involves a crucial concept known as "consensus," where each agent ultimately reaches the same state. This can be expressed as:

$$\lim_{t\to\infty} x_i(t) = \lim_{t\to\infty} x_j(t), \text{ for all } i, j \in \mathcal{V}. \tag{2.3.2}$$

When (2.3.2) is satisfied, we said the consensus is achieved. The mixing matrix plays a vital role in facilitating consensus among the agents. To illustrate this, we consider the following dynamics for the mixing matrix W that satisfies Assumption 2.3.2.

$$\mathbf{x}(t+1) = W\mathbf{x}(t), \tag{2.3.3}$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_m(t)] \in \mathbb{R}^m$. Then we have the following consensus result.

Lemma 2.3.4. *[46] Suppose that the mixing matrix W with respect to an undirected and connected graph $\mathcal{G}$ satisfies Assumption 2.3.2. Then we have the following results.*

1. $\lim_{t\to\infty} W^t = \frac{1}{m}\mathbf{1}\mathbf{1}^T$

2. $\lim_{t\to\infty} \mathbf{x}(t) = (\bar{x}(0), \cdots, \bar{x}(0))$,

where $\bar{x}(0) = \sum_{i=1}^m x_i(0)$.

Lastly, we introduce the following lemma, that extremely useful through the thesis.

Lemma 2.3.5 ([35])**.** *Suppose that the mixing matrix W with respect to an undirected and connected graph $\mathcal{G}$ satisfies Assumption 2.3.2. Let β be the spectral norm of the matrix $W - \frac{1}{m}\mathbf{1}\mathbf{1}^T$. Then there exists a constant $\beta < 1$ such that*

$$\sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij}(x_j - \bar{x}) \right\|^2 \leqslant \beta^2 \sum_{i=1}^m \|x_i - \bar{x}\|^2, \tag{2.3.4}$$

where $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ *for any* $x_i \in \mathbb{R}^{d\times 1}$ *and* $1 \leqslant i \leqslant m$.

Directed Graph

In a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each edge in $\mathcal{E}$ can be represented by ordered set $(i, j) \in \mathcal{E}$ means that i can send messages to j. For a directed graph $\mathcal{G}$, we make the following assumption.

Assumption 2.3.6. Every vertex of $\mathcal{G}$ has a self-loop, i.e.,

$$(i, i) \in \mathcal{E}, (\Leftrightarrow i \in N_i^{\text{in}} \Leftrightarrow i \in N_i^{\text{out}}),\ \forall i \in \mathcal{V}$$

Suppose that a directed graph $\mathcal{G}$ satisfies Assumption 2.3.6. Then we consider the following mixing matrix W.

$$w_{ij} = \begin{cases} 1/|N_j^{\text{out}}| & \text{if } i \in N_j^{\text{out}}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.5}$$

Note that this mixing matrix is a column stochastic matrix, i.e. $\sum_{i=1}^{m} w_{ij} = 1$. Since this mixing matrix W defined by (2.3.5) satisfies Assumption 2.3.6, it follows by Perron-Frobenius theorem that W has a right eigenvector $\pi = [\pi_1, \pi_2, \cdots, \pi_m]^T$ associated with the eigenvalue 1. Importantly, all entries of π are positive, and they satisfy the following condition:

$$\sum_{i=1}^{m} \pi_i = 1. \tag{2.3.6}$$

With this mixing matrix, we have the following lemma.

Lemma 2.3.7 ([39])**.** *Let the mixing matrix* W *be defined by* (2.3.5)*, and* $\pi = [\pi_1, \pi_2, \cdots, \pi_m]^T$ *be the right eigenvector of* W *associated with the eigenvalue 1. Then we have*

$$\lim_{t\to\infty} W^t = \pi \mathbf{1}^T \tag{2.3.7}$$

To illustrate the role of the mixing matrix in a directed graph, we consider the following dynamic.

$$\begin{aligned} \mathbf{x}(t+1) &= W\mathbf{x}(t) \\ \mathbf{y}(t+1) &= W\mathbf{y}(t), \end{aligned} \tag{2.3.8}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_m(t)) \in \mathbb{R}^m$, $\mathbf{y}(t) = (y_1(t), y_2(t), \cdots, y_m(t)) \in \mathbb{R}^m$ and $y(0) = \mathbf{1}^T$. Then by Lemma 2.3.7 it follows that

$$\lim_{t\to\infty} \frac{x_i(t)}{y_i(t)} = \frac{\pi_i \sum_{j=1}^{m} x_j(0)}{\pi_i} = \sum_{j=1}^{m} x_j(0). \tag{2.3.9}$$

Therefore the consensus is achieved, i.e.

$$\lim_{t\to\infty} \frac{\mathbf{x}(t)}{\mathbf{y}(t)} = (\bar{x}(0), \cdots, \bar{x}(0)). \tag{2.3.10}$$

Chapter 3 and Chapter 4 delves into the investigation of decentralized optimization concerning undirected graphs. In contrast, Chapter 5 delves into the examination of decentralized optimization for directed graphs.

2.4 Sequential Estimate

This section is devoted to introducing sequential estimates namely Proposition 2.4.1 and Proposition 2.4.3, which will be mainly used in Chapter 3 and 4. These propositions will serve as crucial tools for deriving convergence estimates based on the sequential inequalities. We mention that the proofs of Proposition 2.4.1 and Proposition 2.4.3 are in Appendix

Proposition 2.4.1. *Let $p \in (0,1]$ and $q > 0$. Take $C_1 > 0$ and $w \geqslant 1$ such that $C_1/w^p < 1$. Suppose that the sequence $\{A(t)\}_{t\geqslant 0}$ satisfies*

$$A(t) \leqslant \left(1 - \frac{C_1}{(t+w-1)^p}\right) A(t-1) + \frac{C_2}{(t+w-1)^{p+q}} \quad \textit{for all } t \geqslant 1. \tag{2.4.1}$$

Set $Q = \left(\frac{w+1}{w}\right)^{p+q}$. Then $A(t)$ satisfies the following bound.

Case 1. If $p < 1$, then we have

$$A(t) \leqslant \delta \cdot ([t/2] + w - 1)^{-q} + \mathcal{R}(t),$$

where $\delta_1 = \frac{QC_2}{C_1} e^{\frac{C_1}{w^p}}$ and

$$\mathcal{R}(t) = e^{-\sum_{s=0}^{t-1} \frac{C_1}{(s+w)^p}} A(0) + QC_2 e^{-\frac{C_1 t}{2(t+w)^p}} \sum_{s=1}^{[t/2]-1} \frac{1}{(s+w)^{p+q}}.$$

Here the second term on the right hand side is assumed to be zero for $1 \leqslant t \leqslant 3$.

Case 2. If $p = 1$, then we have

$$A(t) \leqslant \left(\frac{w}{t+w}\right)^{C_1} A(0) + \mathcal{R}'(t),$$

where

$$\mathcal{R}'(t) = \begin{cases} \frac{w^{C_1-q}}{q-C_1} \cdot \frac{QC_2}{(t+w)^{C_1}} & \textit{if } q > C_1 \\ \log\left(\frac{t+w}{w}\right) \cdot \frac{QC_2}{(t+w)^{C_1}} & \textit{if } q = C_1 \\ \frac{1}{C_1-q} \cdot \left(\frac{w+1}{w}\right)^{C_1} \cdot \frac{QC_2}{(t+w+1)^q} & \textit{if } q < C_1. \end{cases}$$

Remark 2.4.2 (Bound of $\mathcal{R}(t)$). Here we show that

$$\mathcal{R}(t) = e^{-\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p}}A(0) + QC_2 e^{-\frac{C_1 t}{2(t+w)^p}}\sum_{s=1}^{[t/2]-1}\frac{1}{(s+w)^{p+q}} = O(t^{-N}).$$

Since we know that $QC_2 e^{-\frac{C_1 t}{2(t+w)^p}}\sum_{s=1}^{[t/2]-1}\frac{1}{(s+w)^{p+q}} = O(t^{-N})$, it is sufficient to show that $e^{-\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p}}A(0) = O(t^{-N})$. Note that

$$\begin{aligned}\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p} &\geqslant \int_0^t \frac{C_1}{(s+w)^p}ds\\ &= \frac{C_1}{1-p}\Big[(t+w)^{1-p} - w^{1-p}\Big].\end{aligned}$$

Using this we estimate the first term of $\mathcal{R}(t)$ as

$$e^{-\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p}}A(0) \leqslant A(0)e^{\frac{C_1}{1-p}w^{1-p}}e^{-\frac{C_1}{1-p}(t+w)^{1-p}} = O(t^{-N}). \tag{2.4.2}$$

Proposition 2.4.3. *Fix $p \in (0,1]$. Let $a, b,$ and w be positive values and $\beta \in (0,1)$ satisfying $\frac{a}{w^p} < 1$. Assume that a positive sequence $\{B(t)\}_{t\geqslant 0}$ satisfies*

$$B(t+1) \leqslant \Big(1 - \frac{a}{(t+w)^p}\Big)B(t) + \frac{b\beta^t}{(t+w)^p}$$

for $t \geqslant 0$ and $B(0) = 0$. Then we have the following estimates.

1. *If $p = 1$, then*

$$B(t+1) \leqslant J_{a,w,\beta}\cdot\frac{bw^{a-1}}{(t+w)^a} + \frac{b\beta^t}{(t+w)},$$

where $J_{a,w,\beta} = \sum_{j=0}^{\infty}\frac{(j+w)^{a-1}}{w^{a-1}}\beta^j$. Or we have

$$B(t+1) \leqslant \frac{bw^l}{1 - e^{l/w}\beta}\cdot\frac{1}{(t+w)^{l+1}} + \frac{b\beta^t}{(t+w)},$$

where l is any number satisfying $0 \leqslant l \leqslant a-1$ and $e^{l/w}\beta < 1$.

2. *If $p \in (0,1)$, then*

$$B(t+1) \leqslant \frac{b}{w^p}\bigg(e^{-\frac{a}{2}\cdot\frac{t}{(t+w)^p}} + \frac{\beta^{[t/2]}}{(1-\beta)}\bigg).$$

Chapter 3

Unconstrained Decentralized Gradient Descent

3.1 Introduction

We consider the problem (1.0.1) for the unconstrained space, where Ω corresponds to the Euclidean space $\mathbb{R}^d$, i.e.

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

One fundamental algorithm for this problem is the decentralized gradient descent (DGD) which is introduced by Nedić and Ozdaglar [33]. This algorithm consists of a consensus step based on a communication pattern designed by the mixing matrix and a local gradient step which is stated as follows:

$$x_i\left(t+1\right) = \sum_{j=1}^{m} w_{ij} x_j\left(t\right) - \alpha\left(t\right) \nabla f_i\left(x_i\left(t\right)\right). \tag{3.1.1}$$

Here the row vector $x_j(t) \in \mathbb{R}^d$ is the state at time $t \geqslant 0$ handled by agent j and $\alpha(t) > 0$ is the step size. In [49], the authors obtain the exponential convergence rate of the DGD algorithm with a constant step size. However, the results in [49] do not imply the exact convergence result. In order to see why the constant step size causes of inexact convergence, we rewrite (3.1.1) as

$$\mathbf{x}(t+1) = W\mathbf{x}(t) - \alpha \nabla F(\mathbf{x}(t)), \tag{3.1.2}$$

where $\mathbf{x}(t) = \left[x_1(t)^T, x_2(t)^T, \cdots, x_m(t)^T\right]^T \in \mathbb{R}^{m \times d}$ and

$$\nabla F(\mathbf{x}(t)) = \left[\nabla f_1(x_1(t))^T, \nabla f_2(x_2(t))^T, \cdots, \nabla f_m(x_m(t))\right]^T \in \mathbb{R}^{m \times d}.$$

Assuming the existence of the limit, let $\mathbf{x}(\infty) = \lim_{t\to\infty} \mathbf{x}(t)$. We also assume that the consensus is achieved, i.e. $\mathbf{x}(\infty) = W\mathbf{x}(\infty)$. Taking the limit over t on (3.1.2) and using the consensus, it follows that $\nabla F(\mathbf{x}(\infty)) = 0$, which is equivalent to $\nabla f_i(x_i(\infty)) = 0$ for all $i \in \{1, 2, \cdots, m\}$. The consensus implies that $x_i(\infty)$ concurrently minimizes f_i for all $i \in 1, 2, \cdots, m$, which is generally not possible. Furthermore, decentralized optimization aims to identify the minimizer of $f(x) = \sum_{i=1}^{m} f_i(x)$, rather than f_i. That's why we consider the diminishing step size.

The goal in this section is to establish the convergence property of the DGD within a broad range of diminishing step sizes denoted as $\{\alpha(t)\}_{t\geqslant 0}$, where $\alpha(t) = a/(t+w)^p$. Here, $a > 0$, $w \geqslant 1$, and $0 < p \leqslant 1$. Throughout this section, we make the following assumptions for functions and an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Assumption 3.1.1.

(A) For each $i \in \{1, \cdots, m\}$, the local function f_i is L_i-smooth for some $L_i > 0$.

(B) The total cost function f is μ-strongly convex for some $\mu > 0$.

(C) The communication graph $\mathcal{G}$ is connected.

(D) The mixing matrix W is doubly stochastic. In addition, $w_{ii} > 0$ for all $i \in \mathcal{V}$.

Next, we define the following constants which are frequently used throughout this section.

- R, the uniform upper bound for the quantities $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$ and $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|$ for $t \geqslant 0$
- $D = \max_{1\leqslant i\leqslant m} \|\nabla f_i(x_*)\|$ and $L = \max_{1\leqslant i\leqslant m} L_i$
- $d = 2LR + \sqrt{m}D$ and $\eta = \frac{\mu L}{\mu+L}$
- $R_0 = \|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{d\alpha(0)}{1-\beta}$
- $Q_0 = \left(\frac{w+1}{w}\right)^{2p}$ and $Q_1 = \frac{Q_0 L d 2^p}{\eta}$

The contents in this chapter largely follow the results in [12].

3.2 Sequential Estimates

We begin by deriving sequential estimates for both $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$ and $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|$ as part of our effort to establish consensus and convergence results.

To do this, we first rewrite the DGD algorithm in (3.1.1) in terms of $\bar{x}(t)$ and $\mathbf{x}(t)$. By summing up (3.1.1) for $1 \leqslant i \leqslant m$, we have

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha(t)}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)). \tag{3.2.1}$$

Additionally, we may write (3.1.1) in a compact form as

$$\mathbf{x}(t+1) = W\mathbf{x}(t) - \alpha(t)\nabla F(\mathbf{x}(t)). \tag{3.2.2}$$

In the following lemma, we obtain a bound of $\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\|$ in terms of $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$ and $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|$.

Lemma 3.2.1. *Suppose that Assumption 3.1.1 holds. Let $\{x_i(t)\}_{t\geqslant 0}$ be the sequence generated by (3.1.1). If a diminishing sequence $\{\alpha(t)\}_{t\in\mathbb{N}_0}$ satisfies $\alpha(0) \leqslant \frac{2}{\mu+L}$, then we have*

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| \leqslant (1 - \eta\alpha(t))\, \|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + L\alpha(t)\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|. \tag{3.2.3}$$

Proof. Using (3.2.1) and the triangle inequality, it follows that

$$\begin{aligned}
&\|\bar{x}(t+1) - x_*\| \\
&= \left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{m} \sum_{i=1}^{m} \nabla f_i(x_i(t)) \right\| \\
&\leqslant \left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{m} \sum_{i=1}^{m} \nabla f_i(\bar{x}(t)) \right\| + \frac{\alpha(t)}{m} \sum_{i=1}^{m} \|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\|.
\end{aligned} \tag{3.2.4}$$

We begin by estimating the first term in the final inequality of (3.2.4). Since $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$, it follows that

$$\begin{aligned}
&\|\bar{x}(t) - x_* - \alpha(t)\nabla f(\bar{x}(t))\|^2 \\
&= \|\bar{x}(t) - x_*\|^2 - 2\alpha(t)\langle \bar{x}(t) - x_*, \nabla f(\bar{x}(t))\rangle + \alpha(t)^2\|\nabla f(\bar{x}(t))\|^2 \\
&= \|\bar{x}(t) - x_*\|^2 - 2\alpha(t)\langle \bar{x}(t) - x_*, \nabla f(\bar{x}(t)) - \nabla f(x_*)\rangle \\
&\qquad + \alpha(t)^2\|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2.
\end{aligned}$$

Here we used the first optimality condition, $\nabla f(x_*) = 0$, for the last equality. Since f is L-smooth and μ-strongly convex, it follows by Lemma 2.1.9 that

$$\begin{aligned}
&\|\bar{x}(t) - x_* - \alpha(t)\nabla f(\bar{x}(t))\|^2 \\
&\leqslant \left(1 - \frac{2\mu L\alpha(t)}{\mu+L}\right)\|\bar{x}(t) - x_*\|^2 + \left(\alpha(t)^2 - \frac{2\alpha(t)}{\mu+L}\right)\|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2.
\end{aligned}$$

Since $\alpha(t) \leqslant \frac{2}{\mu+L}$ and $2L \geqslant \mu + L$, it follows that

$$\begin{aligned}\left\|\bar{x}(t) - x_* - \frac{\alpha(t)}{m}\sum_{i=1}^{m}\nabla f_i(\bar{x}(t))\right\|^2 &\leqslant \left(1 - \frac{2\mu L\alpha(t)}{\mu + L}\right)\|\bar{x}(t) - x_*\|^2 \\ &\leqslant \Big(1 - \eta\alpha(t)\Big)^2\|\bar{x}(t) - x_*\|^2,\end{aligned} \tag{3.2.5}$$

where $\eta = \frac{2\mu L}{\mu+L}$.

Next, we bound the second term in the last inequality of (3.2.4). By L_i-smoothness of f_i, we have

$$\sum_{i=1}^{m}\|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\| \leqslant L\sum_{i=1}^{m}\|\bar{x}(t) - x_i(t)\| \leqslant L\sqrt{m}\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|. \tag{3.2.6}$$

Here we used $L = \max_{1\leqslant i\leqslant n} L_i$. Putting (3.2.5) and (3.2.6) into (3.2.4), we obtain

$$\|\bar{x}(t+1) - x_*\| \leqslant (1 - \eta\alpha(t))\|\bar{x}(t) - x_*\| + \frac{L\alpha(t)}{\sqrt{m}}\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|.$$

Using the fact that $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| = \sqrt{m}\|\bar{x}(t) - x_*\|$, the inequality above yields the desired estimate:

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| \leqslant (1 - \eta\alpha(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + L\alpha(t)\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|.$$

□

Next we establish a bound of $\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\|$ in terms of $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|$ and $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$.

Lemma 3.2.2. *Suppose that Assumption 3.1.1 holds. Let $\{x_i(t)\}_{t\geqslant 0}$ be the sequence generated by* (3.1.1)*. Then we have*

$$\begin{aligned}&\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| \\ &\leqslant (\beta + L\alpha(t))\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| + L\alpha(t)\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + \sqrt{m}D\alpha(t).\end{aligned} \tag{3.2.7}$$

Proof. Using the scheme (3.2.2), it follows that

$$\begin{aligned}&\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| \\ &= \left\|W\mathbf{x}(t) - \alpha(t)\nabla F(\mathbf{x}(t)) - \bar{\mathbf{x}}(\mathbf{t}) + \frac{\alpha(t)}{m}\left(\mathbf{1}\mathbf{1}^T\nabla F(\mathbf{x}(t))\right)\right\| \\ &\leqslant \|W(\mathbf{x}(t) - \bar{\mathbf{x}}(t))\| + \alpha(t)\left\|\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right)\nabla F(\mathbf{x}(t))\right\| \\ &\leqslant \beta\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| + \alpha(t)\|\nabla F(\mathbf{x}(t))\|.\end{aligned} \tag{3.2.8}$$

Here we used $W\bar{\mathbf{x}}(t) = \bar{\mathbf{x}}(t)$ for the second inequality, and Lemma 2.3.5 and $\lambda_{\max}\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right) < 1$ for the last inequality. By L_i-smoothness of f_i, we have

$$\begin{aligned}\|\nabla f_i(x_i(t))\| &\leqslant \|\nabla f_i(x_i(t)) - \nabla f_i(\bar{x}(t))\| + \|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_*)\| + \|\nabla f_i(x_*)\| \\ &\leqslant L_i\|x_i(t) - \bar{x}(t)\| + L_i\|\bar{x}(t) - x_*\| + D.\end{aligned}$$

We recall that $L = \max_{1\leqslant i\leqslant n} L_i$ and use the Cauchy-Schwarz inequlaity to bound $\|\nabla F(\mathbf{x}(t))\|$ as

$$\begin{aligned}\|\nabla F(\mathbf{x}(t))\| &= \left(\sum_{i=1}^{m}\|\nabla f_i(x_i(t))\|^2\right)^{1/2} \\ &\leqslant L\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| + L\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + \sqrt{m}D.\end{aligned} \tag{3.2.9}$$

Inserting (3.2.9) into (3.2.8), we obtain the desired esitmate. □

3.3 Uniform Boundedness

In this section, we will establish the uniform boundedness of the sequences $\{\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|\}_{t\geqslant 0}$ and $\{\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|\}_{t\geqslant 0}$. These boundedness properties are crucial for deriving our main results. The following theorem states that the sequence $\{x_i(t)\}_{t\in\mathbb{N}_0}$ generated by (3.1.1) are uniformly bounded in the sense that there exists a positive value R such that $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant R$ and $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant R$ for all $t \geqslant 0$.

Theorem 3.3.1 (uniform boundedness). *Suppose that Assumption 3.1.1 holds. If $\{\alpha(t)\}_{t\geqslant 0}$ is non-increasing step size satisfying $\alpha(0) < \frac{\eta(1-\beta)}{L(\eta+L)}$ and set a finite value $R > 0$ as*

$$R = \max\left\{\|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|,\ \frac{L}{\eta}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|,\ \frac{\sqrt{m}D\alpha(0)}{\eta(1-\beta)/L - (\eta + L)\alpha(0)}\right\},$$

then we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant R \quad \textit{and} \quad \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \frac{\eta R}{L} < R, \quad \forall t \geqslant 0. \tag{3.3.1}$$

Proof. By the definition of R, we have

$$\|\bar{\mathbf{x}}(0) - \mathbf{x}_*\| \leqslant R \quad \textit{and} \quad \|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| \leqslant \frac{\eta}{L}R.$$

Next, we assume that the following inequalities

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant R \quad \textit{and} \quad \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \frac{\eta}{L}R \tag{3.3.2}$$

holds true for some fixed $t \geqslant 0$. Then, Substituting these bounds in (3.2.3), it follows that

$$\|\bar{\mathbf{x}}(t+1) - x_*\| \leqslant \Big(1 - \eta\alpha(t)\Big)R + L\alpha(t)\Big(\frac{\eta}{L}\Big)R = R.$$

Plugging (3.3.2) into (3.2.7), we have

$$\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| \leqslant (\beta + L\alpha(t))\Big(\frac{\eta}{L}\Big)R + L\alpha(t)R + \sqrt{m}D\alpha(t).$$

Using the definition of R and the fact that $\alpha(t)$ is non-increasing, it follows that

$$\begin{aligned}\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| &\leqslant \Big[(\beta + L\alpha(t))\Big(\frac{\eta}{L}\Big) + L\alpha(t)\Big]R + \sqrt{m}D\alpha(t) \\ &= \left[\frac{\beta\eta}{L} + (\eta + L)\alpha(t)\right]R + \sqrt{m}D\alpha(t) \\ &\leqslant \left[\frac{\beta\eta}{L} + (\eta + L)\alpha(0)\right]R + \sqrt{m}D\alpha(0).\end{aligned}$$

Notice that the condition $\alpha(0) < \frac{\eta(1-\beta)}{L(\eta+L)}$ implies $\frac{(1-\beta)\eta}{L} - (\eta + L)\alpha(0) > 0$. Thus $R > 0$ is well-defined and the following inequality follows:

$$\left[\frac{\beta\eta}{L} + (\eta + L)\alpha(0)\right]R + \sqrt{m}D\alpha(0) \leqslant \frac{\eta}{L}R.$$

This completes the induction, and so the proof is done. □

Sharpness of the Condition

The uniform boundedness result of Theorem 3.3.1 was also demonstrated in [49] for the case of a constant step size and convex functions. Specifically, it was shown that uniform boundedness holds when $\alpha(0) \leqslant 1/L$, which is a less restrictive condition compared to the one in Theorem 3.3.1, where $\alpha(0) < \frac{\eta(1-\beta)}{L(\eta+L)}$ is required. In contrast, Theorem 3.3.1 permits each individual function to be nonconvex, provided that the total cost is assumed to be strongly convex. Additionally, it allows for a time-varying step size.

To verify the sharpness of the range $\alpha(0) < \frac{\eta(1-\beta)}{L(\mu+L)}$ of the assumptions in Theorem 3.3.1, we construct an example with the following function:

$$f_1(x) = a_1x^2 \quad \text{and} \quad f_2(x) = -a_2x^2, \quad x \in \mathbb{R},$$

where a_1 and a_2 are positive values satisfying $a_1 - a_2 > 0$. Then the total cost $f = 1/2(f_1 + f_2)$ is strongly convex even though the local cost f_2 is non-convex. We take a value $\gamma \in (0, 1/2]$ and set a doubly stochastic matrix W by

$$W = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix}.$$

Let $x_1(t)$ and $x_2(t)$ be the variables that are only known to agents 1 and 2, respectively.

Lemma 3.3.2. *If $\alpha > \frac{\gamma(a_1-a_2)}{2a_1a_2}$, then the sequence $\{(x_1(t), x_2(t))\}_{t\geqslant 0}$ generated by the decentralized gradient descent algorithm* (3.1.1) *with any initial data* $(x_1(0), x_2(0)) \in (\mathbb{R}\backslash\{0\})^2$ *diverges.*

Proof. In this example, the decentralized gradient descent (3.1.1) is written as

$$\begin{aligned} x_1(t+1) &= (1-\gamma)x_1(t) + \gamma x_2(t) - 2\alpha a_1 x_1(t) \\ x_2(t+1) &= \gamma x_1(t) + (1-\gamma)x_2(t) + 2\alpha a_2 x_2(t). \end{aligned}$$

This can be written as follows

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = M \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix},$$

where

$$M = \begin{bmatrix} 1-\gamma-2\alpha a_1 & \gamma \\ \gamma & 1-\gamma+2\alpha a_2 \end{bmatrix}.$$

The eigenvalues of the matrix M are given by $\lambda \in \mathbb{R}$, which satisfy the following equation:

$$\lambda^2 - 2(1-\gamma+(a_2-a_1)\alpha)\lambda + (1-\gamma-2a_1\alpha)(1-\gamma+2a_2\alpha) - \gamma^2 = 0.$$

The solutions to this equation are:

$$(1-\gamma) + (a_2-a_1)\alpha \pm \sqrt{(a_1+a_2)^2\alpha^2+\gamma^2}.$$

This formula allows us to show that the largest eigenvalue is larger than 1 if $\alpha > \frac{\gamma(a_1-a_2)}{2a_1a_2}$, which implies the sequence $\{(x_1(t), x_2(t)\}$ diverges. In fact, it is checked in the following way

$$\begin{aligned} &(1-\gamma) + (a_2-a_1)\alpha + \sqrt{(a_1+a_2)^2\alpha^2+\gamma^2} > 1 \\ &\Leftrightarrow \sqrt{(a_1+a_2)^2\alpha^2+\gamma^2} > \gamma + (a_1-a_2)\alpha \\ &\Leftrightarrow (a_1+a_2)^2\alpha^2+\gamma^2 > \gamma^2 + 2\gamma(a_1-a_2)\alpha + (a_1-a_2)^2\alpha^2 \\ &\Leftrightarrow 4a_1a_2\alpha^2 > 2\gamma(a_1-a_2)\alpha \\ &\Leftrightarrow \alpha > \frac{\gamma(a_1-a_2)}{2a_1a_2}. \end{aligned}$$

The proof is done. □

Now we show that the range $\alpha(0) < \frac{\eta(1-\beta)}{L(\mu+L)}$ is almost sharp for large L and small μ in this example.

Corollary 3.3.3. *If $\alpha(0) > \frac{\mu(1-\beta)}{L(L-2\mu)}$, then the sequences $\{x_i(t)\}_{t\in\mathbb{N}_0}$ may diverge for $i = 1, 2, \cdots, m$. This implies that the condition $\alpha(0) < \frac{\eta(1-\beta)}{L(\mu+L)}$ is sharp in the sense that*

$$\lim_{L\to\infty} \frac{\eta(1-\beta)}{L(\eta+L)} \Big/ \frac{\mu(1-\beta)}{L(L-2\mu)} = 1 \quad \textit{and} \quad \lim_{\mu\to 0} \frac{\eta(1-\beta)}{L(\eta+L)} \Big/ \frac{\mu(1-\beta)}{L(L-2\mu)} = 1.$$

Proof. In the setting of Lemma 3.3.2, we let $a_1 = \frac{L}{2}$ and $a_2 = \frac{L}{2} - \mu$ with a value $\mu > 0$ and a large number $L > 0$. Then f_1 and f_2 are L-smooth functions and $f(x) = (\mu/2)x^2$ is μ-strongly convex. Also we have $\beta = 1 - 2\gamma$ in Lemma 2.3.5. Then the condition on $\alpha(0)$ of Lemma 3.3.2 is written as

$$\alpha > \frac{2\mu\gamma}{L(L-2\mu)}. \tag{3.3.3}$$

On the other hand, the condition of Theorem 3.3.1 is written as

$$\left(\frac{\mu L}{\mu+L} + L\right)\alpha < \frac{\mu L}{\mu+L} \cdot \frac{2\gamma}{L},$$

which is equivalent to

$$\alpha < \frac{2\mu\gamma}{\left(\frac{\mu L}{\mu+L} + L\right)(\mu+L)}. \tag{3.3.4}$$

Thus the condition (3.3.3) is sharp in the sense that the right hand sides of (3.3.3) and (3.3.4) are very close when L is sufficiently large, which also can be seen by the limit

$$\lim_{L\to\infty} \frac{2\mu\gamma}{L(L-2\mu)} \Big/ \frac{2\mu\gamma}{\left(\frac{\mu L}{\mu+L} + L\right)(\mu+L)} = 1.$$

Similarly, for sufficiently small μ, we have

$$\lim_{\mu\to 0} \frac{2\mu\gamma}{L(L-2\mu)} \Big/ \frac{2\mu\gamma}{\left(\frac{\mu L}{\mu+L} + L\right)(\mu+L)} = 1.$$

The proof is done. □

3.4 Consensus Estimate

In the previous section, we established the uniform boundedness of the sequences $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$ and $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|$. Leveraging this result from Theorem 3.3.1, we now derive the following consensus result.

Theorem 3.4.1 (consensus)**.** *Suppose that Assumption 3.1.1 holds, and* $\{\alpha(t)\}_{t\in\mathbb{N}_0}$ *is non-increasing sequence satisfying*

$$\alpha(0) \leqslant \min\left\{\frac{2}{\mu+L}, \frac{\eta(1-\beta)}{L(\eta+L)}\right\}.$$

Let $\{x_i(t)\}_{t\geqslant 0}$ *be the sequence generated by* (3.1.1)*. Then, for all* $t \geqslant 0$ *we have*

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \frac{d}{1-\beta}\alpha(\lfloor t/2\rfloor) + \beta^t\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{\beta^{t/2}d}{1-\beta}\alpha(0). \tag{3.4.1}$$

Proof. We can easily check that the inequality (3.4.1) holds true for $t = 0$. Hence we consider the case $t \geqslant 1$. By Lemma 3.2.2 and Theorem 3.3.1, we have

$$\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| \leqslant \beta\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| + d\alpha(t), \tag{3.4.2}$$

where $d = 2LR + \sqrt{m}D$. Using this iteratively gives

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \beta^t\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + d\sum_{s=0}^{t-1}\beta^{t-1-s}\alpha(s). \tag{3.4.3}$$

Putting $t = 1$ into (3.4.3), we have

$$\begin{aligned}\|\mathbf{x}(1) - \bar{\mathbf{x}}(1)\| &\leqslant \beta^1\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + d\alpha(0)\\ &\leqslant \frac{d}{1-\beta}\alpha(0) + \beta\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{\beta^{1/2}d}{1-\beta}\alpha(0),\end{aligned}$$

which shows that (3.4.1) holds true for $t = 1$. To see that (3.4.1) holds true for $t \geqslant 2$, we observe the following inequality:

$$\begin{aligned}\sum_{s=0}^{t-1}\beta^{t-1-s}\alpha(s) &= \sum_{s=0}^{\lfloor t/2\rfloor-1}\beta^{t-1-s}\alpha(s) + \sum_{s=\lfloor t/2\rfloor}^{t-1}\beta^{t-1-s}\alpha(s)\\ &\leqslant \alpha(0)\sum_{s=0}^{\lfloor t/2\rfloor-1}\beta^{t-1-s} + \alpha(\lfloor t/2\rfloor)\sum_{s=0}^{\infty}\beta^s\\ &\leqslant \alpha(0)\frac{\beta^{t/2}}{1-\beta} + \alpha(\lfloor t/2\rfloor)\frac{1}{1-\beta}.\end{aligned}$$

Inserting this estimate into (3.4.3), it follows that

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \frac{d}{1-\beta}\alpha(\lfloor t/2\rfloor) + \beta^t\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{\beta^{t/2}d}{1-\beta}\alpha(0).$$

□

3.5 Convergence Analysis

In the previous section, we obtained the consensus result. Thanks to Theorem 3.4.1, it is sufficient to focus on the convergence of $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$. Thus, we will now establish its convergence for $p \in (0, 1]$. We first consider the case $p \in (0, 1)$.

Theorem 3.5.1 (convergence)**.** *Suppose that Assumption 3.1.1 holds and consider the sequence $x_i(t)_{t \geqslant 0}$ generated by* (3.1.1)*. Let $p \in (0,1)$ and assume that $\alpha(t) = \frac{a}{(t+w)^p}$ with constants $a > 0$ and $w \geqslant 1$ satisfying*

$$\alpha(0) = \frac{a}{w^p} \leqslant \min\left\{\frac{2}{\mu+L}, \frac{\eta(1-\beta)}{L(\eta+L)}\right\}. \tag{3.5.1}$$

Then we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant \frac{Q_1 ae}{1-\beta} \cdot \Big(\lfloor t/2 \rfloor + w - 1\Big)^{-p} + Y_1(t) + Y_2(t) + Y_3(t), \tag{3.5.2}$$

where

$$\begin{aligned}
Y_1(t) &= e^{-\sum_{s=0}^{t-1} \frac{\eta a}{(s+w)^p}} \|\mathbf{x}(0) - \mathbf{x}_*\|, \\
Y_2(t) &= \frac{Q_1}{1-\beta} e^{-\frac{\eta a}{2} \cdot \frac{t}{(t+w)^p}} \sum_{s=1}^{\lfloor t/2 \rfloor - 1} \frac{a^2}{(w+s)^{2p}}, \\
Y_3(t) &= \frac{aLR_0}{(1-\beta)w^p} \cdot \left(e^{-\frac{\eta a}{2} \frac{t}{(t+w)^p}} + \frac{\sqrt{\beta}^{\lfloor (t-1)/2 \rfloor}}{1-\sqrt{\beta}} \right).
\end{aligned}$$

Here $Q_0 = \left(\frac{w+1}{w}\right)^{2p}$ and $Q_1 = \frac{Q_0 L d 2^p}{\eta}$. It is easy to see that for any fixed $N > 0$, there exists a constant $C_N > 0$ independent of $t \geqslant 0$ such that

$$Y_1(t) + Y_2(t) + Y_3(t) \leqslant C_N t^{-N}.$$

In the estimate (3.5.2), it is observed that when the ratio a/w^p is significant, the exponential components in $Y_1(t), Y_2(t)$, and $Y_3(t)$ diminish rapidly, while the initial term in (3.5.2) become larger. Conversely, if we choose a smaller positive value of a/w^p, the first term in (3.5.2) gets small, but the exponential components decline at a slower rate. These observations indicate that by selecting different values for a and w, we can achieve either rapid convergence with trade-off of significant error in the initial phase, or vice versa

Next we state the result for the case $p = 1$.

Theorem 3.5.2 (convergence)**.** *Suppose that Assumption 3.1.1 holds. Let $\{x_i(t)\}_{t \geqslant 0}$ be the sequence generated by* (3.1.1)*. Let $p = 1$ and assume that $\alpha(t) = \frac{a}{(t+w)}$ with positive constants $a > 1$ and $w \geqslant 1$ satisfying*

$$\alpha(0) = \frac{a}{w} \leqslant \min\left\{\frac{2}{\mu+L}, \frac{\eta(1-\beta)}{L(\eta+L)}\right\}. \tag{3.5.3}$$

If $a > \frac{2}{\eta}$, then we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant \Big(\frac{w}{t+w}\Big)^{\eta a} \|\mathbf{x}(0) - \mathbf{x}_*\| + \frac{2\sqrt{e}Q_1}{(1-\beta)} \cdot \frac{a}{(t+w+1)} + Y_4(t), \tag{3.5.4}$$

where

$$Y_4(t) = \frac{aLR_2}{1-\beta}\left[\frac{w^l}{1-e^{l/w}\sqrt{\beta}} \cdot \frac{1}{(t-1+w)^{l+1}} + \frac{\sqrt{\beta}^t}{(t-1+w)}\right].$$

Here l is any value such that $0 \leqslant l \leqslant a-1$ and $e^{(l/w)}\sqrt{\beta} < 1$.

Similarly as in Theorem 3.5.1, we find that if a/w is small, then the second term in the right-hand side of (3.5.4) is small, but the first term decays slowly. On the contrary, if a/w is large, then the second term is significant while the first term decays fast.

Remark 3.5.3. One of the crucial steps to obtain our main results is to show the uniform boundedness for the sequence $\{x_i(t)\}_{t\geqslant 0}$. And the bound $\frac{\eta(1-\beta)}{L(\eta+L)}$ of (3.5.1) and (3.5.3) is used to obtain uniform boundedness. This implies that if we can guarantee that the sequence $x_i(t)$ is uniformly bounded, then the bound $\frac{2}{\mu+L}$ from equations (3.5.1) and (3.5.3) is sufficient to estimate the convergence, as stated in the aforementioned theorem

Proof of Theorems 3.5.1 and 3.5.2. Define $V(t)$ as

$$V(t) = \beta^t\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{\beta^{t/2}}{1-\beta}(d\alpha(0)). \tag{3.5.5}$$

Then we can write (3.4.1) as

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \leqslant \frac{d}{1-\beta}\alpha(\lfloor t/2 \rfloor) + V(t). \tag{3.5.6}$$

For the reader's convenience, we recall (3.2.3) in Lemma 3.2.1 as

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| \leqslant (1-\eta\alpha(t))\,\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + L\alpha(t)\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|.$$

Inserting (3.5.5) into (3.2.3), we get

$$\begin{aligned}&\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| \\ &\leqslant \Big(1-\eta\alpha(t)\Big)\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + L\alpha(t)\cdot\frac{d}{1-\beta}\alpha(\lfloor t/2\rfloor) + L\alpha(t)V(t).\end{aligned} \tag{3.5.7}$$

Notice that for $w \geqslant 1$,

$$\lfloor t/2 \rfloor + w \geqslant \frac{t-1}{2} + w \geqslant \frac{t+w}{2}$$

for all $t \geqslant 0$. This implies that

$$\alpha(\lfloor t/2 \rfloor) = \frac{a}{(\lfloor t/2 \rfloor + w)^p} \leqslant \frac{2^p a}{(t+w)^p}.$$

Using this observation, we write (3.5.7) as

$$\|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| \leqslant \Big(1 - \frac{C}{(t+w)^p}\Big)\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + \frac{C'}{(t+w)^{2p}} + \frac{aLV(t)}{(t+w)^p},$$

where

$$C = \eta a, \ C' = \frac{Ld2^p a^2}{1-\beta}.$$

In order to estimate $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$ from this sequential inequality, we consider two sequences $\{G(t)\}_{t \geqslant 0}$, and $\{K(t)\}_{t \geqslant 0}$ such that

$$\begin{aligned} G(t+1) &= \Big(1 - \frac{C}{(t+w)^p}\Big)G(t) + \frac{C'}{(t+w)^{2p}}, \quad \text{with} \quad G(0) = \|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|, \\ K(t+1) &= \Big(1 - \frac{C}{(t+w)^p}\Big)K(t) + \frac{aLV(t)}{(t+w)^p}, \quad \text{with} \quad K(0) = 0. \end{aligned}$$

Then we have the following inequality

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| \leqslant G(t) + K(t), \ \text{for } t \geqslant 0. \tag{3.5.8}$$

We first consider the case $p < 1$. Using Proposition 2.4.1 we estimate $G(t)$ as

$$G(t) \leqslant \delta_1 \cdot \Big([t/2] + w - 1\Big)^{-p} + \mathcal{R}_1(t),$$

with constant

$$\delta_1 = \Big(\frac{w+1}{w}\Big)^{2p} \frac{C'}{C} e^{\frac{C}{w^p}} = Q_0 \frac{C'}{C} e^{\frac{C}{w^p}}$$

and

$$\begin{aligned} \mathcal{R}_1(t) &= e^{-\sum_{s=0}^{t-1} \frac{C}{(s+w)^p}} G(0) + Q_0 C' e^{-\frac{\eta a}{2} \cdot \frac{t}{(t+w)^p}} \sum_{s=1}^{[t/2]-1} \frac{1}{(w+s)^{2p}} \\ &=: Y_1(t) + Y_2(t). \end{aligned}$$

Here $Q_0 = \Big(\frac{w+1}{w}\Big)^{2p}$. Next, notice from (3.5.5) that

$$V(t) \leqslant \frac{\beta^{t/2}}{1-\beta} R_2,$$

where $R_2 = \|\mathbf{x}(0) - \overline{\mathbf{x}}(0)\| + \frac{d}{1-\beta}\alpha(0)$. Combining this with Proposition 2.4.3 we find that

$$K(t) \leqslant \frac{aLR_2}{(1-\beta)w^p}\left(e^{-\frac{C}{2}\frac{t}{(t+w)^p}} + \frac{\sqrt{\beta}^{[(t-1)/2]}}{1-\sqrt{\beta}}\right) =: Y_3(t)$$

Putting these estimates in (3.5.8) and using that $c/w^p \leqslant 1$, we get

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| &\leqslant \delta \cdot \Big([t/2] + w - 1\Big)^{-p} + Y_1(t) + Y_2(t) + Y_3(t)\\ &\leqslant \frac{(Q_1 e)a}{1-\beta} \cdot \Big([t/2] + w - 1\Big)^{-p} + Y_1(t) + Y_2(t) + Y_3(t),\end{aligned}$$

which is the desired estimate.

Next, we consider the case $p = 1$. In this case, we assume that $a > \frac{2}{\eta}$ which implies $C = \eta a > 2$. Applying Proposition 2.4.1 again to $G(t)$ we deduce

$$\begin{aligned}G(t) &\leqslant \Big(\frac{w}{t+w}\Big)^C \|\mathbf{x}(0) - \mathbf{x}_*\| + \left(\frac{w+1}{w}\right)^C \cdot \frac{Q_0 C'}{(C-1)(t+w+1)}\\ &\leqslant \Big(\frac{w}{t+w}\Big)^C \|\mathbf{x}(0) - \mathbf{x}_*\| + \left(\frac{w+1}{w}\right)^C \frac{2Q_0 C'}{C(t+w+1)}.\end{aligned}$$

Here we used the inequality $\frac{1}{C-1} \leqslant \frac{2}{C}$ since $C > 2$. Using $1 + x \leqslant e^x$ and the condition $\frac{a}{w} \leqslant \frac{2}{\mu+L}$, we have

$$\left(1 + \frac{1}{w}\right)^C \leqslant e^{\frac{\eta a}{w}} \leqslant e^{\frac{2\mu L}{(\mu+L)^2}} \leqslant \sqrt{e}.$$

We also use Proposition 2.4.3 to find that

$$K(t) \leqslant \frac{aLR_2}{1-\beta}\left[\frac{w^l}{1 - e^{l/w}\sqrt{\beta}} \cdot \frac{1}{(t-1+w)^{l+1}} + \frac{\sqrt{\beta}^t}{(t-1+w)}\right] =: Y_4(t).$$

Combining these estimates with (3.5.8), we have

$$\begin{aligned}&\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|\\ &\leqslant \Big(\frac{w}{t+w}\Big)^C \|\mathbf{x}(0) - \mathbf{x}_*\| + \frac{2\sqrt{e}Q_0 C'}{C(t+w+1)} + Y_4(t)\\ &= \Big(\frac{w}{t+w}\Big)^{\eta a} \|\mathbf{x}(0) - \mathbf{x}_*\| + \frac{2\sqrt{e}Q_1 0}{(1-\beta)} \cdot \frac{a}{(t+w+1)} + Y_4(t),\end{aligned}$$

where $Q_1 = \frac{Q_0 L d 2^p}{\eta}$. This gives the desired bound. The proof is done. □

3.6 Numerical Experiments

In this section, we present a numerical experiment for algorithm (3.1.1) with diminishing step size and investigate the condition for achieving uniform boundedness.

Regression problem

Let m be the number of agents, and for each $1 \leqslant i \leqslant m$, we randomly select a matrix A_i from $\mathbb{R}^{n\times d}$ with elements uniformly distributed in the range [0, 1]. In this simulation, we set the problem dimensions and the number of agents as $n = 5$, $d = 10$ and $m = 30$. In addition, each agent has four neighbors. We construct a connected graph based on the Watts and Strogatz model [44], and employ a Laplacian-based mixing matrix denoted as W. This matrix is defined by

$$W = \mathbf{I} - \frac{1}{\tau}L,$$

where L is the Laplacian matrix and τ is a constant that satisfies $\tau > \frac{1}{2}\lambda_{\max}(L)$. Next, we choose a value $x_* \in \mathbb{R}^d$ and define $y_i = A_i x_* + \epsilon \in \mathbb{R}^n$, where each element of $\epsilon \in \mathbb{R}^n$ follows a normal distribution with mean 0 and variance 0.1. The cost is defined as

$$f(x) = \frac{1}{m}\sum_{k=1}^{m} \|A_k x - y_k\|^2.$$

We choose step sizes $\alpha_i(t) = \frac{a_i}{(t+w)^p}$ for $i = 1, \ldots, 4$ as follows: Setting $w = Z^{1/p}$ with $Z = \frac{16L(\eta+L)}{\mu\eta(1-\beta)}$, we assign the following values to a_i:

$$a_1 = \frac{w^p}{5(\mu+L)}, \quad a_2 = \frac{w^p}{50(\mu+L)}, \quad a_3 = \frac{\eta(1-\beta)w^p}{1.1L(\eta+L)}, \quad a_4 = \frac{\eta(1-\beta)w^p}{2L(\eta+L)}.$$

Then we compute $\alpha_i(0) = \frac{a_i}{w^p}$ as follows:

$$\frac{1}{5(\mu+L)}, \quad \frac{1}{50(\mu+L)}, \quad \frac{\eta(1-\beta)}{1.1L(\eta+L)}, \quad \frac{\eta(1-\beta)}{2L(\eta+L)}.$$

We test the algorithm (3.1.1) with $\alpha_i(t) = \frac{a_i}{(t+w)^p}$ with above choices of a_i and w for $p \in \{0.25, 0.5, 0.75, 1\}$. We measure the error

$$R(t) = \frac{\sum_{i=1}^{n} \|x_i(t) - x_*\|}{\sum_{i=1}^{n} \|x_i(0) - x_*\|}.$$

and the result is presented in Figure 3.1. As we expected in Theorems 3.5.1 and 3.5.2, we get fast convergence for large iterations but slow decay in the early stage if the value a/w^p is small and vice versa for large a/w^p.

Sharpness of the Condition for Achieving Uniform Boundedness.

Next, we provide a simulation result supporting the result of Lemma 3.3.2. We define the cost function $f(x)$ and doubly stochastic W as

$$f(x) = \frac{f_1(x) + f_2(x)}{2} = \frac{a_1x^2 - a_2x^2}{2}$$

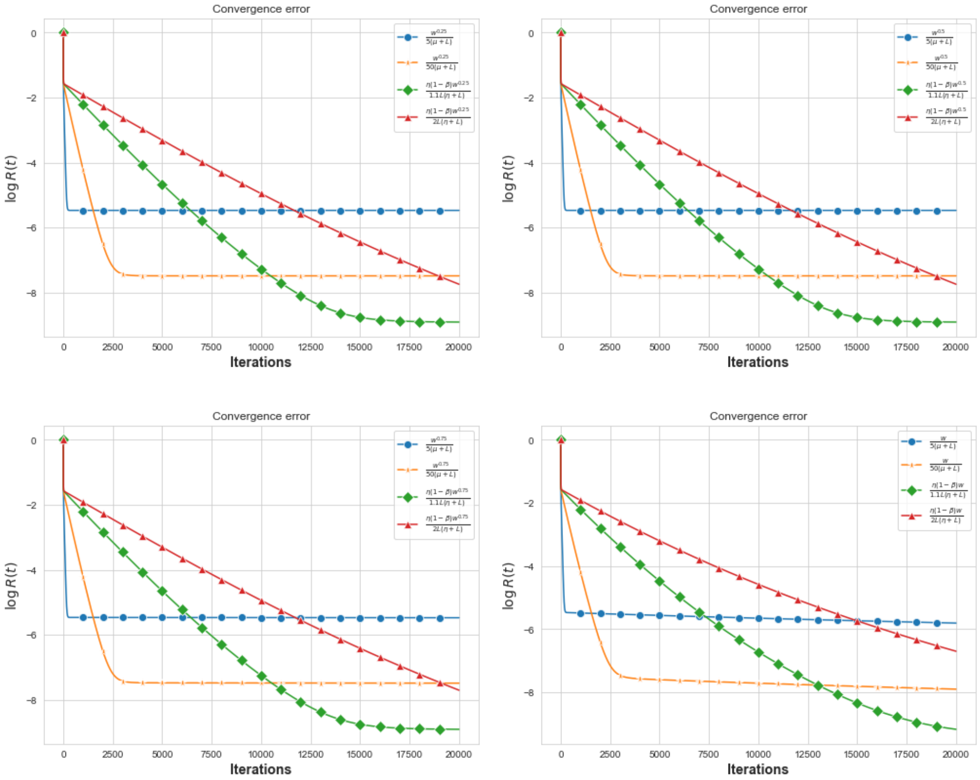


Figure 3.1: Relative error with various choices of p. (Top-Left) $p = 0.25$ (Top-Right) $p = 0.5$ (Bottom-left) $p = 0.75$ (Bottom-Right) $p = 1$

and

$$W = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix}.$$

We note that $x = 0$ is the optimal value. In this simulation, we take $a_1 = 10$, $a_2 = 6$ and $\gamma = 0.2$ and consider the following values of k,

$$k_1 = 2.1,\ k_2 = 2.01,\ k_3 = 2,\ k_4 = 1.99,\ k_5 = 1.9.$$

We test the algorithm (3.1.1) with $\alpha = \frac{\gamma(a_1+a_2)}{ka_1a_2}$ with above choices of k. The initial value $(x_1(0), x_2(0))$ is chosen randomly from $(0, 50) \times (0, 50)$. We measure the quantity $\left(x_1^2(t) + x_2^2(t)\right)^{1/2}$ and the result is presented in Figure 3.2.

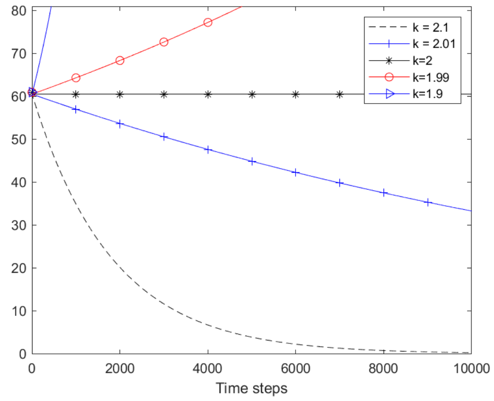


Figure 3.2: Relative error with various choices of step size

As we expected in Lemma 3.3.2 and Corollary 3.3.3, Figure 3.2 shows that the quantities diverge when α is larger than the threshold, $\frac{\gamma(a_1+a_2)}{2a_1a_2}$ and converge when α is smaller than the threshold.

Chapter 4

Constrained Decentralized Gradient Descent

4.1 Introduction

We consider the problem (1.0.1) for constrained space $\Omega \subseteq \mathbb{R}^d$ which is convex and closed. To solve this problem, we use the decentralized projected gradient descent (DPG) [33, 36] as follows:

$$x_i(t+1) = \mathcal{P}_\Omega \left[\sum_{j=1}^n w_{ij} x_j(t) - \alpha(t) \nabla, f_i(x_i(t)) \right] \tag{4.1.1}$$

where $\mathcal{P}_\Omega$ denotes the projection operator onto Ω, as defined in (2.2.1). In the previous section, we study the unconstrained case, i.e. $\Omega = \mathbb{R}^d$. However, when we consider the constrained case, i.e. $\Omega \neq \mathbb{R}^d$, the convergence analysis becomes more challenging due to the projection operator.

To explain the difficulty in the convergence analysis of (4.1.1) compared to the case $\Omega = \mathbb{R}^n$, we note that averaging (3.1.1) gives

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha(t)}{n} \sum_{i=1}^n \nabla f_i(x_i(t)), \tag{4.1.2}$$

where $\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t)$. Then, If we set the step size as $\alpha(t) \leqslant \frac{2}{\mu+L}$, we can derive the following inequality (refer to Lemma 3.2.1) when f is μ-strongly convex and each f_i is L-smooth.:

$$\begin{aligned} \|\bar{x}(t+1) - x_*\| &\leqslant \Big(1 - \frac{\mu L}{\mu + L} \alpha\Big) \|\bar{x}(t) - x_*\| + \frac{L\alpha}{n} \sum_{i=1}^n \|x_i(t) - \bar{x}(t)\| \\ &\leqslant (1 - c\alpha) \|\bar{x}(t) - x_*\| + c\alpha^2, \end{aligned}$$

where c is a proper constant. This inequality is a major ingredient in the convergence estimate of (3.1.1) in the previous section, but the identity (4.1.2) no longer holds for (4.1.1) due to the projection.

The argument of [25]

To overcome this difficulty, the work [25] begins with writing the DPG in the following way:

$$x_i(t+1) = \sum_{j=1}^n w_{ij}x_j(t) - \alpha\nabla f_i(x_i(t)) + \phi_i(t) \tag{4.1.3}$$

where $\phi_i(t)$ is the difference between DGD and DPG defined as follows:

$$\phi_i(t) = \underbrace{\sum_{j=1}^n w_{ij}x_j(t) - \alpha(t)\nabla f_i(x_i(t))}_{\text{DGD}} - \underbrace{\mathcal{P}_\Omega\left[\sum_{j=1}^n w_{ij}x_j(t) - \alpha(t)\nabla f_i(x_i(t))\right]}_{\text{DPG}}.$$

Averaging (4.1.3) for $1 \leqslant k \leqslant n$ one has

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_i(t)) + \frac{1}{n}\sum_{i=1}^n \phi_i(t). \tag{4.1.4}$$

The work [25] treated the last term of (4.1.4) as an additional error term and bound it as

$$\left\|\frac{1}{n}\sum_{i=1}^n \phi_i(t)\right\| = O(\alpha). \tag{4.1.5}$$

Then, the contraction property of the projection operator (refer to Lemma 2.2.3) is applied to achieve the following estimate

$$\begin{aligned} &\|\bar{x}(t+1) - x_*\| \\ &= \left\|P_\Omega[\bar{x}(t+1)] - P_\Omega\left[x_* - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_*)\right]\right\| \\ &\leqslant \left\|\bar{x}(t+1) - \Big(x_* - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_*)\Big)\right\| \\ &= \left\|\bar{x}(t) - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_i(t)) - x_* - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(x_*) + \frac{1}{n}\sum_{i=1}^n \phi_i(t)\right\|. \end{aligned} \tag{4.1.6}$$

Using the smoothness, strong convexity of the costs and (4.1.5), one may estimate the right-hand side of (4.1.6) as follows:

$$\begin{aligned}
&\|\bar{x}(t+1)-x_*\| \\
&\leqslant \left\| \bar{x}(t) - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}(t)) - x_* - \frac{\alpha}{n}\sum_{i=1}^{n}\nabla f_i(x_*) \right\| + \left\| \frac{1}{n}\sum_{i=1}^{n}\phi_i(t) \right\| \\
&\leqslant (1-c\alpha)\|\bar{x}(t)-x_*\| + O(\alpha) + \frac{L\alpha}{n}\sum_{i=1}^{n}\|\bar{x}(t)-x_i(t)\|.
\end{aligned} \tag{4.1.7}$$

We may observe that the estimate (4.1.7) is less effective since there is a $O(\alpha)$ term in the bound while the coefficient for the contraction of $\|\bar{x}(t)-x_*\|$ is $(1-c\alpha)$. More precisely, if we neglect the consensus error $\frac{L\alpha}{n}\sum_{k=1}^{n}\|\bar{x}(t)-x_k(t)\|$ in (4.1.7), we can write (4.1.7) as

$$\begin{aligned}
\|\bar{x}(t+1)-x_*\| &\leqslant (1-c\alpha)\|\bar{x}(t)-x_*\| + C\alpha \\
&\leqslant (1-c\alpha)^{t+1}\|\bar{x}(0)-x_*\| + \sum_{s=0}^{t}(1-c\alpha)^{s}C\alpha \\
&\leqslant (1-c\alpha)^{t+1}\|\bar{x}(0)-x_*\| + \frac{C}{c}(1-(1-c\alpha)^{t+1}).
\end{aligned}$$

Taking the limit $t\to\infty$ in the above estimate gives $\lim_{t\to\infty}\|\bar{x}(t)-x_*\| \leqslant \frac{C}{c}$, which is indepedent of $\alpha > 0$.

The idea for the $O(\sqrt{\alpha})$-convergence

Instead of using the error bound $\phi_i(t)$, we proceed to obtain a sequential estimate of the quantity

$$\sum_{i=1}^{N}\|x_i(t)-x_*\|^2,$$

which enables us to offset the projection operator efficiently using the contraction property of the projection operator. As a result, we obtain a convergence result up to an error $O(\sqrt{\alpha})$. The key part for obtaining $O(\sqrt{\alpha})$ is to use the contraction property of the projection operator in the following inequality

$$\begin{aligned}
&\|x_i(t+1)-x_*\|^2 \\
&= \left\| \mathcal{P}_\Omega\left[\sum_{j=1}^{n} w_{ij}x_j(t) - \alpha(t)\nabla f_i(x_i(t))\right] - \mathcal{P}_\Omega\left[x_* - \alpha(t)\nabla f(x_*)\right] \right\|^2 \\
&\leqslant \left\| \sum_{j=1}^{n} w_{ij}x_j(t) - x_* - \alpha(t)\left(\nabla f_i(x_i(t)) - \nabla f(x_*)\right) \right\|^2.
\end{aligned} \tag{4.1.8}$$

We will then sum this inequality over $1 \leqslant i \leqslant n$ and derive a gain from the right hand side, using the strong convexity and the smoothness property of the costs as well as the property of the mixing matrix W and some manipulations relying on the identity $\sum_{i=1}^n \|a+b_i\|^2 = n\|a\|^2 + \sum_{i=1}^n b_i^2$ which holds when $\sum_{i=1}^n b_i = 0$. Then, roughly we will get the following type of inequality

$$\|\mathbf{x}(t+1) - \mathbf{x}_*\|^2 \leqslant (1-c\alpha)\|\mathbf{x}(t) - \mathbf{x}_*\|^2 + C\alpha^2 \tag{4.1.9}$$

for some constants $c > 0$ and $C > 0$. Using this estimate recursively, one has

$$\begin{aligned} \|\mathbf{x}(t) - \mathbf{x}_*\|^2 &\leqslant (1-c\alpha)^t \|\mathbf{x}(0) - \mathbf{x}_*\|^2 + C\alpha^2 \sum_{k=0}^{t-1} (1-c\alpha)^k \\ &\leqslant (1-c\alpha)^t \|\mathbf{x}(0) - \mathbf{x}_*\|^2 + \frac{C\alpha}{c}, \end{aligned} \tag{4.1.10}$$

which justifies the $O(\sqrt{\alpha})$-convergence of the sequence $\mathbf{x}(t+1)$ towards the optimal point $\mathbf{x}_*$.

Towards the $O(\alpha)$-convergence

We recall from [49] that the sequence of the algorithm (3.1.1) on the whole space $\mathbb{R}^d$ converges to an $O(\alpha)$-neighborhood of x_*. This naturally leads us to pose the following question.

Question: *Is the convergence error $O(\sqrt{\alpha})$ in Theorem 4.6.1 optimal? or can we improve the convergence error to $O(\alpha)$?*

We give a partial answer to this question in Section 4.7. Precisely, we find a one-dimensional and a half-space examples such that the algorithm (4.1.1) converges to an $O(\alpha)$ neighborhood of the optimal point. Prior to introducing the concept of $O(\alpha)$-convergence in these examples, we briefly highlight a distinction between an argument presented in [49] and that in our work. Following arguments [49], it turns out that the following sequential estimates hold:

$$\begin{aligned} \|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| &\leqslant \beta \|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\| + C\alpha. \\ \|\bar{\mathbf{x}}(t+1) - \mathbf{x}_*\| &\leqslant (1-c\alpha)\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| + C\alpha \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| \end{aligned} \tag{4.1.11}$$

Here $c > 0$ and $C > 0$ are suitable constants and $\beta < 1$ is a value related to the mixing matrix W. Then, similarly to (4.1.10), we may deduce from each of these two estimates the following estimates

$$\begin{aligned} \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\| &\leqslant \beta^t \|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| + \frac{C\alpha}{1-\beta} \\ \|\bar{\mathbf{x}}(t) - \mathbf{x}_*\| &\leqslant (1-c\alpha)^t \|\bar{\mathbf{x}}(0) - \mathbf{x}_*\| + C\alpha \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|. \end{aligned} \tag{4.1.12}$$

Combining these two estimates then gives the $O(\alpha)$-convergence result of the DGD. By contrast with the approach that utilizes the separate two estimates from (4.1.11), our methodology relies on the sequential estimate presented in (4.1.9), which can be interpreted as the sum of the two aforementioned estimates from (4.1.11). However, this approach imposes constraints on obtaining precise estimates for both $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2$ and $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2$. Overcoming this limitation may prove challenging if the proof commences with the estimate (4.1.8) for handling the projection operator.

To circumvent this challenge, we have devised a new approach that analyzes the convergence of the sequence $x_k(t)$ by partitioning the coordinates into two distinct segments: one influenced by the projection operator and the other unaffected by it. Further details can be found in Section 4.7. While our current work focuses on the domain $\Omega = \mathbb{R}^{d-1} \times \mathbb{R}+$, we are confident that the fundamental concept of the argument can be extended to more general domains, including $\mathbb{R}^q \times \mathbb{R}+^{d-q}$ and any smooth convex domain.

Throughout this section, we make the following assumptions for functions and an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Assumption 4.1.1.

(A) $\Omega \subseteq \mathbb{R}^d$ is closed and convex

(B) For each $i \in \{1, \cdots, m\}$, the local function f_i is L_i-smooth for some $L_i > 0$.

(C) The total cost function f is μ-strongly convex for some $\mu > 0$.

(D) The communication graph $\mathcal{G}$ is connected.

(E) The mixing matrix W is doubly stochastic. In addition, $w_{ii} > 0$ for all $i \in \mathcal{V}$.

Next, we define the following constants which are frequently used throughout this section.

- R, the uniform upper bound for the quantities $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2$ for $t \geqslant 0$
- $D = \max_{1 \leqslant i \leqslant m} \|\nabla f_i(x_*)\|$ and $L = \max_{1 \leqslant i \leqslant m} L_i$
- Fix a variable $\delta > 0$ such that $(1+\delta)\beta^2 < 1$ and let $\tilde{\beta} := (1+\delta)\beta^2$
- $c_1 := 3L^2\left(1 + \frac{1}{\delta}\right),\ c_2 := 3nD^2\left(1 + \frac{1}{\delta}\right),\ c_3 := c_1 + L^2,\ c_4 := \frac{4L^2}{\mu}$

4.2 Preliminary Results

In this section, we introduce some estimates that are frequently used to obtain consensus and convergence results for the projected DGD.

Lemma 4.2.1. *Suppose that* $(A), (B)$ *and* (C) *of Assumption 4.1.1 hold. If* $\alpha(t) \leqslant \frac{2}{L+\mu}$ *for all* $t \geqslant 0$*, then we have*

$$\left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{n} \sum_{i=1}^{n} (\nabla f_i(\bar{x}(t)) - \nabla f_i(x_*)) \right\|^2 \leqslant \left(1 - \frac{\mu\alpha(t)}{2}\right)^2 \|\bar{x}(t) - x_*\|^2. \quad (4.2.1)$$

Proof. We expand the left hand side in (4.2.1) as

$$\begin{aligned} \|\bar{x}(t) - x_*\|^2 - 2\alpha(t)\Big\langle \bar{x}(t) - x_*, \nabla f(\bar{x}(t)) - \nabla f(x_*)\Big\rangle \\ + \alpha(t)^2 \|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2 . \end{aligned} \quad (4.2.2)$$

Since f is L-smooth and μ-strongly convex, it follows by Lemma 2.1.9 that

$$\Big\langle \bar{x}(t) - x_*, \ \nabla f(\bar{x}(t)) - \nabla f(x_*) \Big\rangle \geqslant \frac{L\mu}{L+\mu} \|\bar{x}(t) - x_*\|^2 + \frac{1}{L+\mu} \|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2.$$

Putting the above inequality in (4.2.2), we get

$$\begin{aligned} &\left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{n} \sum_{i=1}^{n} (\nabla f_i(\bar{x}(t)) - \nabla f_i(x_*)) \right\|^2 \\ &\leqslant \left(1 - \frac{2L\mu\alpha(t)}{L+\mu}\right) \|\bar{x}(t) - x_*\|^2 + \alpha(t)\left(\alpha(t) - \frac{2}{L+\mu}\right) \|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2. \end{aligned}$$

Using the assumption $\alpha(t) \leqslant \frac{2}{L+\mu}$ and $2L \geqslant L + \mu$, we then have

$$\begin{aligned} \left\| \bar{x}(t) - x_* - \frac{\alpha(t)}{n} \sum_{i=1}^{n} (\nabla f_i(\bar{x}(t)) - \nabla f_i(x_*)) \right\|^2 &\leqslant \left(1 - \frac{2L\mu\alpha(t)}{L+\mu}\right) \|\bar{x}(t) - x_*\|^2 \\ &\leqslant \left(1 - \frac{\mu\alpha(t)}{2}\right)^2 \|\bar{x}(t) - x_*\|^2. \end{aligned}$$

The proof is done. □

Lemma 4.2.2. *Suppose that (B) of Assumption 4.1.1 holds. For* $(x_1, \cdots, x_n) \in \mathbb{R}^{dn}$ *and* $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ *we have*

$$\sum_{i=1}^{n} \|\nabla f_i(x_i) - \nabla f_i(\bar{x})\|^2 \leqslant L^2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \quad (4.2.3)$$

and

$$\sum_{i=1}^{n}\left\|\nabla f_i(x_i)-\frac{1}{n}\sum_{l=1}^{n}\nabla f_l(x_l)\right\|^2 \leqslant 3L^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2+3L^2\|\bar{\mathbf{x}}-\mathbf{x}_*\|^2+3nD^2. \quad (4.2.4)$$

Proof. Since f_i is L_i-smooth function, it follows that

$$\sum_{i=1}^{n}\|\nabla f_i(x_i)-\nabla f_i(\bar{x})\|^2 \leqslant L^2\sum_{i=1}^{n}\|x_i-\bar{x}\|^2,$$

which directly implies (4.2.3). Next, we prove (4.2.4). Note that for any $a = (a_1,\cdots,a_n)\in\mathbb{R}^n$, we have

$$\begin{aligned}\sum_{i=1}^{n}\left\|a_i-\frac{1}{n}\sum_{l=1}^{n}a_l\right\|^2 &= \sum_{i=1}^{n}\left(\|a_i\|^2-2\left\langle a_i,\frac{1}{n}\sum_{l=1}^{n}a_l\right\rangle+\frac{1}{n^2}\left\|\sum_{l=1}^{n}a_l\right\|^2\right)\\ &= \sum_{i=1}^{n}\|a_i\|^2+\left(\frac{1}{n^2}-\frac{2}{n}\right)\left\|\sum_{l=1}^{n}a_l\right\|^2\\ &\leqslant \sum_{i=1}^{n}\|a_i\|^2.\end{aligned}$$

Using this, it follows that

$$\sum_{i=1}^{n}\left\|\nabla f_i(x_i)-\frac{1}{n}\sum_{l=1}^{n}\nabla f_l(x_l)\right\|^2 \leqslant \sum_{i=1}^{n}\|\nabla f_i(x_i(t))\|^2. \quad (4.2.5)$$

By the triangle inequality, one has

$$\|\nabla f_i(x_i)\|^2 \leqslant 3\|\nabla f_i(x_i)-\nabla f_i(\bar{x})\|^2+3\|\nabla f_i(\bar{x})-\nabla f_i(x_*)\|^2+3\|\nabla f_i(x_*)\|^2.$$

This, together with L-smoothness and (4.2.3), gives

$$\sum_{i=1}^{n}\|\nabla f_i(x_i)\|^2 \leqslant 3L^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2+3L^2\|\bar{\mathbf{x}}-\mathbf{x}_*\|^2+3nD^2.$$

Combining this with (4.2.5) gives the desired estimate. □

Lemma 4.2.3. *Suppose that* (*A*), (*B*) *and* (*C*) *of Assumption 4.1.1 hold. If the diminishing sequence* $\{\alpha(t)\}_{t\geqslant 0}$ *satisfy* $\alpha(0)\leqslant\frac{2}{L+\mu}$*, then the sequence* $\{x_i(t)\}_{t\geqslant 0}$ *generating by* (4.1.1) *for all* $1\leqslant i\leqslant n$ *satisfies the following inequality*

$$\begin{aligned}&n\left\|\bar{x}(t)-x_*-\alpha(t)\left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i(t))-\nabla f(x_*)\right)\right\|^2\\ &\leqslant\left(1-\frac{\mu\alpha(t)}{2}\right)\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+\left(L^2\alpha(t)^2+\frac{4L^2\alpha(t)}{\mu}\right)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2.\end{aligned} \quad (4.2.6)$$

Proof. By Young's inequality, for any x and y in $\mathbb{R}^n$, the following inequality holds for all $\eta > 0$:

$$\|x+y\|^2 \leqslant (1+\eta)\|x\|^2 + \left(1+\frac{1}{\eta}\right)\|y\|^2 \tag{4.2.7}$$

For notational convenience, we let $H(t) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i(t))$. Using (4.2.7), we obtain the following inequality:

$$\begin{aligned}
&\|\bar{x}(t) - x_* - \alpha(t)\left(H(t) - \nabla f(x_*)\right)\|^2 \\
&= \|\bar{x}(t) - x_* - \alpha(t)\left(\nabla f(\bar{x}(t)) - \nabla f(x_*)\right) + \alpha(t)\left(\nabla f(\bar{x}(t)) - H(t)\right)\|^2 \\
&\leqslant (1+\eta)\left\|\bar{x}(t) - x_* - \alpha(t)\left(\nabla f(\bar{x}(t)) - \nabla f(x_*)\right)\right\|^2 \\
&\qquad\qquad + \left(1+\frac{1}{\eta}\right)\alpha(t)^2\left\|\nabla f(\bar{x}(t)) - H(t)\right\|^2.
\end{aligned}$$

Now we estimate the right-hand side of the last inequality. By Lemma 4.2.1, we can establish the following inequality:

$$\|\bar{x}(t) - x_* - \alpha(t)\left(\nabla f(\bar{x}(t)) - \nabla f(x_*)\right)\|^2 \leqslant \left(1 - \frac{\mu\alpha(t)}{2}\right)^2 \|\bar{x}(t) - x_*\|^2. \tag{4.2.8}$$

Utilizing the Cauchy-Schwarz inequality and Lemma 4.2.2, we can also derive the following result:

$$\begin{aligned}
\|\nabla f(\bar{x}(t)) - H(t)\|^2 &= \frac{1}{n^2}\left\|\sum_{i=1}^n \left(\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\right)\right\|^2 \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\bar{x}(t)) - \nabla f_i(x_i(t))\right\|^2. \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \|\bar{x}(t) - x_i(t)\|^2.
\end{aligned} \tag{4.2.9}$$

Let's set $\eta = \frac{\mu\alpha(t)}{4}$. Combining (4.2.8) and (4.2.9), it follows that

$$\begin{aligned}
&n\left\|\bar{x}(t) - x_* - \frac{\alpha(t)}{n}\sum_{l=1}^n \left(\nabla f_l(x_l(t)) - \nabla f_l(x_*)\right)\right\|^2 \\
&\leqslant n\left(1+\frac{\mu\alpha(t)}{4}\right)\left(1-\frac{\mu\alpha(t)}{2}\right)^2\|\bar{x}(t) - x_*\|^2 \\
&\qquad\qquad + n\left(1+\frac{4}{\mu\alpha(t)}\right)\frac{L^2\alpha(t)^2}{n}\sum_{i=1}^n \|\bar{x}(t) - x_i(t)\|^2 \\
&= \left(1+\frac{\mu\alpha(t)}{4}\right)\left(1-\frac{\mu\alpha(t)}{2}\right)^2\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 + \left(1+\frac{4}{\mu\alpha(t)}\right)L^2\alpha(t)^2\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2.
\end{aligned}$$

Here we used

$$n\|\bar{x}(t)-x_*\|^2 = \|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2 \text{ and } \sum_{i=1}^{n}\|\bar{x}(t)-x_i(t)\|^2 = \|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2$$

for the last equality. Note that since $\alpha(t) \leqslant \frac{2}{\mu}$ by the assumption, it follows that

$$\begin{aligned}\Big(1+\frac{\mu\alpha(t)}{4}\Big)\Big(1-\mu\alpha(t)+\frac{\mu^2\alpha(t)^2}{4}\Big) &= 1-\frac{3}{4}\mu\alpha(t)+\frac{\mu^3}{16}\alpha(t)^3\\ &= 1-\frac{\mu\alpha(t)}{2}-\Big(\frac{\mu\alpha(t)}{4}-\frac{\mu^3\alpha(t)^3}{16}\Big)\\ &\leqslant 1-\frac{\mu\alpha(t)}{2}.\end{aligned}$$

Using this result, we obtain the desired estimate. The proof is done. □

4.3 Sequential Estimates

In this section, we establish sequential estimates crucial for deriving the convergence results of the algorithm (4.1.1). As previously discussed in Section 4.1, the presence of the projection operator complicates the process of averaging Equation (4.1.1) to obtain (4.1.2). To address this challenge, we estimate the quantity $\|\mathbf{x}(t+1)-\mathbf{x}_*\|^2$ instead of $\|\bar{\mathbf{x}}(t+1)-\mathbf{x}_*\|$ to analyze the sequence of (4.1.1). Specifically, we aim to establish an estimate of $\|\mathbf{x}(t+1)-\mathbf{x}_*\|^2$ in terms of $\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2$ and $\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2$ by utilizing the contraction property of the projection operator. We start this section by deriving an estimate of $\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2$.

Proposition 4.3.1. *Suppose that Assumption 4.1.1 holds. If $\{\alpha(t)\}_{t\geqslant 0}$ satisfies $\alpha(0) \leqslant \frac{2}{L+\mu}$, then the sequence $\{x_i(t)\}_{t\geqslant 0}$ generated by (4.1.1) for all $1 \leqslant i \leqslant n$ satisfies the following inequality.*

$$\begin{aligned}&\|\mathbf{x}(t+1)-\bar{\mathbf{x}}(t+1)\|^2\\ &\leqslant (c_1\alpha(t)^2+\tilde{\beta})\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2 + c_1\alpha(t)^2\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2 + c_2\alpha(t)^2.\end{aligned}$$

Proof. For the reader's convenience, we recall the DPG algorithm as

$$x_i(t+1) = \mathcal{P}_\Omega\Bigg[\sum_{j=1}^{n} w_{ij}x_j(t) - \alpha(t)\nabla f_i(x_i(t))\Bigg].$$

Additionally, we recall Young's inequality, which is useful for our derivations: For any vectors u and v in $\mathbb{R}^d$, we have

$$\|u+v\|^2 \leqslant (1+\delta)\|u\|^2 + \left(1+\frac{1}{\delta}\right)\|v\|^2$$

Using the DPG algorithm, we can write $\|x_i(t+1)-\bar{x}(t+1)\|^2$ as

$$\left\|\mathcal{P}_\Omega\Bigg[\sum_{j=1}^n w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))\Bigg]-\frac{1}{n}\sum_{k=1}^n\mathcal{P}_\Omega\Bigg[\sum_{j=1}^n w_{kj}x_j(t)-\alpha(t)\nabla f_i(x_k(t))\Bigg]\right\|^2.$$

Summing this $i=1$ to n and applying Lemma 2.2.3, it follows that

$$\begin{aligned}&\sum_{i=1}^n\|x_i(t+1)-\bar{x}(t+1)\|^2\\ &\leqslant\sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}\left(x_j(t)-\bar{x}(t)\right)-\alpha(t)\nabla f_i(x_i(t))+\frac{\alpha(t)}{n}\sum_{l=1}^n\nabla f_l(x_l(t))\right\|^2.\end{aligned}\tag{4.3.1}$$

where Applying Young's inequality with

$$u=\sum_{j=1}^n w_{ij}\left(x_j(t)-\bar{x}(t)\right),\ v=\alpha(t)\left(\nabla f_i(x_i(t))+\frac{1}{n}\sum_{l=1}^n\nabla f_l(x_l(t))\right),$$

the last term is bounded as

$$(1+\delta)\left\|W(\mathbf{x}(t)-\bar{\mathbf{x}}(t))\right\|^2+\sum_{i=1}^n\left(1+\frac{1}{\delta}\right)\alpha(t)^2\left\|\nabla f_i(x_i(t))+\frac{1}{n}\sum_{l=1}^n\nabla f_l(x_l(t))\right\|^2.$$

Here we used the fact that $\sum_{i=1}^n\|x_i(t)-\bar{x}(t)\|^2=\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2$. Now, applying Lemma 2.3.5, we bound the first term as:

$$(1+\delta)\beta^2\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2,$$

Using Lemma 4.2.2, the second term is bounded as

$$\left(1+\frac{1}{\delta}\right)\alpha(t)^2\left(3L^2\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2+3L^2\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+3nD^2\right).$$

Putting all together, we obtain the desired estimate:

$$\begin{aligned}&\sum_{i=1}^n\|x_i(t+1)-\bar{x}(t+1)\|^2\\ \leqslant&(1+\delta)\left\|W(\mathbf{x}(t)-\bar{\mathbf{x}}(t))\right\|^2+\sum_{i=1}^n\left(1+\frac{1}{\delta}\right)\alpha(t)^2\left\|\nabla f_i(x_i(t))+\frac{1}{n}\sum_{l=1}^n\nabla f_l(x_l(t))\right\|^2\\ \leqslant&\left(3L^2\left(1+\frac{1}{\delta}\right)\alpha(t)^2+(1+\delta)\beta^2\right)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2\\ &\qquad+3L^2\left(1+\frac{1}{\delta}\right)\alpha(t)^2\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+3nD^2\left(1+\frac{1}{\delta}\right)\alpha(t)^2\\ =&(c_1\alpha(t)^2+(1+\delta)\beta^2)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2+c_1\alpha(t)^2\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+c_2\alpha(t)^2,\end{aligned}$$

which completes the proof. □

Next, we give an estimate of the quantity $\|\mathbf{x}(t)-\mathbf{x}_*\|^2$. Since $\frac{1}{n}\sum_{i=1}^n x_i(t)-\bar{x}(t)=0$, it follows directly that

$$\|\mathbf{x}(t)-\mathbf{x}_*\|^2=\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2+\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2. \tag{4.3.2}$$

Using the above relation, we state the following proposition.

Proposition 4.3.2. *Suppose that Assumptions 4.1.1 and 4.4.3 hold. If the diminishing sequence* $\{\alpha(t)\}_{t\geqslant 0}$ *satisfy* $\alpha(0)\leqslant\frac{2}{L+\mu}$, *then the sequence* $\{x_i(t)\}_{t\geqslant 0}$ *generated by* (4.1.1) *for all* $1\leqslant i\leqslant n$ *satisfes the following inequality.*

$$\begin{aligned}&\|\mathbf{x}(t+1)-\bar{\mathbf{x}}(t+1)\|^2+\|\bar{\mathbf{x}}(t+1)-\mathbf{x}_*\|^2\\&\leqslant\left(c_3\alpha(t)^2+c_4\alpha(t)+\tilde{\beta}\right)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2\\&\qquad\qquad+\left(1-\frac{\mu}{2}\alpha(t)+c_1\alpha(t)^2\right)\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+c_2\alpha(t)^2.\end{aligned}$$

Proof. By the DPG algorithm and using the contraction property of the projection operator in Lemma 2.2.3, we deduce

$$\begin{aligned}\|x_i(t+1)-x_*\|^2&=\left\|\mathcal{P}_\Omega\Big[\sum_{j=1}^n w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))\Big]-\mathcal{P}_\Omega\left[x_*-\alpha(t)\nabla f(x_*)\right]\right\|^2\\&\leqslant\left\|\sum_{j=1}^n w_{ij}x_j(t)-x_*-\alpha(t)\left(\nabla f_i(x_i(t))-\nabla f(x_*)\right)\right\|^2.\end{aligned}$$

For the first equality, we used the fact that $x_*=\mathcal{P}_\Omega\left[x_*-\alpha(t)\nabla f(x_*)\right]$, which follows from $x_*=\arg\min_{x\in\Omega}f(x)$. Summing up the above inequality from $i=1$ to n, we have

$$\sum_{i=1}^n\|x_i(t+1)-x_*\|^2\leqslant\sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}x_j(t)-x_*-\alpha(t)\left(\nabla f_i(x_i(t))-\nabla f(x_*)\right)\right\|^2. \tag{4.3.3}$$

Denoting $P(t)$ as

$$P(t)=\bar{x}(t)-\frac{\alpha(t)}{n}\sum_{l=1}^n\nabla f_l(x_l(t)),$$

we find the following identity of the right-hand side of (4.3.3):

$$\begin{aligned}&\sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}x_j(t)-x_*-\alpha(t)\left(\nabla f_i(x_i(t))-\nabla f(x_*)\right)\right\|^2\\&=\sum_{i=1}^n\left\|P(t)-(x_*-\alpha(t)\nabla f(x_*))-P(t)+\sum_{j=1}^n w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))\right\|^2\\&=n\left\|P(t)-(x_*-\alpha(t)\nabla f(x_*))\right\|^2+\sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))-P(t)\right\|^2.\end{aligned} \tag{4.3.4}$$

For the last equality, we used

$$\begin{aligned}&\sum_{i=1}^{n}\Big(\sum_{j=1}^{n} w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))-P(t)\Big)\\&=\sum_{j=1}^{n} x_j(t)-\alpha(t)\sum_{i=1}^{n}\nabla f_i(x_i(t))-nP(t)=0,\end{aligned}$$

and the fact that $\sum_{i=1}^n \|a+b_i\|^2 = n\|a\|^2+\sum_{i=1}^n b_i$ if $\sum_{i=1}^n b_i = 0$. Now we estimate the first term on the right-hand side of the last equality in (4.3.4) as follows:

$$\begin{aligned}&n\Bigg\|P(t)-(x_*-\alpha(t)\nabla f(x_*))\Bigg\|^2\\&=n\left\|\bar{x}(t)-x_*-\alpha(t)\left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i(t))-\nabla f(x_*)\right)\right\|^2\\&\leqslant\left(1-\frac{\mu\alpha(t)}{2}\right)\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+\left(L^2\alpha(t)^2+\frac{4L^2\alpha(t)}{\mu}\right)\|\bar{\mathbf{x}}(t)-\mathbf{x}(t)\|^2,\end{aligned}$$

where we apply Lemma 4.2.3 for the last inequality. The second term of the right-hand side of the last equality in (4.3.4) is equal to the right hand side of (4.3.1) and the same estimate gives

$$\begin{aligned}&\sum_{i=1}^{n}\Bigg\|\sum_{j=1}^{n} w_{ij}x_j(t)-\alpha(t)\nabla f_i(x_i(t))-P(t)\Bigg\|^2\\&=\sum_{i=1}^{n}\Bigg\|\sum_{j=1}^{n} w_{ij}\Big(x_j(t)-\bar{x}(t)\Big)-\alpha(t)\Big(\nabla f_i(x_i(t))-\frac{1}{n}\sum_{l=1}^{n}\nabla f_l(x_l(t))\Big)\Bigg\|^2\\&\leqslant(c_1\alpha(t)^2+(1+\delta)\beta^2)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2+c_1\alpha(t)^2\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+c_2\alpha(t)^2.\end{aligned}$$

Putting the above two estimates in the last term of (4.3.4), we get

$$\begin{aligned}\sum_{i=1}^{n}\|x_i(t+1)-x_*\|^2&=\|\mathbf{x}(t+1)-\bar{\mathbf{x}}(t+1)\|^2+\|\bar{\mathbf{x}}(t+1)-\mathbf{x}_*\|^2.\\&\leqslant\left(\Big(c_1+L^2\Big)\alpha(t)^2+\frac{4L^2\alpha(t)}{\mu}+\tilde{\beta}\right)\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2\\&\quad+\left(1-\frac{\mu\alpha(t)}{2}+c_1\alpha(t)^2\right)\|\bar{\mathbf{x}}(t)-\mathbf{x}_*\|^2+c_2\alpha(t)^2,\end{aligned}$$

which completes the proof. □

4.4 Uniform Boundedness

In this section, we introduce the uniform boundedness of the sequence $\{x_i(t)\}_{t\geqslant 0}$ in the sense that $\|\mathbf{x}(t)-\bar{\mathbf{x}}(t)\|^2\leqslant R$ for all $t\geqslant 0$.

Theorem 4.4.1 (Conditions for uniform bounedness)**.** *There exists a constant $R>0$ such that*

$$\|\mathbf{x}(t)-\mathbf{x}_*\|^2\leqslant R$$

holds for all $t\geqslant 0$ if at least one of the following statements holds true:

1. *Ω is bounded (no restriction on the stepsize).*
2. *Each local cost function f_i is convex and satisfies (B) of Assumption 4.1.1. In addition, the stepsize is constant, i.e., $\alpha(t)\equiv\alpha$, satisfying $\alpha\leqslant\frac{1+\lambda_n(W)}{L}$.*
3. *Each local cost function f_i satisfies (B) of Assumption 4.1.1 and the total cost function f satisfies (C) of Assumption 4.1.1. In addition, the stepsize $\{\alpha(t)\}_{t\geqslant 0}$ is non-increasing and satisfies*

$$\alpha(t)<\min\left\{Z,\frac{\mu}{4\tilde{c}_1},\frac{2}{L+\mu}\right\}.$$

Here we have set the positive constant Z by

$$Z:=\frac{1}{2\tilde{c}_3}\Big[-\Big(\tilde{c}_4+\frac{\mu}{4}\Big)+\sqrt{\Big(\tilde{c}_4+\frac{\mu}{4}\Big)^2+4\tilde{c}_3(1-\beta)}\Big]\leqslant 1-\beta,$$

where the constants $\tilde{c}_1,\tilde{c}_2,\tilde{c}_3,\tilde{c}_4$ are defined as follows:

$$\tilde{c}_1:=\frac{3L^2}{1-\beta},\ \tilde{c}_2:=\frac{3nD^2}{1-\beta},\ \tilde{c}_3:=\tilde{c}_1+L^2,\ \tilde{c}_4:=\frac{4L^2}{\mu}. \tag{4.4.1}$$

The uniform boundedness of the sequence is straightforward in the first case, where Ω is assumed to be bounded. We will now proceed to prove the theorem for the second case.

Proof of Theorem 4.4.1 for case 2. Consider the following functional $E_\alpha:(\mathbb{R}^d)^n\to\mathbb{R}$ defined as

$$E_\alpha(x)=\frac{1}{2}\Big(\sum_{k=1}^n\|x_k\|^2-\sum_{k=1}^n\sum_{j=1}^n w_{kj}\langle x_k,x_j\rangle\Big)+\alpha\sum_{k=1}^n f_k(x_k).$$

Then

$$x(t+1)=P_{\Omega^n}\Big(x(t)-\nabla E_\alpha(x(t))\Big).$$

The function E_α is convex and smooth with constant $1 - \lambda_n(W) + \alpha L$ (refer to [49]). Then, we may use the general result for the projected gradient descent (see e.g., [6]) to conclude that the sequence $\{x(t)\}_{t\geqslant 0}$ is uniformly bounded if

$$1 \leqslant \frac{2}{1 - \lambda_n(W) + \alpha L},$$

which is equivalent to $\alpha \leqslant \frac{1+\lambda_n(W)}{L}$. □

Next, we consider the third case of Theorem 4.4.1. To handle this case of Theorem 4.4.1, we set

$$A(t) = \|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2, \quad B(t) = \|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2, \quad C(t) = \|\mathbf{x}(t) - \mathbf{x}_*\|^2. \tag{4.4.2}$$

We can easily check that $C(t) = A(t) + B(t)$ using (4.3.2). In the following lemma, we find a sequential inequality for the sequence $\{C(t)\}_{t\geqslant 0}$ and find the uniform boundedness of $\{C(t)\}_{t\geqslant 0}$, which also implies the uniform boundedness of $\{A(t)\}_{t\geqslant 0}$ and $\{B(t)\}_{t\geqslant 0}$. It contains the proof of Theorem 4.4.1 for case 3.

Lemma 4.4.2. *Suppose that Assumption 4.1.1 holds and the step size $\{\alpha(t)\}_{t\geqslant 0}$ is nonincreasing and satisfies*

$$\alpha(0) \leqslant \min\left\{ Z, \frac{\mu}{4c_1}, \frac{2}{L+\mu} \right\} \tag{4.4.3}$$

where

$$Z := \frac{1}{2c_3}\left[-\left(c_4 + \frac{\mu}{4}\right) + \sqrt{\left(c_4 + \frac{\mu}{4}\right)^2 + 4c_3(1-\tilde{\beta})} \right].$$

Suppose also that $\tilde{\beta} := (1+\delta)\beta^2 < 1$. Then the sequence $\{x_i(t)\}_{t\in\mathbb{N}_0}$ generated by (4.1.1) *satisfies the following statements.*

1. *We have*
$$C(t+1) \leqslant \left(1 - \frac{\mu}{4}\alpha(t)\right) C(t) + c_2\alpha(t)^2. \tag{4.4.4}$$

2. *There exists $R > 0$ such that*
$$C(t) \leqslant R, \ \text{for all } t \in \mathbb{N}.$$
In fact, we may set $R = \max\left\{ \frac{4c_2\alpha(0)}{\mu}, C(0) \right\}$

Proof. We first prove (4.4.4). By Proposition 4.3.2, we have the following estimate

$$C(t) \leqslant (c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})A(t) + \left(1 - \frac{\mu}{2}\alpha(t) + c_1\alpha(t)^2\right)B(t) + c_2\alpha(t)^2.$$

Note that since $\alpha(0) \leqslant \frac{\mu}{4c_1}$ by (4.4.3), it follows that $\mu/2 - c_1\alpha(0) \geqslant \frac{\mu}{4}$. Suppose that the following inequality hold.

$$1 - \frac{\mu}{2}\alpha(t) + c_1\alpha(t)^2 < 1 - \frac{\mu}{4}\alpha(t), \tag{4.4.5}$$

$$c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta} < 1 - \frac{\mu}{4}\alpha(t). \tag{4.4.6}$$

Then we obtain (4.4.4) as follows.

$$\begin{aligned} C(t+1) &\leqslant (c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})A(t) + \left(1 - \frac{\mu}{2}\alpha(t) + c_1\alpha(t)^2\right)B(t) + c_2\alpha(t)^2 \\ &\leqslant \left(1 - \frac{\mu}{4}\alpha(t)\right)C(t) + c_2\alpha(t)^2. \end{aligned}$$

Now we show that (4.4.5) and (4.4.6) hold under the assumption (4.4.3).

Since $\{\alpha(t)\}_{t\geqslant 0}$ is diminishing sequence and $\alpha(0) \leqslant \frac{\mu}{4c_1}$, we obtain (4.4.5) as follows:

$$1 - \frac{\mu}{2}\alpha(t) + c_1\alpha(t)^2 = 1 - \left(\frac{\mu}{2} - c_1\alpha(t)\right)\alpha(t) \leqslant 1 - \frac{\mu}{4}\alpha(t).$$

We note that (4.4.6) is equivalent to

$$c_3\alpha(t)^2 + \left(c_4 + \frac{\mu}{4}\right)\alpha(t) + \tilde{\beta} - 1 \leqslant 0. \tag{4.4.7}$$

Therefore, the following inequality is a sufficient condition for (4.4.7):

$$\alpha(t) \leqslant \alpha(0) \leqslant Z := \frac{1}{2c_3}\left[-\left(c_4 + \frac{\mu}{4}\right) + \sqrt{\left(c_4 + \frac{\mu}{4}\right)^2 + 4c_3(1-\tilde{\beta})}\right].$$

Since $\tilde{\beta} < 1$, we have $Z > 0$. This proves the first estimate of the lemma.

In order to show the second estimate, we argue by induction. Fix a value $R > 0$ and assume that $C(t) \leqslant R$ for some $t \in \mathbb{N}_0$. Then, it follows from (4.4.4) that

$$C(t+1) \leqslant (1 - \frac{\mu}{4}\alpha(t))R + c_2\alpha(t)^2 = R - (\frac{\mu}{4}R - c_2\alpha(t))\alpha(t).$$

If we set

$$R = \frac{4c_2\alpha(0)}{\mu},$$

then we have

$$\frac{\mu}{4}R - c_2\alpha(t) \geqslant \frac{\mu}{4}R - c_2\alpha(0) = 0.$$

This implies $C(t+1) \leqslant R$. Therefore we have $C(t) \leqslant R$ for any $t \geqslant 0$. The proof is done. □

The following assumption formulates the above uniform boundedness property:

Assumption 4.4.3. There exists a constant $R > 0$ such that

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2 \leqslant R$$

holds for all $t \geqslant 0$.

Although the result of this assumption is proved in Theorem 4.4.1, we proposed this assumption because the result may hold for larger ranges of $\alpha(t)$ than that guaranteed by Theorem 4.4.1. Proving a sharper range of $\alpha(t)$ for the uniform boundedness property would be an interesting future work.

4.5 Consensus Estimates

In this section, we state the consensus results for (4.1.1) both for the constant step size and diminishing step size.

Theorem 4.5.1 (consensus)**.** *Suppose that Assumptions 4.1.1 and 4.4.3 hold. If* $\{\alpha(t)\}_{t\geqslant 0}$ *satisfies* $\alpha(0) \leqslant \frac{2}{L+\mu}$, *then we have*

$$\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2 \leqslant \beta^t \|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{J\alpha(t)^2}{(1-\beta)^2}.$$

Here for constant step size, i.e. $\alpha(t) \equiv \alpha$, *we set*

$$J := 3(L^2R^2 + nD^2).$$

For a diminishing step size, we set

$$J := 3(L^2R^2 + nD^2) \cdot \sup_{s\geqslant 0} \frac{\alpha(0)^2\beta^s + \alpha([s/2])^2}{\alpha(s)^2}.$$

Proof. Let $\delta = \frac{1}{\beta} - 1$ so that $\tilde{\beta} = \beta < 1$. Then, the estimate from Proposition 4.3.1 becomes:

$$\begin{aligned}&\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\|^2\\ &\leqslant (\tilde{c}_1\alpha(t)^2 + \beta)\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2 + \tilde{c}_1\alpha(t)^2\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 + \tilde{c}_2\alpha(t)^2.\end{aligned}$$

Since $\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2 < R$, it follows that

$$\begin{aligned}&\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\|^2\\ &\leqslant \beta\|\mathbf{x}(t) - \bar{\mathbf{x}}(t)\|^2 + \frac{3\alpha(t)^2}{1-\beta}(L^2R^2 + nD^2)\\ &\leqslant \beta^{t+1}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3}{1-\beta}(L^2R^2 + nD^2)\sum_{s=0}^{t}\alpha(s)^2\beta^{t-s}.\end{aligned} \tag{4.5.1}$$

For a constant stepszie $\alpha(t) \equiv \alpha$, we can estimate (4.5.1) as

$$\begin{aligned}
\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\|^2 &\leqslant \beta^{t+1}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3\alpha^2}{1-\beta}(L^2R^2 + nD^2)\sum_{s=0}^{t}\beta^{t-s} \\
&\leqslant \beta^{t+1}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3\alpha^2}{(1-\beta)^2}(L^2R^2 + nD^2) \\
&= \beta^{t+1}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{J\alpha^2}{(1-\beta)^2},
\end{aligned}$$

where $J = 3(L^2R^2 + nD^2)$. For a diminishing step size, we have

$$\begin{aligned}
\sum_{s=0}^{t}\alpha(s)^2\beta^{t-s} &= \sum_{s=0}^{\lfloor t/2\rfloor - 1}\alpha(s)^2\beta^{t-s} + \sum_{s=\lfloor t/2\rfloor}^{t}\alpha(s)^2\beta^{t-s} \\
&\leqslant \alpha(0)^2\frac{\beta^t}{1-\beta} + \alpha(\lfloor t/2\rfloor)^2\frac{1}{1-\beta}.
\end{aligned} \tag{4.5.2}$$

Inserting this inequality to (4.5.1), we have

$$\|\mathbf{x}(t+1) - \bar{\mathbf{x}}(t+1)\|^2 \leqslant \beta^{t+1}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{J\alpha(t)^2}{(1-\beta)^2},$$

where

$$J = 3(L^2R^2 + nD^2)\cdot\sup_{s\geqslant 0}\frac{\alpha(0)^2\beta^s + \alpha([s/2])^2}{\alpha(s)^2}.$$

The proof is done. □

4.6 Convergence Analysis

In the previous section, we have achieved the consensus result, and thanks to this, we can concentrate on the convergence of $\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|$. In this section, we establish the convergence results for the sequence $\{\bar{x}(t)\}_{t\geqslant 0}$ towards the optimal point. Before stating the results, we further introduce the following constants:

- $G_1 = 2R\Big(\tilde{c}_3\alpha(0)^2 + \tilde{c}_4\alpha(0) + \beta\Big)$
- $G_2 = \Big[(\tilde{c}_3\alpha(0)^2 + \tilde{c}_4\alpha(0) + \beta)\frac{2J^2}{(1-\beta)^2} + \tilde{c}_1R + \tilde{c}_2\Big]$
- $\tilde{c}_1 := \frac{3L^2}{1-\beta},\ \tilde{c}_2 := \frac{3nD^2}{1-\beta},\ \tilde{c}_3 := \tilde{c}_1 + L^2,\ \tilde{c}_4 := \frac{4L^2}{\mu}$

These constants are derived by substituting $\delta = (1-\beta)/\beta$ into the expressions defined on page 35. The following convergence result holds when the step size is given by a constant.

Theorem 4.6.1 (convergence result for the constant step size). *Suppose that Assumptions 4.1.1 and 4.4.3 hold. If the step size is given by a constant $\alpha > 0$ such that $\alpha \leqslant \frac{2}{L+\mu}$, then we have*

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 \leqslant \Big(1 - \frac{\mu\alpha}{2}\Big)^t \|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|^2 + \frac{2G_2}{\mu}\alpha + \frac{2}{\mu\alpha}\bigg(\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1} + \beta^{\frac{t-1}{2}}\bigg).$$

Theorem 4.6.1 implies the sequence generated by (4.1.1) converges to an $O(\sqrt{\alpha})$-neighborhood of the optimal point exponentially fast. In the below, we provide the convergence results when the step size is given by the diminishing step size $\alpha(t) = v/(t+w)^p$ for $p \in (0,1]$. We first consider the case $p \in (0,1)$.

Theorem 4.6.2 (convergence result for the diminishing step size). *Suppose that Assumptions 4.1.1 and 4.4.3 hold. Let $p \in (0,1)$ and assume that $\alpha(t) = \frac{v}{(t+w)^p}$ with $v, w > 0$ satisfying*

$$\alpha(0) = \frac{v}{w^p} \leqslant \frac{2}{L+\mu}.$$

Then we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 \leqslant \frac{2eQ\Big(\rho G_1 + G_2\Big)v}{\mu}(\lfloor t/2 \rfloor + w - 1)^{-p} + \mathcal{R}_1(t) + \mathcal{R}_2(t),$$

where $Q = \Big(\frac{w+1}{w}\Big)^{2p}$, $\rho = \sup_{t \geqslant 0} \beta^t/\alpha(t)^2$ and

$$\mathcal{R}_1(t) = e^{-\sum_{s=0}^{t-1} \frac{\mu v}{2(s+w)^p}} \|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|^2,$$
$$\mathcal{R}_2(t) = Q\Big(\rho G_1 + G_2\Big)v^2 e^{-\frac{\mu v t}{4(t+w)^p}} \sum_{s=1}^{\lfloor t/2 \rfloor - 1} \frac{1}{(s+w)^{2p}}.$$

In the above result, we easily see that for any fixed $N > 0$, there exists a constant $C_N > 0$ independent of $t \geqslant 0$ such that

$$\mathcal{R}_1(t) + \mathcal{R}_2(t) \leqslant C_N t^{-N}.$$

Therefore, the convergence rate depends on p, and we conclude that the sequence generated by the DPG algorithm converges to the optimal point with a rate of $O(t^{-p/2})$. Next we state the result for the case $p = 1$.

Theorem 4.6.3 (convergence result for the diminishing step size). *Suppose that Assumptions 4.1.1 and 4.4.3 hold. Let $\alpha(t) = \frac{v}{(t+w)}$ with $v, w > 0$ satisfying*

$$\alpha(0) = \frac{v}{w} \leqslant \frac{2}{L+\mu}.$$

Also, choose $v > 0$ such that $\mu v/2 > 1$. Then we have

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 \leqslant \left(\frac{w}{t+w}\right)^{\mu v/2} \|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|^2 + \mathcal{R}_3(t), \tag{4.6.1}$$

where $Q = \left(\frac{w+1}{w}\right)^2$, $\rho = \sup_{t \geqslant 0} \beta^t/\alpha(t)^2$ *and*

$$\mathcal{R}_3(t) = \frac{Q}{(\mu v/2) - 1}\Big(\frac{w+1}{w}\Big)^{\mu v/2} \frac{\Big(\rho G_1 + G_2\Big)v^2}{(t+w-1)}.$$

We observe that the convergence rate of the first term on the right-hand side of (4.6.1) is $O(t^{-\mu v/2})$, while that of $\mathcal{R}_3(t)$ is $O(t^{-1})$. Since $\mu v/2 > 1$ as chosen in Theorem 4.6.3, we conclude that the sequence converges to the optimal point with a rate of $O(t^{-\mu v/4})$.

Proof of Theorems 4.6.1, 4.6.2 and 4.6.3. Using the notation (4.4.2), we can rewrite Theorem 4.5.1 and Proposition 4.3.2 as

$$A(t) \leqslant \beta^t A(0) + \frac{J\alpha(t)^2}{(1-\beta)^2} \tag{4.6.2}$$

and

$$\begin{aligned} &A(t+1) + B(t+1) \\ &\leqslant (c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})A(t) + \left(1 - \frac{\mu}{2}\alpha(t) + c_1\alpha(t)^2\right)B(t) + c_2\alpha(t)^2. \end{aligned} \tag{4.6.3}$$

Combining (4.6.2) and (4.6.3), we get

$$\begin{aligned} B(t+1) &\leqslant A(t+1) + B(t+1) \\ &\leqslant \left(1 - \frac{\mu}{2}\alpha(t)\right) B(t) + (c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})A(0)\beta^t \\ &\quad + \left((c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})\frac{J}{(1-\beta)^2} + c_1 B(t) + c_2\right)\alpha(t)^2. \end{aligned}$$

Since $A(t) + B(t) < R$ and $\alpha(t) \leqslant \alpha(0)$, it follows that

$$B(t+1) \leqslant \left(1 - \frac{\mu}{2}\alpha(t)\right) B(t) + G_1\beta^t + G_2\alpha(t)^2, \tag{4.6.4}$$

where

$$\begin{aligned} G_1 &= (c_3\alpha(0)^2 + c_4\alpha(0) + \beta)R, \\ G_2 &= (c_3\alpha(t)^2 + c_4\alpha(t) + \tilde{\beta})\frac{J}{(1-\beta)^2} + c_1 R + c_2. \end{aligned}$$

Case $\alpha(t) \equiv \alpha$**.** To estimate the sequence $B(t)$ we consider the following two sequences $\{q_1(t)\}_{t\geqslant 0}$ and $\{q_2(t)\}_{t\geqslant 0}$ satisfying

$$q_1(t+1) = \Big(1 - \frac{\mu}{2}\alpha\Big)q_1(t) + G_1\beta^t, \quad q_1(0) = 0,$$
$$q_2(t+1) = \Big(1 - \frac{\mu}{2}\alpha\Big)q_2(t) + G_2\alpha^2, \quad q_2(0) = B(0).$$

It then easily follows that $B(t) \leqslant q_1(t) + q_2(t)$ for all $t \geqslant 0$. Similarly as in (4.5.2), we have

$$q_1(t) = \sum_{s=0}^{t-1} G_1\beta^s\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1-s} \leqslant \frac{2}{\mu\alpha}\left(\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1} + \beta^{\frac{t-1}{2}}\right)$$

Note that

$$\sum_{s=0}^{t-1}\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1-s} \leqslant \frac{2}{\mu\alpha}.$$

Using this, we estimate $q_2(t)$ as

$$q_2(t) \leqslant \Big(1 - \frac{\mu\alpha}{2}\Big)^t q_2(0) + \frac{2G_2}{\mu}\alpha.$$

Combining the above estimates, we obtain

$$B(t) \leqslant \Big(1 - \frac{\mu\alpha}{2}\Big)^t B(0) + \frac{2G_2}{\mu}\alpha + \frac{2}{\mu\alpha}\left(\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1} + \beta^{\frac{t-1}{2}}\right),$$

which verifies the result of Theorem 4.6.1.

We next consider diminishing step size $\alpha(t) = \frac{v}{(t+w)^p}$, where $0 < p \leqslant 1$. For diminishing step size, we set

$$\rho := \sup_{t\geqslant 0}\frac{\beta^t}{\alpha(t)^2}.$$

Then it follows that

$$B(t+1) \leqslant \Big(1 - \frac{\mu}{2}\alpha(t)\Big)B(t) + \Big(\rho G_1 + G_2\Big)\alpha(t)^2,$$

Case $p \in (0,1)$**.** The estimate (4.6.4) reads as

$$B(t+1) \leqslant \Big(1 - \frac{\mu v}{2(t+w)^p}\Big)B(t) + \Big(\rho G_1 + G_2\Big)\frac{v^2}{(t+w)^{2p}}.$$

By applying Proposition 2.4.1, it follows that

$$B(t) \leqslant \frac{4Q\Big(\rho G_1 + G_2\Big)v}{\mu}([t/2] + w - 1)^{-p} + \mathcal{R}_1(t) + \mathcal{R}_2(t),$$

where

$$\begin{aligned}\mathcal{R}_1(t) &= e^{-\sum_{s=0}^{t-1}\frac{\mu v}{2(s+w)^p}}B(0)\\ \mathcal{R}_2(t) &= Q\Big(\rho G_1 + G_2\Big)v^2 e^{-\frac{\mu v t}{4(t+w)^p}}\sum_{s=1}^{[t/2]-1}\frac{1}{(s+w)^{2p}}\end{aligned}$$

with constant $Q = \left(\frac{w+1}{w}\right)^{2p}$. The proof of Theorem 4.6.2 is done.

Case $p = 1$**.** The estimate (4.6.4) gives

$$B(t+1) \leqslant \Big(1 - \frac{\mu v}{2(t+w)}\Big)B(t) + \Big(\rho G_1 + G_2\Big)\frac{v^2}{(t+w)^2}.$$

Choose $v > 0$ so that $C_1 = \mu v/2 > 1$. Then we use Proposition 4.3.2 to derive the following estimate

$$B(t) \leqslant \Big(\frac{w}{t+w}\Big)^{C_1}B(0) + \frac{1}{C_1 - 1}\Big(\frac{w+1}{w}\Big)^{C_1}\frac{Q\Big(\rho G_1 + G_2\Big)v^2}{(t+w-1)},$$

where $Q = \left(\frac{w+1}{w}\right)^2$. It proves Theorem 4.6.3. □

4.7 Improved Convergence Estimate

In this section, we show that the convergence result of Theorem 4.6.1 can be improved for one dimensional example. Recall the following estimate in Theorem 4.6.1:

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}_*\|^2 \leqslant \Big(1 - \frac{\mu\alpha}{2}\Big)^t\|\bar{\mathbf{x}}(0) - \mathbf{x}_*\|^2 + \frac{2G_2}{\mu}\alpha + \frac{2}{\mu\alpha}\left(\Big(1 - \frac{\mu\alpha}{2}\Big)^{t-1} + \beta^{\frac{t-1}{2}}\right).$$

This implies that the sequence $\{\bar{x}(t)\}_{t\geqslant 0}$ converges to an $O(\sqrt{\alpha})$-neighborhood of the optimal point. However, if $\Omega = \mathbb{R}^n$, it is known by the work [49] that the sequence converges to an $O(\alpha)$-neighborhood of the optimal point. Hence, it is natural to ask the following question: While our approach for Theorem 4.6.1 has led to converge to an $O(\sqrt{\alpha})$-neighborhood, can we improve empirical performance up to an $O(\alpha)$-neighborhood? We answer this question by constructing a specific example.

One-dimensional example

Let us consider the functions $g_1, g_2 : [1,\infty) \to \mathbb{R}$ defined by

$$g_1(x) = 5x^2 \quad\text{and}\quad g_2(x) = -3x^2, \quad x \in [1,\infty) \tag{4.7.1}$$

and the mixing matrix $\tilde{W}$ defined by

$$\tilde{W} = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix} \tag{4.7.2}$$

satisfying (E) of Assumption 4.1.1. We note that the total cost function $g = (g_1 + g_2)/2$ has the optimal point at $x = 1$. Then we can represent the projected decentralized gradient descent algorithm with a constant step size α explicitly as follows:

$$\begin{aligned} x_1(t+1) &= \max\left[\frac{2}{3}x_1(t) + \frac{1}{3}x_2(t) - 10\alpha x_1(t),\ 1\right], \\ x_2(t+1) &= \max\left[\frac{1}{3}x_1(t) + \frac{2}{3}x_2(t) + 6\alpha x_2(t),\ 1\right]. \end{aligned} \tag{4.7.3}$$

We first establish that the state $(x_1(t), x_2(t))$ generated by the algorithm (4.7.3) will be confined to a certain region after a finite number of iterations. The proof for the following lemma will be provided at the end of this section.

Lemma 4.7.1. *Let $g_1(x)$, $g_2(x)$ and the mixing matrix $\tilde{W}$ be defined by* (4.7.1) *and* (4.7.2) *and let $\mathbf{x}(t) = (x_1(t), x_2(t))$ be the state at $t \geqslant 0$ generated by* (4.7.3)*. Then for any initial state $\mathbf{x}(0) = (x_1(0), x_2(0))$ and $\alpha \in (0, 1/45)$, there exists $t_0 \leqslant \log \|\mathbf{x}(0)\|_2 / \log(1/\lambda_+) - 1$ such that $x_1(t_0+1) = 1$ and $x_2(t_0+1) \leqslant 1 + 30\alpha$, where $\lambda_+ \in (0,1)$ is defined as*

$$\lambda_+ = \frac{2}{3} - 2\alpha + \sqrt{\frac{1}{9} + 64\alpha^2}.$$

Now, we demonstrate that the state $(x_1(t), x_2(t))$ generated by the algorithm (4.7.3) converges to an $O(\alpha)$ neighborhood of the optimal point $(1,1)$.

Theorem 4.7.2. *Let $g_1(x)$, $g_2(x)$ and the mixing matrix $\tilde{W}$ be defined by* (4.7.1) *and* (4.7.2) *and let $\mathbf{x}(t) = (x_1(t), x_2(t))$ be the state at $t \geqslant 0$ generated by* (4.7.3)*. Then for any initial state $\mathbf{x}(0) = (x_1(0), x_2(0))$ and $\alpha \in (0, 1/45)$, the state $\mathbf{x}(t)$ converges exponentially fast to the point $(1, 1/(1-18\alpha))$ which belongs to an $O(\alpha)$ neighborhood of the optimal point $(1,1)$.*

Proof. By Lemma 4.7.1, we can choose t_0 satisfying $x_1(t_0+1) = 1$ and $x_2(t_0+1) \leqslant 1 + 30\alpha$. Note that if $\alpha < 1/45$ then $1/(1-18\alpha) < 1 + 30\alpha$. Since $x_1(t_0+1) = 1$ and $x_2(t_0+1) < 1 + 30\alpha$, it follows that

$$\frac{2}{3} + \frac{1}{3}x_2(t_0+1) - 10\alpha < 1,$$

which implies $x_1(t_0+2) = 1$. We have

$$\begin{aligned} x_2(t_0+2) &= \frac{1}{3}x_1(t_0+1) + \frac{2}{3}x_2(t_0+1) + 6\alpha x_2(t_0+1) \\ &= \frac{1}{3}x_1(t_0+1) + \frac{2+18\alpha}{3}x_2(t_0+1). \end{aligned} \tag{4.7.4}$$

We can further write (4.7.4) as

$$x_2(t_0+2) - \frac{1}{1-18\alpha} = \frac{2+18\alpha}{3}\left[x_2(t_0+1) - \frac{1}{1-18\alpha}\right]. \tag{4.7.5}$$

Since $(2+18\alpha)/3 < 1$, it follows that $x_2(t_0+2) < x_2(t_0+1) < 1+30\alpha$ for $x_2(t_0+1) > \frac{1}{1-18\alpha}$, and $1 \leqslant x_2(t_0+2) \leqslant \frac{1}{1-18\alpha}$ for $1 \leqslant x_2(t_0+1) \leqslant \frac{1}{1-18\alpha}$. We can conclude that $x_1(t) = 1$ and $x_2(t) < 1+30\alpha$ for all $t \geqslant t_0+1$. In addition, (4.7.5) implies that $x_2(t)$ converges to $1/(1-18\alpha)$. The proof is done. □

The above result will be also verified by numerical test in the next section. This result suggests that the sequence $\{\bar{x}(t)\}$ converges to an $O(\alpha)$-neighborhood of the optimal point which is a stronger result than the convergence result to an $O(\sqrt{\alpha})$-neighborhood of Theorem 4.6.1. We guess that the result of Theorem 4.7.2 could be extended to more general examples. Before ending this section, we give proof of Lemma 4.7.1.

Proof of Lemma 4.7.1. Notice that if $x_1(t+1) > 1$ and $x_2(t+1) > 1$, it should hold that

$$\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \end{pmatrix} = \begin{pmatrix} \frac{2}{3} - 10\alpha & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} + 6\alpha \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}. \tag{4.7.6}$$

The eigenvalues of the matrix in the right hand side of (4.7.6) are

$$\lambda_\pm = \frac{2}{3} - 2\alpha \pm \sqrt{\frac{1}{9} + 64\alpha^2}$$

which are positive and less than 1 for $\alpha \in (0, 1/45)$. Therefore

$$\|(x_1(t+1), x_2(t+1))\|_2 \leqslant \lambda_+ \|(x_1(t), x_2(t))\|_2,$$

and so

$$\|(x_1(t), x_2(t))\|_2 \leqslant \lambda_+^t \|(x_1(0), x_2(0))\|_2.$$

Thus we can find a smallest integer $t_0 \leqslant \log\|\mathbf{x}(0)\|_2 / \log(1/\lambda_+) - 1$ such that $x_1(t_0+1) = 1$. By (4.7.3), it follows that

$$\frac{2}{3}x_1(t_0) + \frac{1}{3}x_2(t_0) - 10\alpha x_1(t_0) \leqslant 1.$$

This leads to

$$\begin{aligned}\frac{1}{3}x_2(t_0) &\leqslant 1-\Big(\frac{2}{3}-10\alpha\Big)x_1(t_0)\\ &\leqslant 1-\Big(\frac{2}{3}-10\alpha\Big)=\frac{1}{3}+10\alpha,\end{aligned} \tag{4.7.7}$$

which implies that $x_2(t_0)\leqslant 1+30\alpha$. Now we want to show that $x_2(t_0+1)\leqslant 1+30\alpha$. Note that

$$\begin{aligned}\frac{1}{3}x_1(t_0)+\Big(\frac{2}{3}+6\alpha\Big)x_2(t_0) &\leqslant \frac{1}{3}\Big(\frac{3-x_2(t_0)}{2-30\alpha}\Big)+\Big(\frac{2}{3}+6\alpha\Big)x_2(t_0)\\ &=\frac{1}{2-30\alpha}+\Big(\frac{2}{3}+6\alpha-\frac{1}{6-90\alpha}\Big)x_2(t_0).\end{aligned}$$

Here we used $x_1(t_0)\leqslant(3-x_2(t_0))/(2-30\alpha)$ from (4.7.7) for the first inequality. Inserting this into (4.7.3) we find

$$\begin{aligned}x_2(t_0+1) &= \max\left[\frac{1}{3}x_1(t_0)+\Big(\frac{2}{3}+6\alpha\Big)x_2(t_0),1\right]\\ &= \max\left[\frac{1}{2-30\alpha}+\Big(\frac{2}{3}+6\alpha-\frac{1}{6-90\alpha}\Big)x_2(t_0),1\right].\end{aligned}$$

Combining this with $x_2(t_0)\leqslant 1+30\alpha$, we get

$$x_2(t_0+1)\leqslant \max\left[\frac{1}{2-30\alpha}+\Big(\frac{2}{3}+6\alpha-\frac{1}{6-90\alpha}\Big)(1+30\alpha),1\right].$$

Note that

$$\begin{aligned}&\frac{1}{2-30\alpha}+\Big(\frac{2}{3}+6\alpha-\frac{1}{6-90\alpha}\Big)(1+30\alpha)\leqslant 1+30\alpha\\ &\Leftrightarrow 1\leqslant(1-22\alpha+180\alpha^2)(1+30\alpha)\\ &\Leftrightarrow 0\leqslant \alpha(1-45\alpha)(1-15\alpha),\end{aligned}$$

which holds true for $\alpha\in(0,1/45)$. Applying this, we finally get

$$x_2(t_0+1)\leqslant 1+30\alpha.$$

□

Half-space example

Throughout this section, we will use the following notations:

- $\tilde{x}_k(t)\in\mathbb{R}^{d-1}$ denotes the first $d-1$ coordinates of $x_k(t)$.

- $y(t) \in \mathbb{R}^{d-1}$ denotes the first $d-1$ coordinates of $\bar{x}(t)$.

- $\tilde{x}_* \in \mathbb{R}^{d-1}$ denotes the first $d-1$ coordinates of x_*.

- $\tilde{\nabla} f(\cdot) \in \mathbb{R}^{d-1}$ denotes the first $d-1$ coordinates of $\nabla f(\cdot)$.

We investigate the case where the domain is a half-space $\Omega = \{(\tilde{x}, x[d]) \,|\, \tilde{x} \in \mathbb{R}^{d-1}, x[d] \geqslant 0\} \subseteq \mathbb{R}^d$. Here $x[k] \in \mathbb{R}$ denotes the *k-th* component of the vector $x \in \mathbb{R}^d$. Let $x_* = (\tilde{x}_*, x_*[d]) \in \mathbb{R}^d$ be a solution, i.e. $x_* = \arg\min_{x \in \Omega} f(x)$. We assume the following assumption for a solution.

Assumption 4.7.3. We assume that the minimizer of f is on the boundary of Ω, specifically, $x_* = (\tilde{x}_*, 0)$ with some $\tilde{x}_* \in \mathbb{R}^{d-1}$.

Additionally, we assume the following:

Assumption 4.7.4. The total cost function and each local cost function satisfy (B) and (C) of Assumption 4.1.1. In addition each local cost function f_i is continuously differentiable function, and the minimizer $x_* = (\tilde{x}_*, 0)$ satisfies $\partial_d f(x_*) = \frac{1}{n}\sum_{i=1}^n \partial_d f_i(x_*) \geqslant \omega > 0$. Here $\partial_k f$ denotes the *k-th* component of the gradient ∇f.

Lemma 4.7.5. *Let $\{x_k(t)\}_{t \geqslant 0}$ be the sequence of the decentralized projected gradient* (4.1.1) *with constant stepsize $\alpha > 0$. There is a small value $J > 0$ such that for any $\alpha \in (0, J)$ there exists a time instant $T_\alpha \in \mathbb{N}$ satisfying the following estimate:*

$$\|x_k(t) - x_*\| \leqslant \frac{w}{2L} \tag{4.7.8}$$

and

$$\frac{1}{n}\sum_{k=1}^n \partial_d f_k(x_k(t)) \geqslant \frac{w}{2}, \tag{4.7.9}$$

for any $t \geqslant T_\alpha$.

Proof. By Theorem 4.6.1, there exist a time $T_\alpha \in \mathbb{N}$ such that $\|x_k(t) - x_*\| = O(\sqrt{\alpha})$ for $t \geqslant T_\alpha$. Thus, we have $\|x_k(t) - x_*\| \leqslant \frac{w}{2L}$ for $\alpha \in (0, J)$ with a suitable value $J > 0$.

Since a local cost function f_i is L-smooth and $x_k(t) \in \mathbb{R}^d$ satisfies $\|x_k(t) - x_*\| \leqslant \frac{w}{2L}$ for $1 \leqslant k \leqslant n$, we have

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n \partial_d f_k(x_k(t)) &= \frac{1}{n}\sum_{k=1}^n \partial_d f_k(x_*) + \frac{1}{n}\sum_{k=1}^n \Big(\partial_d f_k(x_k(t)) - \partial f_k(x_*)\Big) \\ &\geqslant w - L\Big(\frac{w}{2L}\Big) = \frac{\omega}{2}.\end{aligned} \tag{4.7.10}$$

The proof is done. □

To demonstrate that a sequence $x_k(t)$ converges to an $O(\alpha)$-neighborhood of the optimal point, we first show that there exists a time T such that for all $t > T$ we have $\bar{x}(t)[d] \leqslant O(\alpha)$. To achieve this, we introduce the following lemma.

Lemma 4.7.6. *For all $t \geqslant T_\alpha$, one of the following two cases holds.*

Case 1. Assume that $\sum_{j=1}^n w_{kj}x_j(t) - \eta\nabla f_k(x_k(t))$ belongs to Ω for all $1 \leqslant k \leqslant n$. Then we have

$$\bar{x}(t+1)[d] \leqslant \bar{x}(t)[d] - \frac{\omega\alpha}{2}. \tag{4.7.11}$$

Case 2. Assume that $\sum_{j=1}^n w_{kj}x_j(t) - \eta\nabla f_k(x_k(t))$ does not belong to Ω for some $1 \leqslant k \leqslant n$. Then we have

$$\bar{x}(t+1)[d] \leqslant \Big(\beta^{(t+1)}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}\Big)^{1/2}. \tag{4.7.12}$$

Proof. We begin by proving the first case, where the projection operator can be eliminated:

$$x_k(t+1) = \sum_{j=1}^n w_{kj}x_j(t) - \alpha\nabla f_k(x_k(t)) \quad \forall\ 1 \leqslant k \leqslant n.$$

This scheme can be written as follows.

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha}{n}\sum_{k=1}^n \nabla f_k(x_k(t)).$$

Considering the d-th coordinate of the above formula and using (4.7.10), it follows that

$$\bar{x}(t+1)[d] = \bar{x}(t)[d] - \frac{\alpha}{n}\sum_{k=1}^n \partial_d f_k(x_k(t)) \leqslant \bar{x}(t)[d] - \frac{\omega\alpha}{2}.$$

Next, we prove the second case. This case corresponds to the case that for some $1 \leqslant k \leqslant n$,

$$\Big(\sum_{j=1}^n w_{kj}x_j(t) - \eta\nabla f_k(x_k(t))\Big)[d] \leqslant 0, \tag{4.7.13}$$

and so we have

$$x_k(t+1)[d] = P_\Omega\Big(\sum_{j=1}^n w_{kj}x_j(t) - \eta\nabla f_k(x_k(t))\Big)[d] = 0. \tag{4.7.14}$$

Therefore

$$\begin{aligned}(\bar{x}(t+1)[d])^2 &= |\bar{x}(t+1)[d] - x_k(t+1)[d]|^2 \\ &\leqslant \|\bar{x}(t+1) - x(t+1)\|^2 \\ &\leqslant \beta^{(t+1)}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}.\end{aligned} \tag{4.7.15}$$

Here we used Theorem 4.5.1 for the last inequality. □

Now we establish our first goal: There exists a time T such that for all $t > T$, we have $\bar{x}(t)[d] \leqslant O(\alpha)$.

Proposition 4.7.7. *Let $t_0 \in \mathbb{N}$ be the smallest number such that $\frac{w\alpha}{2}t_0 > K$ with $K > 0$ defined as*

$$K = \Big\{\bar{x}(T_\alpha)[d],\ \Big(\beta\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}\Big)^{1/2}\Big\}, \tag{4.7.16}$$

and let $t_1 \in \mathbb{N}$ satisfy $\beta^{t_1/2}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\| \leqslant \frac{3(L^2R^2+nD^2)\alpha}{1-\beta}$. Then for all $t > T = t_0 + \max\{T_\alpha, t_1\}$, we have

$$\bar{x}(t+1)[d] \leqslant \frac{6(L^2R^2 + nD^2)\alpha}{1-\beta}. \tag{4.7.17}$$

Proof. For each $t \geqslant T_\alpha$, the result of Lemma 4.7.6 gives the following estimate

$$\bar{x}(t+1)[d] \leqslant \max\Big\{\bar{x}(t)[d] - \frac{w\alpha}{2},\ \Big(\beta^{(t+1)}\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}\Big)^{1/2}\Big\}. \tag{4.7.18}$$

From this we find that for $t \geqslant \max\{T_\alpha, t_1\}$,

$$\bar{x}(t)[d] \leqslant \max\Big\{\bar{x}(0)[d],\ \Big(\beta\|\mathbf{x}(0) - \bar{\mathbf{x}}(0)\|^2 + \frac{3(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}\Big)^{1/2}\Big\} = K. \tag{4.7.19}$$

Also, using (4.7.18) for $t \geqslant \max\{T_\alpha, t_1\}$ we have

$$\bar{x}(t+1)[d] \leqslant \max\Big\{\bar{x}(t)[d] - \frac{w\alpha}{2},\ \frac{\sqrt{6(L^2R^2 + nD^2)}\alpha}{1-\beta}\Big\}. \tag{4.7.20}$$

Combining this with $\bar{x}(\max\{T_\alpha, t_1\})[d] \leqslant K$ yields that

$$\bar{x}(t+1)[d] \leqslant \frac{\sqrt{6(L^2R^2 + nD^2)}\alpha}{1-\beta} \quad \text{for } t \geqslant t_0 + \max\{T_\alpha, t_1\}. \tag{4.7.21}$$

It completes the proof. □

Now we investigate the convergence property $\tilde{x}_k(t)$ towards the point $\tilde{x}_*$ for each $1 \leqslant k \leqslant n$. We first investigate the convexity property of the function f in terms of its first $d-1$ coordinates in the following lemma.

Lemma 4.7.8. *Let T be defined in Proposition 4.7.7 and $y(t)$ be the first $d-1$ coordinates of $\bar{x}(t)$. Then, for all $t \geqslant T$, we have the following inequality:*

$$\begin{aligned}
&\Big\langle y(t) - \tilde{x}_*,\ \tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*)\Big\rangle \\
&\geqslant \frac{L\mu}{2(L+\mu)}\|\bar{x}(t) - x_*\|^2 + \frac{1}{L+\mu}\|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2 \\
&\qquad\qquad - \frac{2(L+\mu)}{\mu} \cdot \frac{9L(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2}.
\end{aligned}$$

Proof. Using the fact that f is L-smooth and Proposition 4.7.7, we deduce for $t > T$ the following estimate

$$\begin{aligned}
&\Big|\Big\langle \bar{x}(t)[d] - x_*[d],\ (\partial_d f(\bar{x}(t)) - \partial_d f(x_*))\Big\rangle\Big| \\
&= \Big|\Big\langle \bar{x}(t)[d],\ (\partial_d f(\bar{x}(t)) - \partial_d f(x_*))\Big\rangle\Big| \\
&\leqslant \frac{\sqrt{6L(L^2R^2 + nD^2)}\alpha}{1-\beta}\|\bar{x}(t) - x_*\| \\
&\leqslant \frac{2(L+\mu)}{\mu} \cdot \frac{6L(L^2R^2 + nD^2)\alpha^2}{4(1-\beta)^2} + \frac{L\mu}{2(L+\mu)}\|\bar{x}(t) - x_*\|^2,
\end{aligned}$$

where we used Young's inequality $2ab \leqslant \kappa a^2 + \frac{b^2}{\kappa}$ with $\kappa = \frac{2(L+\mu)}{\mu}$ for the last inequality. Using this, we have

$$\begin{aligned}
&\Big\langle \bar{x}(t) - x_*,\ \nabla f(\bar{x}(t)) - \nabla f(x_*)\Big\rangle \\
&= \Big\langle \tilde{x}(t) - \tilde{x}_*,\ \tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*)\Big\rangle + \Big\langle \bar{x}(t)[d] - x_*[d],\ \partial_d f(\bar{x}(t)) - \partial_d f(x_*)\Big\rangle \\
&\leqslant \Big\langle \tilde{x}(t) - \tilde{x}_*,\ \tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*)\Big\rangle + \frac{(L+\mu)}{\mu} \cdot \frac{3L(L^2R^2 + nD^2)\alpha^2}{(1-\beta)^2} \\
&\qquad\qquad + \frac{L\mu}{2(L+\mu)}\|\bar{x}(t) - x_*\|^2.
\end{aligned} \tag{4.7.22}$$

Since the total cost function f is L-smooth and μ-strongly convex, it follows by Lemma 2.1.9 that

$$\Big\langle \bar{x}(t) - x_*,\ \nabla f(\bar{x}(t)) - \nabla f(x_*)\Big\rangle \geqslant \frac{L\mu}{L+\mu}\|\bar{x}(t) - x_*\|^2 + \frac{1}{L+\mu}\|\nabla f(\bar{x}(t)) - \nabla f(x_*)\|^2.$$

Combining this with (4.7.22), it follows that

$$\begin{aligned}&\left\langle y(t)-\tilde{x}_*,\ \tilde{\nabla} f(\bar{x}(t))-\tilde{\nabla} f(x_*)\right\rangle \\ &\geqslant \frac{L\mu}{2(L+\mu)}\|\bar{x}(t)-x_*\|^2+\frac{1}{L+\mu}\|\nabla f(\bar{x}(t))-\nabla f(x_*)\|^2 \\ &\qquad\qquad -\frac{(L+\mu)}{\mu}\cdot\frac{3L(L^2R^2+nD^2)\alpha^2}{(1-\beta)^2},\end{aligned}$$

which proves the lemma. □

Now we show that the sequence $\{y(t)\}_{t\geqslant 0}$ converges to an $O(\alpha)$-neighborhood of the optimal point $\tilde{x}_*$.

Proposition 4.7.9. *Let T be defined in Proposition 4.7.7. Then there exists $c_1, c_2 > 0$ such that for $t \geqslant T$ we have*

$$\left\|y(t)-\tilde{x}_*\right\|^2 \leqslant (1-c_1\alpha)^t\left\|y(0)-\tilde{x}_*\right\|^2+\frac{c_2}{c_1}\alpha^2. \tag{4.7.23}$$

Proof. We express the first $d-1$ coordinates of formula (4.1.1) as:

$$\tilde{x}_k(t+1)=\sum_{j=1}^n w_{kj}\tilde{x}_j(t)-\alpha\tilde{\nabla} f_k(x_k(t)). \tag{4.7.24}$$

Averaging this for $1\leqslant k\leqslant n$, we find (4.7.24) as

$$y(t+1)=y(t)-\alpha\tilde{\nabla} f(\bar{x}(t))+\alpha\left(\tilde{\nabla} f(\bar{x}(t))-\frac{1}{n}\sum_{k=1}^n\tilde{\nabla} f_k(x_k(t))\right). \tag{4.7.25}$$

Using the fact that f_k is L-smooth and Theorem 4.5.1, we have

$$\begin{aligned}\left\|\tilde{\nabla} f(\bar{x}(t))-\frac{1}{n}\sum_{k=1}^n\tilde{\nabla} f_k(x_k(t))\right\|^2 &\leqslant \frac{1}{n}\sum_{k=1}^n\left\|\tilde{\nabla} f_k(\bar{x}(t)-\tilde{\nabla} f_k(x_k(t))\right\|^2 \\ &\leqslant \frac{L^2}{n}\sum_{k=1}^n\|\bar{x}(t)-x_k(t)\|^2 \\ &\leqslant L^2\beta^t\|\mathbf{x}(0)-\bar{\mathbf{x}}(0)\|^2+\frac{3L^2(L^2R^2+nD^2)\alpha^2}{1-\beta} \\ &\leqslant \frac{6L^2(L^2R^2+nD^2)\alpha^2}{1-\beta}.\end{aligned} \tag{4.7.26}$$

Using (4.7.25), Young's inequality and (4.7.26), it follows that

$$
\begin{aligned}
&\|y(t+1) - \tilde{x}_*\|^2 \\
=&\left\| y(t) - \alpha f(\bar{x}(t)) + \alpha \left(\tilde{\nabla} f(\bar{x}(t)) - \frac{1}{n} \sum_{k=1}^{n} \tilde{\nabla} f_k(x_k(t)) \right) \right\|^2 \\
\leqslant&(1+c\alpha) \left\| y(t) - \alpha \tilde{f}(\bar{x}(t)) - \tilde{x}_* \right\|^2 + \left(1 + \frac{1}{c\alpha}\right) \frac{6L^2(L^2R^2 + nD^2)}{(1-\beta)^2} \alpha^4,
\end{aligned}
\tag{4.7.27}
$$

where we have let $c = \frac{L\mu}{2(L+\mu)}$. Next, we estimate the first term in the last inequality in (4.7.27). By the first optimality condition, we have $\tilde{\nabla} f(x_*) = 0$. Using this it follows that

$$
\begin{aligned}
&\|y(t) - \alpha \tilde{\nabla} f(\bar{x}(t)) - \tilde{x}_*\|^2 \\
&= \left\| y(t) - \tilde{x}_* - \alpha \Big(\tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*) \Big) \right\|^2 \\
&= \left\| y(t) - \tilde{x}_* \right\|^2 + \alpha^2 \left\| \tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*) \right\|^2 - 2\alpha \Big\langle y(t) - \tilde{x}_*,\ \tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*) \Big\rangle.
\end{aligned}
$$

Applying Lemma 4.7.8 to the last equality, we get

$$
\begin{aligned}
&\|y(t) - \alpha \tilde{\nabla} f(\bar{x}(t)) - \tilde{x}_*\|^2 \\
&= \Big(1 - \frac{L\mu}{(L+\mu)} \eta \Big) \left\| y(t) - \tilde{x}_* \right\|^2 + \Big(\alpha^2 - \frac{2\alpha}{L+\mu} \Big) \left\| \Big(\tilde{\nabla} f(\bar{x}(t)) - \tilde{\nabla} f(x_*) \Big) \right\|^2 \\
&\quad + \frac{2(L+\mu)}{\mu} \cdot \frac{3L(L^2R^2 + nD^2)\alpha^3}{(1-\beta)^2}.
\end{aligned}
$$

This inequality with the condition $\alpha \leqslant \frac{2}{L+\mu}$ gives

$$
\begin{aligned}
&\|y(t) - \eta \tilde{\nabla} f(\bar{x}(t)) - \tilde{x}_*\|^2 \\
&= \Big(1 - \frac{L\mu}{(L+\mu)} \alpha \Big) \left\| y(t) - \tilde{x}_* \right\|^2 + \frac{2(L+\mu)}{\mu} \cdot \frac{3L(L^2R^2 + nD^2)\alpha^3}{(1-\beta)^2}.
\end{aligned}
\tag{4.7.28}
$$

Putting (4.7.28) into (4.7.27) we obtain

$$
\begin{aligned}
&\|y(t+1) - \tilde{x}_*\|^2 \\
&\leqslant (1+c\alpha) \Big(1 - \frac{L\mu}{(L+\mu)} \alpha \Big) \left\| y(t) - \tilde{x}_* \right\|^2 + (1+c\alpha) \frac{2(L+\mu)}{\mu} \cdot \frac{3L(L^2R^2 + nD^2)\alpha^3}{(1-\beta)^2} \\
&\qquad + \left(1 + \frac{1}{c\alpha}\right) \frac{6L^2(L^2R^2 + nD^2)}{(1-\beta)^2} \alpha^4.
\end{aligned}
$$

This estimate can be written in the following form

$$
\|y(t+1) - \tilde{x}_*\|^2 \leqslant (1 - c_1\alpha) \left\| y(t) - \tilde{x}_* \right\|^2 + c_2\alpha^3, \tag{4.7.29}
$$

where $c_1 = \frac{L\mu}{2(L+\mu)}$ and

$$\begin{aligned}\tilde{c}_2 &= (1+c\alpha)\frac{2(L+\mu)}{\mu}\cdot\frac{3L(L^2R^2+nD^2)}{(1-\beta)^2} + \left(\alpha+\frac{1}{c}\right)\frac{6L^2(L^2R^2+nD^2)}{(1-\beta)^2} \\ &\leqslant (1+c)\frac{2(L+\mu)}{\mu}\cdot\frac{3L(L^2R^2+nD^2)}{(1-\beta)^2} + \left(1+\frac{1}{c}\right)\frac{6L^2(L^2R^2+nD^2)}{(1-\beta)^2} =: c_2.\end{aligned} \tag{4.7.30}$$

Using this iteratively, we get

$$\begin{aligned}\|y(t)-\tilde{x}_*\|^2 &\leqslant (1-c_1\alpha)^t\|y(0)-\tilde{x}_*\|^2 + c_2\alpha^3\sum_{k=0}^{t-1}(1-c_1\alpha)^k \\ &\leqslant (1-c_1\alpha)^t\|\tilde{x}(0)-\tilde{x}_*\|^2 + \frac{c_2}{c_1}\alpha^2,\end{aligned} \tag{4.7.31}$$

which completes the proof. □

Theorem 4.7.10. *Suppose that the domain Ω is the half space $\mathbb{R}^{d-1}\times\mathbb{R}_+$ with any dimension $d \geqslant 1$ and Assumption 4.7.3 holds. Suppose also that Assumptions 4.1.1 and 4.4.3 hold. If the stepsize is given by a constant $\alpha > 0$ such that $\alpha \leqslant \frac{2}{L+\mu}$, we have*

$$\mathit{limsup}_{t\to\infty}\|x_k(t)-x_*\| = O(\alpha) \quad \forall\ 1 \leqslant k \leqslant n. \tag{4.7.32}$$

Proof. First we consider the case that $\alpha \in (0, J)$. We have the following formula

$$\|x_k(t)-x_*\|^2 = \|\tilde{x}_k(t)-\tilde{x}_*\|^2 + \|x_k(t)[d]\|^2. \tag{4.7.33}$$

By Theorem 4.5.1 and Proposition 4.7.9, for any $t \geqslant T$, we have

$$\begin{aligned}&\|\tilde{x}_k(t)-\tilde{x}_*\|^2 \\ &\leqslant 2\|\tilde{x}_k(t)-y(t)\|^2 + 2\|y(t)-\tilde{x}_*\|^2 \\ &\leqslant 2\beta^t\|\mathbf{x}(0)-\bar{\mathbf{x}}(0)\|^2 + 2(1-c_1\alpha)^t\|y(0)-\tilde{x}_*\|^2 + \left(\frac{6(L^2R^2+nD^2)}{(1-\beta)^2}+\frac{2c_2}{c_1}\right)\alpha^2,\end{aligned}$$

By Proposition 4.7.7,

$$\begin{aligned}\|x_k(t)[d]\|^2 &\leqslant 2\|x_k(t)[d]-\bar{x}(t)\|^2 + 2\|\bar{x}(t)\|^2 \\ &\leqslant 2\beta^t\|\mathbf{x}(0)-\bar{\mathbf{x}}(0)\|^2 + \frac{12(L^2R^2+nD^2)}{(1-\beta)^2}\alpha^2.\end{aligned}$$

Inserting the above two estimates in (4.7.33) we get

$$\begin{aligned}\|x_k(t)-x_*\|^2 \leqslant{}& 4\beta^t\|\mathbf{x}(0)-\bar{\mathbf{x}}(0)\|^2 + 2(1-c_1\alpha)^t\|y(0)-\tilde{x}_*\|^2 + \\ &+\left(\frac{6(L^2R^2+nD^2)}{(1-\beta)^2}+\frac{2c_2}{c_1}\right)\alpha^2.\end{aligned} \tag{4.7.34}$$

This gives for $\alpha \in (0, J)$ the following estimate

$$\limsup_{t\to\infty} \|x_k(t) - x_*\|^2 = O(\alpha^2). \tag{4.7.35}$$

For $J \leqslant \alpha \leqslant \frac{2}{L+\mu}$, we may use the result of Theorem 4.6.1 to derive

$$\limsup_{t\to\infty} \|x_k(t) - x_*\|^2 \leqslant C\alpha \leqslant \frac{C}{J}\alpha^2. \tag{4.7.36}$$

Combining the above two estimates completes the proof. □

4.8 Nuemrical Experiments

In this section, we conduct numerical experiments to validate the DPG algorithm. These experiments include non-negative least squares, constrained logistic regression, and a one-dimensional example. For non-negative least squares and constrained logistic regression, we construct a connected graph based on the Watts and Strogatz model [44], and employ a Laplacian-based constant edge weight matrix denoted as W [46]. This matrix is defined by $W = \mathbf{I} - L/\tau$ where L is the Laplacian matrix and τ is a constant that satisfies $\tau > \frac{1}{2}\lambda_{\max}(L)$. If $\lambda_{\max}(L)$ is not availiabe, one can use $\tau = \max_{1\leqslant i\leqslant n} D_{ii}(L)$ instead. This construction satisfies (D) and (E) of Assumption 4.1.1. In addition, we compare the DPG algorithm with the algorithm presented in [16], a version of the DIGing algorithm [32] for the constrained problem. We refer to the algorithm in [16] as P-DIGing.

Non-negative least squares

We consider the following decentralized non-negative least squares problem with n agents

$$\min_{x\in\Omega} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\|q_i - p_i^T x\|^2,$$

where $\Omega = \{(x[1], x[2], \cdots, x[d]) \mid x[k] \in \mathbb{R}_+\}$. In this case, the projection operator $\mathcal{P}_\Omega$ for a point $x = (x[1], x[2], \cdots, x[d])$ is defined by

$$\mathcal{P}_\Omega[x] = \begin{cases} 0 \text{ if } x[k] < 0, \\ x[k], \text{ otherwise,} \end{cases}$$

for all $1 \leqslant k \leqslant d$. We initialized the points $x_i(0)$ as independent random variables generated from a standard Gaussian distribution and then apply the projection operator. The variables $p_i \in \mathbb{R}^{d\times p}$ and $q_i \in \mathbb{R}^p$ are randomly chosen from the uniform distribution on $[0,1]$. In this simulation, we set the problem dimensions and the number of agents as $d = 10$, $p = 5$, and $n = 30$. In addition, each agent has four neighbors based on Watts and Strogatz model. Then we consider the relative convergence error

$$R(t) = \frac{\sum_{i=1}^{n} \|x_i(t) - x_*\|}{\sum_{i=1}^{n} \|x_i(0) - x_*\|}.$$

We examine the dynamics of $R(t)$ under various constant stepsizes and diminishing stepsizes. Note that the constant w can be determined by $w = (L + \mu)^{1/p}$. Here the values for the constants L and μ can be determined by

$$L = \max_{1 \leqslant i \leqslant n} \|p_i p_i^T\| \text{ and } \mu = \frac{|\lambda_n(p_i p_i^T)|}{n}.$$

We observe that with a constant stepsize, the value of $R(t)$ converges to a neighborhood of the solution, and the size of the neighborhood depends on the chosen stepsize as long as it meets the condition outlined in Theorem 4.6.1. However, if the stepsize exceeds $2/(L + \mu)$, the value of $R(t)$ diverges. Additionally, the right figure in Figure 4.1 shows that with diminishing stepsizes, the values of $R(t)$ converge to zero, which is expected in Theorem 4.6.2 and 4.6.3.

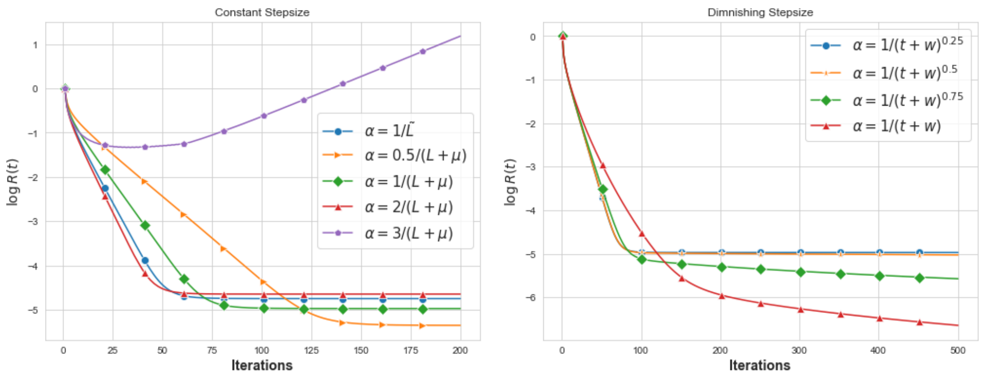


Figure 4.1: The behavior of $\log R(t)$ under various choices of constant stepsizes (left), diminishing stepsizes (right). The stepsize $1/\tilde{L}$ corresponds to the theoretical critical stepsize mentioned in [25], where $\tilde{L} = \frac{1}{n}\sum_{i=1}^{n} \|p_i p_i^T\|$.

Subsequently, we conduct a comparative analysis between the constant stepsize, the diminishing stepsize, and the P-DIGing algorithm . Note that the P-DIGing algorithm is the method to achieve exact convergence with a constant stepsize. In addition, we investigate the DPG+P-DIGing algorithm, which employs the DPG with a constant stepsize during the initial stages, transitioning to the P-DIGing algorithm. We empirically determine an optimal constant step size, specifically setting α to 0.2521 for P-DIGing. This choice ensures faster convergence compared to step sizes either larger or smaller than 0.2521. The numerical results for constant and diminishing stepsizes are illustrated in Figure 4.2. We observe that

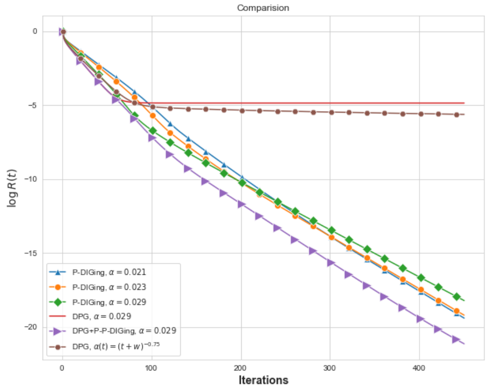


Figure 4.2: The behavior of $\log R(t)$ with P-DIGing, DPG using a constant step size, DPG+P-DIGing, and DPG with a diminishing step size.

while the DPG with a constant stepsize converges to a solution's neighborhood, it demonstrates faster convergence during the initial stages. Therefore, combining the strength of DPG and P-DIGing properly, the numerical results demonstrate that the DPG+P-DIGing algorithm not only converges faster than P-DIGing but also achieves a more precise solution compared to DPG.

Constrained logistic regression

We consider the following decentralized logistic regression problem with the MNIST dataset [14]:

$$\min_{x \in \Omega} \sum_{i=1}^{n} f_i(x),$$

where $f_i(x) = \sum_{j=1}^{k} \log[1+\exp\big((-x^T\tau_j)\phi_j\big)]+\frac{\alpha}{2}\|x\|^2$. Here k represents the number of data points, $\tau_j \in \mathbb{R}^{784}$ is the feature vector and $\phi_j \in \{-1, 1\}$ is the corresponding class. It should be noted that we only select the number 1 and 2 classes in the MNIST data and artificially rename these classes as -1 and 1, respectively. In this experiment, we set the number of agents as $n = 20$, with each agent having four neighbors. Additionally, we randomly select 1000 data points from each of class 1 and class 2, respectively, and equally divide this data among the agents. Consequently, each agent is assigned 50 data points from both class 1 and class 2. To obtain x_*, we employ the centralized projected gradient descent algorithm. We measure the quantity $R(t)$ for DPG with a constant stepsize as well as the P-DIGing algorithm (see Figure 4.3). Similar to the non-negative least squares, we artificially select an appropriate constant step size for P-DIGing. As we expected, even with a smaller step size, the DPG algorithm exhibits a faster convergence rate than the P-DIGing algorithms in the initial stages. Consequently, the DPG+P-DIGing algorithm demonstrates a faster convergence compared to the P-DIGing algorithm.

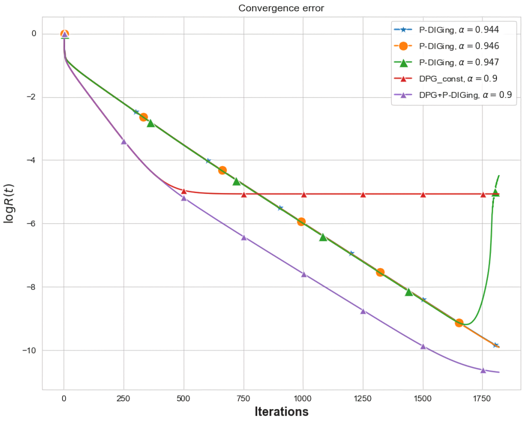


Figure 4.3: The behavior of $\log R(t)$ with P-DIGing, DPG using a constant step size, DPG+P-DIGing.

One-dimensional example

We provide a numerical test for the example considered in Section 4.7. To verify the result of Theorem 4.7.2, we consider the sequence $(x_1(t), x_2(t))$ of (4.7.3) and the following measure

$$R'(t) = \frac{(x_1(t)-1)^2 + \left(x_2(t) - \frac{1}{1-18\alpha}\right)^2}{(x_1(0)-1)^2 + \left(x_2(0) - \frac{1}{1-18\alpha}\right)^2} \quad \text{for } t \geqslant 0. \tag{4.8.1}$$

We test with the stepsizes $\{1/200, 1/100, 1/46, 1/45, 1/43\}$ and the initial values given as

$$(x_1(0), x_2(0)) \in \{(5,10), (100,5)\}. \tag{4.8.2}$$

The graph of the measure $R'(t)$ is provided in Figure 4.4. The result shows that the algorithm (4.7.3) converges to the value $(1, 1/(1-18\alpha))$ for the stepsizes $\{1/200, 1/100, 1/46\}$ as expected by Theorem 4.7.2. Meanwhile, the algorithm (4.7.3) does not converge for the stepsize $\alpha \in \{1/43, 1/45\}$ which is not supported in the interval $(0, 1/45)$ guaranteed by Theorem 4.7.2. These results verify the sharpness of the result of Theorem 4.7.2.

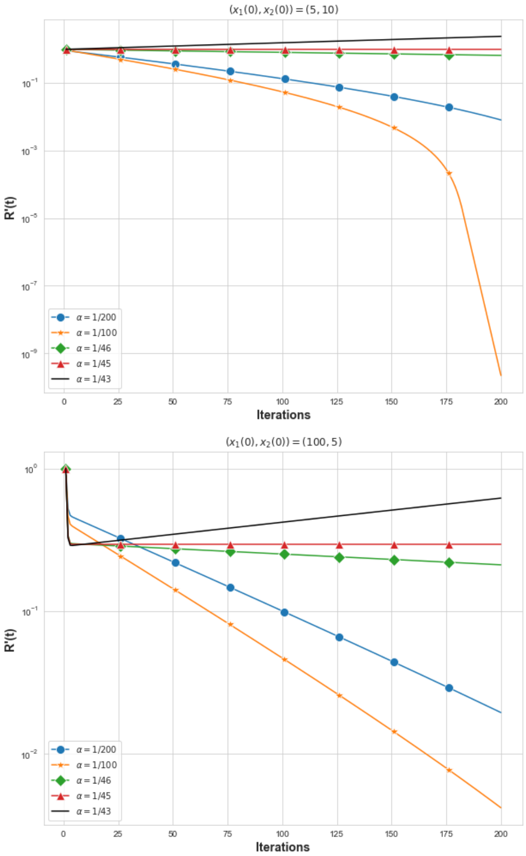


Figure 4.4: The left and right graphs represent the value $R'(t)$ for the initial values $(x_1(0), x_2(0)) = (5, 10)$ and $(100, 5)$ respectively.

Chapter 5

Gradient-push algorithm with even-triggered communication

5.1 Introduction

We again consider the problem (1.0.1) for the unconstrained space, where Ω corresponds to the Euclidean space $\mathbb{R}^d$ but the communication graph is direct and time-varying. In this case, it is hard to apply the DGD algorithm directly since the consensus is not achieved. Instead of the DGD algorithm, we consider the gradient-push (GP) algorithm [31] which is stated as follows:

$$\begin{aligned} w_i(t+1) &= \sum_{j=1}^{m} a_{ij}(t)x_j(t), \\ y_i(t+1) &= \sum_{j=1}^{m} a_{ij}(t)y_j(t), \\ z_i(t+1) &= \frac{w_i(t+1)}{y_i(t+1)}, \\ x_i(t+1) &= w_i(t+1) - \alpha(t+1)\nabla f_i(z_i(t+1)). \end{aligned} \tag{5.1.1}$$

Here $x_i(t) \in \mathbb{R}^d$ and $y_i(t) \in \mathbb{R}$ are the states at time $t \geqslant 0$ handled by the agent i and $y_i(0) = 1$ for all i.

In this chapter, the communication pattern among agents in (1.0.1) at each time $t \in \mathbb{N} \cup \{0\}$ is characterized by a time-varying directed graph $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$. We define in-neighbors and out-neighbors of node i, respectively, as $N_i^{\text{in}}(t) = \{j | (j,i) \in \mathcal{E}(t)\} \cup \{i\}$ and $N_i^{\text{out}}(t) = \{j | (i,j) \in \mathcal{E}(t)\} \cup \{i\}$. Also the out-degree of node i is defined as $d_i^{\text{out}}(t) = |N_i^{\text{out}}(t)|$. Define the mixing matrix $A(t)$ such that

$[A(t)]_{ij} = a_{ij}(t)$, where

$$a_{ij}(t) = \begin{cases} 1/d_j^{\text{out}}(t), \text{ if } i \in N_j^{\text{out}}(t), \\ 0, \text{ otherwise.} \end{cases} \tag{5.1.2}$$

Here $a_{ij}(t)$ is a weight that agent i uses when it receives the state information of agent j.

Simple Case

To gain a more detailed understanding of the GP algorithm, let's consider a simplified case where $A(t) \equiv A$. According to Lemma 2.3.7, we have:

$$A^{\infty} = \lim_{t\to\infty} A^t = \pi \mathbf{1}^T.$$

Here, $\pi = [\pi_1, \pi_2, \cdots, \pi_m]^T$ satisfies (2.3.6). Let $\bar{w}(t) = \frac{1}{m}\sum_{i=1}^{m} w_i(t)$. Then, we can write the GP algorithm (5.1.1) as

$$\begin{aligned} \bar{w}(t+1) &= \bar{w}(t) - \frac{\alpha(t)}{n}\sum_{i=1}^{n} \nabla f_i(z_i(t)) \\ &= \bar{w}(t) - \alpha(t) f(\bar{w}(t)) + \frac{\alpha(t)}{n}\sum_{i=1}^{n} (\nabla f_i(\bar{w}(t)) - \nabla f_i(z_i(t))) \end{aligned}$$

Additionally, we can observe that $w_i(t)/(n\pi_i)$ converges to $\bar{w}(t)$ with an error of $O(\alpha(t))$ (see [11]).

From this expression, it becomes apparent that $\bar{w}(t)$ approximately follows the gradient descent method. Consequently, if we aim for $\bar{x}(t)$ to converge to the minimizer of f, a natural choice for $z_i(t)$ would be $w_i(t)/(n\pi_i)$, which can be estimated by $w_i(t)/y_i(t)$

GP with Event-triggered Communication

As we mentioned in Chapter 1, the GP algorithm requires each agent to communicate with their neighbors at every iteration. Henceforth, we use an event-triggered communication as follows: Each agent i maintains the current states $x_i(t) \in \mathbb{R}^d$ and $y_i(t) \in \mathbb{R}$ and the latest sent states $\hat{x}_i(t) \in \mathbb{R}^d$ and $\hat{y}_i(t) \in \mathbb{R}$ at time t. In a non-time-varying graph, the agents send their current states to neighbors simultaneously only when the differences between the current states and the latest sent states are larger than certain thresholds, which are called trigger times. For each time t, we denote by $\hat{x}_j(t) \in \mathbb{R}^d$ and $\hat{y}_j(t) \in \mathbb{R}$ the latest sent states that agent j

sent to its neighbors at the latest trigger time $\kappa_j^x(t), \kappa_j^y(t)$ up to time t. Then we have

$$\hat{x}_j(s) = \hat{x}_j(\kappa_j^x(t)) = x_j(\kappa_j^x(t)), \quad \text{for all } \kappa_j^x(t) \leqslant s \leqslant t$$

and

$$\hat{y}_j(s) = \hat{y}_j(\kappa_j^y(t)) = y_j(\kappa_j^y(t)), \quad \text{for all } \kappa_j^y(t) \leqslant s \leqslant t.$$

We use the latest sent states to update $x_i(t+1)$ and $y_i(t+1)$ with $a_{ij}(t) = a_{ij}(0)$ for all $t \geqslant 0$, where the value $a_{ij}(t)$ is designed by the agent j as in (5.1.2) depending on the edge information at j. Each agent i sends the states $x_i(t+1)$ and $y_i(t+1)$ to its neighbors respectively if

$$\|x_i(t+1) - \hat{x}_i(t)\| > \tau(t) \tag{5.1.3}$$

and

$$|y_i(t+1) - \hat{y}_i(t)| > \zeta(t), \tag{5.1.4}$$

where $\tau(t), \zeta(t) > 0$ are the thresholds. In a time-varying graph, even though a time t is not a trigger time in the sense of (5.1.3) and (5.1.4), the agent j has to send its current states to new neighbors if the neighbors $N_j^{out}(t)$ is changed. Furthermore, the value $a_{ij}(t)$ should be transmitted to agent $i \in N_j^{out}(t) \cup N_j^{out}(t-1)$ by the agent j. Covering these cases and trying to reduce the communication as much as possible, we impose the following additional rule for transmission in the time-varying graph case:

- If the $N_j^{out}(t)$ is not changed at time t, then agent j does not send $a_{ij}(t)$ to its neighbor $i \in N_j^{out}(t)$.
- If the $N_j^{out}(t)$ is changed at time t, then we follow the rules below.
  - The agent j sends the updated value $a_{ij}(t)$ to its neighbors $k \in N_j^{out}(t)$ and inform the agent $k \in N_j^{out}(t-1) \backslash N_j^{out}(t)$ of that the weight is updated as $a_{kj}(t) = 0$.
  - The agent j sends its latest sent states $\hat{x}_j(t)$ and $\hat{y}_j(t)$ to new neighbors $i \in N_j^{out}(t) \backslash N_j^{out}(t-1)$ which have not received the states since the latest triggering time.

Algorithm 1 is the GP algorithm with event-triggered communication.

Algorithm 1 Distributed Event-Triggered gradient-push algorithm on directed graph

```
Require: Initialize $x_i(0)$ arbitraily and $y_i(0) = \hat{y}_i(0) = 1$ for all $i \in \{1, \cdots, m\}$.
    Set $\hat{x}_i(0) = x_i(0)$
for $t = 0, 1, \cdots,$ do
    if $t = 0$, or $\kappa_i^x(t)$ is updated then
        Send $\hat{x}_i(t)$ to its negihbors simultaneously.
    else
        Send information using the transmission rules
    end if
    if $t = 0$, or $\kappa_i^y(t)$ is updated then
        Send $\hat{y}_i(t)$ to its negihbors simultaneously.
    else
        Send information using the transmission rules
    end if
    Compute the new action as
        $\widehat{w}_i(t+1) = \sum_{j=1}^{m} a_{ij}(t)\hat{x}_j(t),$    (5.1.5)
        $y_i(t+1) = \sum_{j=1}^{m} a_{ij}(t)\hat{y}_j(t),$    (5.1.6)
        $\widehat{z}_i(t+1) = \frac{\widehat{w}_i(t+1)}{y_i(t+1)},$    (5.1.7)
        $x_i(t+1) = \widehat{w}_i(t+1) - \alpha(t+1)\nabla f_i(\widehat{z}_i(t+1)).$    (5.1.8)
    if $\|x_i(t+1) - \hat{x}_i(t)\| \geqslant \tau(t+1)$ then
        Set $\hat{x}_i(t+1) = x_i(t+1)$ and update $\kappa_i^x(t+1)$.
    else
        Set $\hat{x}_i(t+1) = \hat{x}_i(t)$ and do not send
    end if
    if $|y_i(t+1) - \hat{y}_i(t)| \geqslant \zeta(t+1)$ then
        Set $\hat{y}_i(t+1) = y_i(t+1)$ and update $\kappa_i^y(t+1)$
    else
        Set $\hat{y}_i(t+1) = \hat{y}_i(t)$ and do noet send
    end if
end for
```

5.2 Preliminaries

In this section, we introduce the assumptions and constants that will be used throughout. We begin with the following assumptions.

Assumption 5.2.1.

(A) For each $i \in \{1, \cdots, m\}$, there exists $D_i > 0$ such that $\|\nabla f_i(x)\| \leqslant D_i$ for all $x \in \mathbb{R}^d$.

(B) The sequence of graph $\{\mathcal{G}(t)\}_{t\in\mathbb{N}}$ is uniformly strongly connected, i.e., there exists a value $B \in \mathbb{N}$ such that the graph with edge set $\cup_{i=kB}^{(k+1)B-1}\mathcal{E}(i)$ is strongly connected for any $k \geqslant 0$.

(C) The sequence of step size $\{\alpha(t)\}_{t\in\mathbb{N}}$ is monotonically non-increasing and satisfies

$$\sum_{t=1}^{\infty} \alpha(t) = \infty, \ \sum_{t=1}^{\infty} \alpha(t)^2 < \infty.$$

(D) The sequence of event-triggering thresholds $\{\tau(t)\}_{t\in\mathbb{N}}$ is monotonically non-increasing for $t \geqslant 1$ and we set $\tau(0) = 0$. In addition, the sequence satisfies

$$\sum_{t=0}^{\infty} \tau(t) < \infty.$$

(E) The sequence of event-triggering thresholds $\{\zeta(t)\}_{t\in\mathbb{N}}$ is monotonically non-increasing for $t \geqslant 0$ and we set $\zeta(t) = 0$. In addition, the sequence satisfies

$$\sum_{t=0}^{\infty} t^{3/2}\zeta(t) < \infty, \quad \sum_{t=0}^{\infty} \zeta(t) < 1.$$

Note that $\sum_{t=0}^{\infty} t^{3/2}\zeta(t) < \infty$ implies that there exists a finite M such that $\sum_{t=0}^{\infty} \zeta(t) = M$. If we set a new sequence $\{\tilde{\zeta}(t)\}_{t\in\mathbb{N}}$ by $\tilde{\zeta}(t) = \zeta(t)/(M+1)$, then it satisfy $\sum_{t=0}^{\infty} \tilde{\zeta}(t) < 1$. Hence if we have a sequence $\{\zeta(t)\}_{t\in\mathbb{N}}$ satisfying $\sum_{t=0}^{\infty} t^{3/2}\zeta(t) < \infty$, then we may divide the sequence by a positive constant to satisfy (E) of Assumption 5.2.1. One example of the sequence that satisfies (E) of Assumption 5.2.1 is $\zeta(t) = \frac{1}{3t^3}$ for $t \geqslant 1$. In addition the assumption on the triggering thresholds include the case that thresholds have exponential decays. This exponential decay assumption is commonly used in the literature [29, 29, 30]. However, we mention that the above assumptions imposed for the convergence analysis may not be optimal and it will be interesting to weaken these assumptions. Next, we define the following constants which are frequently used throughout this chapter.

- $D = \max_{1 \leqslant i \leqslant m} D_i$.
- $Q := \inf_{t=0,1,\cdots} \min_{1 \leqslant i \leqslant m} [A(t:0)\mathbf{1}]_i$
- $E_\tau(T) = \sum_{t=0}^T \tau(t), \quad E_{\tau,2}(T) = \sum_{t=0}^T \tau(t)^2, \quad E_\tau = \sum_{t=0}^\infty \tau(t)$
- $F_\zeta(T) = \sum_{t=0}^T \zeta(t), \quad F_\zeta = \sum_{t=0}^\infty \zeta(t), \quad F_{\zeta_{3/2}} = \sum_{t=0}^\infty t^{3/2}\zeta(t)$
- $m_\zeta = \mathbf{1}_m^T y(0) + \sum_{s=1}^\infty \mathbf{1}_m^T \theta(s) = m + \sum_{s=1}^\infty \mathbf{1}_m^T \theta(s)$, where $\theta(s) = \hat{y}(s) - y(s)$.
- $\beta(t) = m\Big(\Big(F_\zeta - F_\zeta(t)\Big) + C_0\lambda^t + C_0\lambda^{t/2}F_\zeta(t) + \frac{\zeta([t/2]+1)}{1-\lambda}\Big)$
- $B_\zeta = \frac{m_\zeta}{m}$. and $K(t) = \frac{\beta(t)}{m_\zeta}$
- $\delta := \min_{1 \leqslant i \leqslant m} \inf_{t \in \mathbb{N}} y_i(t) > 0$

Note that positivity of δ is proved in Lemma 5.3.2 under (E) of Assumption 5.2.1. Lastly, we introduce the following lemma for the mixing matrix defined by (5.1.2).

Lemma 5.2.2 ([31])**.** *Suppose that the graph sequence $\{\mathcal{G}(t)\}$ is uniformly strongly connected. Then, the following statements are valid.*

1. *For each integer $s \geqslant 0$, there is a stochastic vector $\phi(s)$ such that for all i, j and $t \geqslant s$*
$$|[A(t:s)]_{ij} - \phi_i(t)| \leqslant C_0\lambda^{t-s} \tag{5.2.1}$$
for some values $C_0 \geqslant 1$ and $\lambda \in (0,1)$ depending on the graph sequence.

2. *The following inequality holds.*
$$\inf_{t=0,1,\cdots} \min_{1 \leqslant i \leqslant m} [A(t:0)\mathbf{1}]_i \geqslant 1/n^{nB}. \tag{5.2.2}$$

Here we denote by $A(t:s)$ the matrix given as

$$A(t:s) = A(t)A(t-s)\cdots A(s) \quad \textit{for all } t \geqslant s \geqslant 0.$$

5.3 Sequence for Achieving Consensus

A convergence property of $\{y_i(t)\}_{t\in\mathbb{N}}$ and their positive uniform lower bound are key points in proving our main results. Let us first look at the case without event-triggering ($\zeta(t) = 0$), which means all in-neighbors of agent i share the $y_i(t)$ with this agent for every time step. In this case, since $y(t) = \hat{y}(t)$ for all $t \in \mathbb{N}$, it holds that

$$y(t) = A(t-1:0)\mathbf{1}_m \tag{5.3.1}$$

by (5.1.6) in Algorithm 1. Hence we can directly show that $y(t)$ converges to $\phi(t)$ and has a uniform lower bound Q using Lemma 5.2.2. In the event-triggered case, $y(t)$ can be written as

$$y(t) = A(t-1:0)\mathbf{1}_m + \sum_{s=1}^{t-1} A(t-1:s)\theta(s). \tag{5.3.2}$$

Therefore the convergence and uniform lower boundedness property may not hold due to the additional term

$$\sum_{s=1}^{t-1} A(t-1:s)\theta(s).$$

The following lemma shows that $y(t)$ has a positive uniform lower bound δ and converges to $m_\zeta \phi(t)$ instead of $\phi(t)$ under (E) of Assumption 5.2.1.

Lemma 5.3.1. *Suppose that (E) of Assumption 5.2.1 holds. Then we have*

$$m_\zeta \geqslant (1-F_\zeta)m \tag{5.3.3}$$

and the following estimate holds:

$$\left\| y(t+1) - m_\zeta \phi(t) \right\|_\infty \leqslant \beta(t) \quad \forall\ t \in \mathbb{N}. \tag{5.3.4}$$

In addition, we have

$$\lim_{t\to\infty} t^{3/2}\beta(t) = 0.$$

Proof. Observe that $|\theta_i(s)| = |\hat{y}_i(s) - y_i(s)| \leqslant \zeta(s)$ for $s \geqslant 1$ by the event-triggering condition, and so

$$\Big| \sum_{s=1}^{\infty} 1_m^T \theta(s) \Big| \leqslant m \sum_{s=1}^{\infty} \zeta(s) = mF_\zeta.$$

Using this, we get

$$m_\zeta = m + \sum_{s=1}^{\infty} 1_m^T \theta(s) \geqslant (1-F_\zeta)m,$$

which proves (5.3.3).

Next we prove (5.3.4). For each $s \geqslant 0$, by definition we have

$$y(s+1) = A(s)\hat{y}(s) = A(s)(y(s) + \theta(s)),$$

where we have set $\theta(0) = 0$. Using this iteratively gives the following formula

$$\begin{aligned} y(t+1) &= A(t:0)y(0) + \sum_{s=1}^{t} A(t:s)\theta(s) \\ &= \phi(t)\Big[\mathbf{1}_m^T y(0) + \sum_{s=1}^{t} \mathbf{1}_m^T \theta(s)\Big] \\ &\quad + (A(t:0) - \phi(t)1_m^T)y(0) + \sum_{s=1}^{t} \Big[(A(t:s) - \phi(t)\mathbf{1}_m^T)\theta(s)\Big]. \end{aligned}$$

Since $|\theta_j(s)| \leqslant \zeta(s)$, we find

$$\sum_{s=t+1}^{\infty} |\theta_j(s)| \leqslant \sum_{s=t+1}^{\infty} \zeta(s) = F_\zeta - F_\zeta(t).$$

Using the above inequality, we obtain

$$\Big|m_\zeta - 1_m^T y(0) - \sum_{s=0}^{t} 1_m^T \theta(s)\Big| \quad = \Big|\sum_{s=t+1}^{\infty} 1_m^T \theta(s)\Big| \leqslant \Big(F_\zeta - F_\zeta(t)\Big)m. \tag{5.3.5}$$

Hence we have

$$\begin{aligned} \Big\|y(t+1) - m_\zeta \phi(t)\Big\|_\infty \leqslant \Big(F_\zeta - F_\zeta(t)\Big)m + \Big\|(A(t:0) - \phi(t)1_m^T)y(0)\Big\|_\infty \\ + \Big\|\sum_{s=1}^{t} \Big[(A(t:s) - \phi(t)1_m^T)\theta(s)\Big]\Big\|_\infty. \end{aligned} \tag{5.3.6}$$

Now we estimate the second and third terms in the right hand side of (5.3.6). Using (5.2.1) we have

$$\Big\|(A(t:0) - \phi(t)1_m^T)y(0)\Big\|_\infty \leqslant mC_0\lambda^t \tag{5.3.7}$$

and

$$\begin{aligned} \Big\|\sum_{s=1}^{t} \Big[(A(t:s) - \phi(t)1_m^T)\theta(s)\Big]\Big\|_\infty &\leqslant mC_0 \sum_{s=1}^{t} \lambda^{t-s}\zeta(s) \\ &\leqslant mC_0\bigg(\sum_{s=1}^{\lfloor t/2 \rfloor} \lambda^{t-s}\zeta(s) + \sum_{s=\lfloor t/2 \rfloor+1}^{t} \lambda^{t-s}\zeta(s)\bigg) \\ &\leqslant mC_0\bigg(\lambda^{t/2}F_\zeta(t) + \zeta(\lfloor t/2 \rfloor + 1) \sum_{s=\lfloor t/2 \rfloor+1}^{t} \lambda^{t-s}\bigg) \\ &\leqslant mC_0\bigg(\lambda^{t/2}F_\zeta(t) + \frac{\zeta(\lfloor t/2 \rfloor + 1)}{1-\lambda}\bigg). \end{aligned} \tag{5.3.8}$$

Putting the estimates (5.3.7) and (5.3.8) in (5.3.6), we get

$$\left\| y(t+1) - \phi(t)m_\zeta \right\|_\infty \leqslant m\bigg(\Big(F_\zeta - F_\zeta(t)\Big) + C_0\lambda^t + C_0\lambda^{t/2}F_\zeta(t) + \frac{\zeta([t/2]+1)}{1-\lambda} \bigg).$$

This proves the second assertion of the lemma.

Now we shall show that $\lim_{t\to\infty} t^{3/2}\beta(t) = 0$. Since $\lambda \in (0,1)$, it suffices to show that

$$\lim_{t\to\infty} t^{3/2}(F_\zeta - F_\zeta(t) + \zeta([t/2])) = 0.$$

This fact follows directly from the fact that $\sum_{t=0}^{\infty} t^{3/2}\zeta(t) < \infty$ and the following inequality

$$t^{3/2}(F_\zeta - F_\zeta(t)) = t^{3/2}\sum_{s=t+1}^{\infty}\zeta(s) \leqslant \sum_{s=t+1}^{\infty} s^{3/2}\zeta(s).$$

The proof is done. □

Lemma 5.3.2. *Suppose that* (*B*) *and* (*E*) *of Assumption 5.2.1 hold. Then the value* δ *is positive.*

Proof. Note that by Lemma 5.2.2, it follows that

$$\begin{aligned} m\phi_i(t) &= \sum_{j=1}^{m}[A(t:0)]_{ij} + \sum_{j=1}^{m}\Big(\phi_i(t) - [A(t:0)]_{ij}\Big) \\ &\geqslant Q - mC_0\lambda^t. \end{aligned}$$

Using the above inequality and Lemma 5.3.1, we deduce for each $1 \leqslant i \leqslant m$ the following estimate

$$\begin{aligned} y_i(t+1) &\geqslant m_\zeta\phi_i(t) - \beta(t) \\ &\geqslant (m_\zeta/m)Q - m_\zeta C_0\lambda^t - \beta(t). \end{aligned}$$

Since $\beta(t)$ converges to zero as t goes to infinity and $\lambda \in (0,1)$, there exists a time $T \in \mathbb{N}$ and a constant $\tilde{\delta} > 0$ such that for any $t \geqslant T$,

$$y_i(t+1) \geqslant \tilde{\delta}. \tag{5.3.9}$$

Note that by (B) of Assumption 5.2.1, each matrix $A(t)$ has no zero row. This fact, together with the definition of $\hat{y}$ and (5.1.6), for any $t \in \mathbb{N}$ we have

$$\min_{1\leqslant i\leqslant m} y_i(t) > 0. \tag{5.3.10}$$

Therefore, combining (5.3.9) with (5.3.10), we conclude that δ satisfies

$$\delta \geqslant \min_{1\leqslant i\leqslant m}\{y_i(0), y_i(1), \cdots, y_i(T), \tilde{\delta}\} > 0.$$

The proof is done. □

5.4 Consensus Estimate

In this section, we derive a bound of the disagreement in agent estimates $\{\hat{z}_i(t)\}_{i=1}^m$ that will be used in the proofs of the main theorems. In the case without event-triggering ($\tau(t) = \zeta(t) = 0$), the paper [31] proved that $\|\hat{z}_i(t+1) - \bar{x}(t)\|$ converges to zero for the step size satisfying (C) of Assumption 5.2.1 as t goes to infinity. For the event-triggered case, the following proposition shows that the values $\{\hat{z}_i(t)\}_{i=1}^m$ approach $B_\zeta \bar{x}(t)$ instead of $\bar{x}(t)$ as t goes to infinity due to the effect of the threshold $\zeta(t)$ for the triggering condition of $\{y_i(t)\}_{i=1}^m$.

Proposition 5.4.1. *Suppose that Assumption 5.2.1 holds. Then for any $t \geqslant 1$ we have*

$$\begin{aligned}\|\hat{z}_i(t+1) - B_\zeta \bar{x}(t)\| \leqslant \frac{1}{\delta}\Big(C_0\lambda^t + K(t)\Big)\|x(0)\|_1 \\ + \frac{m}{\delta}\sum_{s=0}^{t-1}\Big[C_0\lambda^{t-s-1} + K(t)\Big]\Big(\alpha(s+1)D + \tau(s)\Big) \\ + \frac{d_i(t)\tau(t)}{\delta},\end{aligned}$$

and for $t = 0$ we have

$$\|\hat{z}_i(1) - \bar{x}(0)\| \leqslant \frac{2C_0}{\delta}\|x(0)\|,$$

where the constant $\delta > 0$ satisfies $y_i(t) > \delta$ for all $t > 0$.

To prove Proposition 5.4.1, we consider a variable $w_i(t+1) \in \mathbb{R}^d$ which is a companion to the variable $\hat{w}_i(t+1) \in \mathbb{R}^n$ defined as

$$w_i(t+1) = \sum_{j=1}^m a_{ij}(t)x_j(t), \tag{5.4.1}$$

and their difference

$$e_i(t+1) = \hat{w}_i(t+1) - w_i(t+1). \tag{5.4.2}$$

Then we may rewrite the gradient step (5.1) as

$$x_i(t+1) = w_i(t+1) - \alpha(t+1)\nabla f_i(\hat{z}_i(t+1)) + e_i(t+1). \tag{5.4.3}$$

Summing up (5.4.3) for $1 \leqslant i \leqslant m$ and using that $A(t)$ is column-stochastic, we have

$$\bar{x}(t+1) = \bar{x}(t) - \frac{\alpha(t+1)}{m}\sum_{i=1}^m \nabla f_i(\hat{z}_i(t+1)) + \frac{1}{m}\sum_{i=1}^m e_i(t+1). \tag{5.4.4}$$

Now we find a bound of $e_i(t+1)$ which is the difference between $w_i(t+1)$ and $\hat{w}_i(t+1)$ associated to the event-triggering $\tau(t)$.

Lemma 5.4.2. *Suppose that* (B) *of Assumption 5.2.1 holds. The quantity* $e_i(t+1)$ *defined in* (5.4.2) *satisfies*

$$\|e_i(t+1)\| \leqslant d_i(t)\tau(t), \tag{5.4.5}$$

where $d_i(t) = \sum_{j=1}^m a_{ij}(t)$. *In addition, we have*

$$\sum_{i=1}^m \|e_i(t+1)\| \leqslant m\tau(t). \tag{5.4.6}$$

Proof. By using the triggering condition, we have

$$\begin{aligned}\|e_i(t+1)\| &\leqslant \|\hat{w}_i(t+1) - w_i(t+1)\| \\ &\leqslant \left\| \sum_{j=1}^m a_{ij}(t)(\hat{x}_j(t) - x_j(t)) \right\| \\ &\leqslant \sum_{j=1}^m a_{ij}(t)\|\hat{x}_j(t) - x_j(t)\| \leqslant d_i(t)\tau(t),\end{aligned}$$

which proves (5.4.5). Summing this over $1 \leqslant i \leqslant m$ and using that $A(t)$ is column stochastic, we find

$$\begin{aligned}\sum_{i=1}^m \|e_i(t+1)\| &\leqslant \sum_{i=1}^m \sum_{j=1}^m \Big(a_{ij}(t)\tau(t)\Big) \\ &= \sum_{j=1}^m \Big(\sum_{i=1}^m a_{ij}(t)\Big)\tau(t) = m\tau(t).\end{aligned}$$

The proof is finished. □

Now we are ready to prove Proposition 5.4.1.

Proof of Proposition 5.4.1. We regard $x_k(t)$ as a row vector in $\mathbb{R}^{1\times d}$ and define the variables $x(t) \in \mathbb{R}^{m\times d}$, $\nabla f(\hat{z}(t)) \in \mathbb{R}^{m\times d}$, and $e(t) \in \mathbb{R}^{m\times d}$ as

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{bmatrix}, \ \nabla f(\hat{z}(t)) = \begin{bmatrix} \nabla f_1(\hat{z}_1(t)) \\ \vdots \\ \nabla f_m(\hat{z}_m(t)) \end{bmatrix}, \ e(t) = \begin{bmatrix} e_1(t) \\ \vdots \\ e_m(t) \end{bmatrix}.$$

Note that by (5.4.2) and (5.1.7), we have

$$\hat{z}_i(t+1) = \frac{w_i(t+1) + e_i(t+1)}{y_i(t+1)}.$$

Also we see that

$$B_\zeta \bar{x}(t) = \frac{m\bar{x}(t)}{m_\zeta} = \frac{\mathbf{1}_m^T x(t)}{m_\zeta}.$$

Using these formulas and (5.4.1) we have

$$\begin{aligned}\hat{z}_i(t+1) - B_\zeta \bar{x}(t) &= \frac{w_i(t+1) + e_i(t+1)}{y_i(t+1)} - \frac{\mathbf{1}_m^T x(t)}{m_\zeta} \\ &= \frac{1}{y_i(t+1)}\left([A(t)x(t)]_i - y_i(t+1)\frac{1_m^T x(t)}{m_\zeta}\right) + \frac{e_i(t+1)}{y_i(t+1)}.\end{aligned} \tag{5.4.7}$$

To estimate the first term on the right hand side of the last equality, we rewrite (5.4.3) as

$$x(t+1) = A(t)x(t) - \alpha(t+1)\nabla f(\hat{z}(t+1)) + e(t+1).$$

Using this formula recursively, for $t \geqslant 1$ we have

$$A(t)x(t) = A(t:0)x(0) - \sum_{s=0}^{t-1} A(t:s+1)\varepsilon(s), \tag{5.4.8}$$

where we have let

$$\varepsilon(s) = \alpha(s+1)\nabla f(\hat{z}(s+1)) - e(s+1).$$

Using (A) of Assumption 5.2.1 and (5.4.6) we have the following bound

$$\|\varepsilon(s)\|_1 \leqslant m\Big(\alpha(s+1)D + \tau(s)\Big). \tag{5.4.9}$$

Since $A(t)$ is column stochastic we have $1_m^T A(t) = 1_m^T$, and combine this with (5.4.8) to have

$$\mathbf{1}_m^T x(t) = \mathbf{1}_m^T x(0) - \sum_{s=0}^{t-1} \mathbf{1}_m^T \varepsilon(s). \tag{5.4.10}$$

Combining (5.4.8) and (5.4.10) yields

$$\begin{aligned}&A(t)x(t) \\ &= \phi(t)\mathbf{1}_m^T x(t) + \Big(A(t:0) - \phi(t)\mathbf{1}_m^T\Big)x(0) - \sum_{s=0}^{t-1}\Big(A(t:s+1) - \phi(t)\mathbf{1}_m^T\Big)\varepsilon(s),\end{aligned} \tag{5.4.11}$$

where $\phi(t)$ is the stochastic vector satisfying (5.2.1). By Lemma 5.3.1, for $y(t) := [y_1(t), \cdots, y_m(t)]^T \in \mathbb{R}^{m\times 1}$ we have

$$y(t+1) = m_\zeta \phi(t) + r(t),$$

where $r(t)$ satisfies $\|r(t)\|_\infty \leqslant \beta(t)$. Combining this with (5.4.11), we obtain

$$\begin{aligned}
&[A(t)x(t)]_i - y_i(t+1)\frac{1_m^T x(t)}{m_\zeta} \\
&= \phi_i(t)1_m^T x(t) + [(A(t:0) - \phi(t)1_m^T)x(0)]_i - \sum_{s=0}^{t-1}\Big[(A(t:s+1) - \phi(t)1_m^T)\varepsilon(s)\Big]_i \\
&\qquad - [m_\zeta \phi_i(t) + r_i(t)]\frac{1_m^T x(t)}{m_\zeta} \\
&= [(A(t:0) - \phi(t)\mathbf{1}_m^T)x(0)]_i - \sum_{s=0}^{t-1}\Big[(A(t:s+1) - \phi(t)\mathbf{1}_m^T)\varepsilon(s)\Big]_i - r_i(t)\frac{1_m^T x(t)}{m_\zeta}.
\end{aligned}$$

By applying (5.2.1) here, we deduce

$$\begin{aligned}
&\left\| [A(t)x(t)]_i - y_i(t+1)\frac{1_m^T x(t)}{m} \right\| \\
&\leqslant C_0\lambda^t \|x(0)\|_1 + \sum_{s=0}^{t-1} C_0 \lambda^{t-s-1}\|\varepsilon(s)\|_1 + K(t)\|1_m^T x(t)\|,
\end{aligned} \tag{5.4.12}$$

where $K(t) = \beta(t)/m_\zeta$. From (5.4.10) we find the following estimate

$$\|1_m^T x(t)\| \leqslant \|x(0)\|_1 + \sum_{s=0}^{t-1}\|\varepsilon(s)\|_1.$$

Combining this with (5.4.12) and using (5.4.9), we obtain

$$\begin{aligned}
&\left\| [A(t)x(t)]_i - y_i(t+1)\frac{1_m^T x(t)}{m} \right\| \\
&\leqslant C_0\lambda^t \|x(0)\|_1 + \sum_{s=0}^{t-1} C_0 \lambda^{t-s-1}\|\varepsilon(s)\|_1 + K(t)\Big(\|x(0)\|_1 + \sum_{s=0}^{t-1}\|\varepsilon(s)\|_1\Big) \\
&\leqslant \Big(C_0\lambda^t + K(t)\Big)\|x(0)\|_1 + m\sum_{s=0}^{t-1}\Big[C_0\lambda^{t-s-1} + K(t)\Big]\Big(\alpha(s+1)D + \tau(s)\Big).
\end{aligned} \tag{5.4.13}$$

By applying Lemma 5.3.2, (5.4.5) and the above inequality to the norm of (5.4.7), we obtain

$$\begin{aligned}
&\|\hat{z}_i(t+1) - B_\zeta \bar{x}(t)\| \\
&\leqslant \frac{1}{\delta}\Big(C_0\lambda^t + K(t)\Big)\|x(0)\|_1 + \frac{m}{\delta}\sum_{s=0}^{t-1}\Big[C_0\lambda^{t-s-1} + K(t)\Big]\Big(\alpha(s+1)D + \tau(s)\Big) \\
&\qquad + \frac{d_i(t)\tau(t)}{\delta},
\end{aligned}$$

It remains to estimate the case $t = 0$. By the algorithm, we have

$$\hat{z}_i(1) - \bar{x}(0) = \frac{\hat{w}_i(1)}{y_i(1)} - \bar{x}(0) = \frac{\sum_{j=1}^m a_{ij}(0)x_j(0)}{\sum_{j=1}^m a_{ij}(0)} - \bar{x}(0).$$

Using this we find

$$\|z_i(1) - \bar{x}(0)\| \leqslant \frac{1}{\delta}\|x(0)\|_1 + \frac{1}{m}\|x(0)\|_1 \leqslant 2\|x(0)\|_1 \leqslant \frac{2C_0}{\delta}\|x(0)\|.$$

The proof is finished. □

By utilizing Proposition 5.4.1, we analyze the relation between $\hat{z}_i(t)$ and $B_\zeta\bar{x}(t)$ under the assumptions on $\{\alpha(t)\}_{t\in\mathbb{N}}$, $\{\tau(t)\}_{t\in\mathbb{N}}$ and $\{\zeta(t)\}_{t\in\mathbb{N}}$ of the main theorems. To do this, we first recall a useful lemma from [36].

Lemma 5.4.3 ([36])**.** *If* $\lim_{k\to\infty}\gamma_k = \gamma$ *and* $0 < \beta < 1$*, then*

$$\lim_{k\to\infty}\sum_{l=0}^{k}\beta^{k-l}\gamma_l = \frac{\gamma}{1-\beta}$$

Corollary 5.4.4. *Suppose that* $(A), (B), (D)$ *and (E) of Assumption 5.2.1 hold. Also, assume that the step size* $\alpha(t)$ *satisfies* (B) *of Assumption 5.2.1 or* $\alpha(t) = 1/\sqrt{t}$*. Then we have*

$$\lim_{t\to\infty}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\| = 0 \text{ for all } i.$$

Proof. We recall from Proposition 5.4.1 the following inequality

$$\begin{aligned}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\| \leqslant \frac{1}{\delta}\Big(C_0\lambda^t + K(t)\Big)\|x(0)\|_1 \\ + \frac{m}{\delta}\sum_{s=0}^{t-1}\Big[C_0\lambda^{t-s-1} + K(t)\Big]\Big(\alpha(s+1)D + \tau(s)\Big) \quad (5.4.14)\\ + \frac{d_i(t)\tau(t)}{\delta},\end{aligned}$$

We notice that $\lim_{t\to\infty} t^{3/2}K(t) = 0$ by Lemma 5.3.1. From this and the boundedness of $\alpha(s)$ and $\tau(s)$, it easily follows that

$$\lim_{t\to\infty}\frac{1}{\delta}K(t)\|x(0)\|_1 + m\sum_{s=0}^{t-1}K(t)\Big(\alpha(s+1)D + \tau(s)\Big) = 0.$$

In addition, by $(C), (D)$ and (E) of Assumption 5.2.1, we know that $\lim_{s\to\infty}\alpha(s+1) = 0$, $\lim_{s\to\infty}\tau(s) = 0$ and $\lim_{s\to\infty}\zeta(s) = 0$. Using this fact with Lemma 5.4.3 in the right hand side of (5.4.14), we deduce

$$\lim_{t\to\infty}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\| = 0,$$

which completes the proof. □

Corollary 5.4.5. *Suppose that Assumption 5.2.1 hold. Let* $\alpha(t) = \frac{1}{\sqrt{t}}$. *Then we have*

$$\begin{aligned}
&\sum_{t=0}^{T} \alpha(t+1)\|\hat{z}_i(t+1) - B_\varsigma \bar{x}(t)\| \\
&\leqslant \frac{C_0}{\delta(1-\lambda)}\|x(0)\|_1 + \frac{4mC_0 E_\tau(T)}{\delta(1-\lambda)} + \frac{C_0 m D}{\delta(1-\lambda)}(1 + \ln(T)) \\
&\quad + \frac{1}{\delta}\sum_{t=0}^{T} K(t)\alpha(t+1)\Big[\|x(0)\|_1 + \sum_{s=0}^{t-1}(\alpha(s+1)D + \tau(s))\Big].
\end{aligned}$$

Proof. By Proposition 5.4.1, we have

$$\begin{aligned}
&\delta \sum_{t=0}^{T} \alpha(t+1)\|\hat{z}_i(t+1) - B_\varsigma \bar{x}(t)\| \leqslant \\
&\quad \|x(0)\|_1 \sum_{t=0}^{T} (C_0\lambda^t + K(t))\alpha(t+1) \\
&+ m\sum_{t=0}^{T} \alpha(t+1) \sum_{s=0}^{t-1}(C_0\lambda^{t-s-1} + K(t))(\alpha(s+1)D + \tau(s)) \\
&+ \sum_{t=0}^{T} \alpha(t+1)(d_i(t)\tau(t)).
\end{aligned} \tag{5.4.15}$$

The terms involving $K(t)$ are fit to the inequality of the corollary. Let us estimate each summation not involving $K(t)$ in the right hand side. Using that $\alpha(t) \leqslant 1$, the first term is bounded with

$$\sum_{t=0}^{T} \alpha(t+1)\lambda^t \leqslant \sum_{t=0}^{T} \lambda^t \leqslant \frac{1}{1-\lambda}. \tag{5.4.16}$$

The last term is bounded using

$$\sum_{t=0}^{T} \alpha(t+1)(d_i(t)\tau(t)) \leqslant m\sum_{t=0}^{T} \tau(t) = mE_\tau(T). \tag{5.4.17}$$

We estimate the second term using

$$\begin{aligned}\sum_{t=0}^{T}\alpha(t+1)\sum_{s=0}^{t-1}\lambda^{t-s-1}\alpha(s+1) &= \sum_{t=1}^{T+1}\frac{1}{\sqrt{t}}\sum_{s=1}^{t}\lambda^{t-s}\frac{1}{\sqrt{s}}\\ &\leqslant \sum_{t=1}^{T+1}\sum_{s=1}^{t}\lambda^{t-s}\frac{1}{s}\\ &= \sum_{s=1}^{T+1}\frac{1}{s}\sum_{t=s}^{T+1}\lambda^{t-s}\\ &\leqslant \frac{1+\ln(T+1)}{1-\lambda}.\end{aligned} \tag{5.4.18}$$

In order to estimate the third term, we estimate

$$\begin{aligned}\sum_{s=0}^{t-1}\lambda^{t-s-1}\tau(s) &= \sum_{s=0}^{\lfloor (t-1)/2\rfloor}\lambda^{t-s-1}\tau(s) + \sum_{s=\lfloor (t-1)/2\rfloor+1}^{t-1}\lambda^{t-s-1}\tau(s)\\ &\leqslant \lambda^{(t-1)/2}\sum_{s=0}^{\lfloor (t-1)/2\rfloor}\tau(s) + \tau(\lfloor t/2\rfloor)\sum_{s=\lfloor (t-1)/2\rfloor+1}^{t-1}\lambda^{t-s-1}\\ &\leqslant \lambda^{(t-1)/2}E_\tau(T) + \frac{\tau(\lfloor t/2\rfloor)}{1-\lambda}.\end{aligned}$$

Using this we derive

$$\begin{aligned}\sum_{t=0}^{T}\alpha(t+1)\Big[\sum_{s=0}^{t-1}\lambda^{t-s-1}\tau(s)\Big] &\leqslant E_\tau(T)\sum_{t=0}^{T}\frac{\lambda^{(t-1)/2}}{\sqrt{t+1}} + \frac{1}{1-\lambda}\sum_{t=0}^{T}\frac{\tau(\lfloor t/2\rfloor)}{\sqrt{t+1}}\\ &\leqslant \frac{E_\tau(T)}{1-\sqrt{\lambda}} + \frac{E_\tau(T)}{1-\lambda} < \frac{3E_\tau(T)}{1-\lambda}.\end{aligned} \tag{5.4.19}$$

Putting the above estimates (5.4.16)-(5.4.19) in (5.4.15), we obtain

$$\begin{aligned}&\sum_{t=0}^{T}\alpha(t+1)\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|\\ &\leqslant \frac{C_0}{\delta(1-\lambda)}\|x(0)\|_1 + \frac{4mC_0E_\tau(T)}{\delta(1-\lambda)} + \frac{C_0mD}{\delta(1-\lambda)}(1+\ln(T))\\ &\quad + \frac{1}{\delta}\sum_{t=0}^{T}K(t)\alpha(t+1)\Big[\|x(0)\|_1 + \sum_{s=0}^{t-1}(\alpha(s+1)D + \tau(s))\Big].\end{aligned}$$

which finishes the proof. □

5.5 Convergence Estimates

In this section we prove our main results, namely Theorems 5.5.1 and 5.5.2. In the previous section, we obtained consensus estimates. Especially, Corollary 5.4.4 and 5.4.5 investigate the difference between the variable $\hat{z}_i(t)$ in the Algorithm 1 and $B_\zeta \bar{x}(t)$. Based upon these results, Theorem 5.5.1 and 5.5.2 can be proved by comparing the cost values computed at the points $B_\zeta \bar{x}(t)$ and x^*. Our first result establishes the convergence of $\hat{z}_i(t)$ to the optimal solutions for an arbitrary step size $\alpha(t)$ satisfying (C) of Assumption 5.2.1, and event-triggering thresholds $\tau(t)$ and $\zeta(t)$ satisfying (D) and (E) of Assumption 5.2.1.

Theorem 5.5.1. *Suppose that Assumption 5.2.1 holds. Then the sequence $\{\hat{z}_i(t)\}_{t\in\mathbb{N}}$ for $1 \leqslant i \leqslant n$ of the Algorithm 1 satisfies the following property:*

$$\lim_{t\to\infty} \hat{z}_i(t) = x^* \text{ for all } i \text{ and for some } x^* \in X^*.$$

Next we consider the Algorithm 1 with specific step size $\alpha(t) = 1/\sqrt{t}$. This step size does not satisfy (C) of Assumption 5.2.1, but we may obtain an explicit convergence rate as in the following result.

Theorem 5.5.2. *Suppose that $(A), (B), (D)$ and (E) of Assumption 5.2.1 hold. Let $\alpha(t) = \frac{1}{\sqrt{t}}$ for $t \geqslant 1$. Define $H(-1) = 1$ and $H(t) := \prod_{k=0}^{t}(1 + \tau(k))$. Moreover, suppose that every node i maintains the variable $\tilde{z}_i(t) \in \mathbb{R}^d$ initialized at time $t = 0$ with $\tilde{z}_i(0) \in \mathbb{R}^d$ and updated by*

$$\tilde{z}_i(t+1) = \frac{\frac{\alpha(t+1)}{H(t)}\hat{z}_i(t+1) + S(t)\tilde{z}_i(t)}{S(t+1)},$$

where $S(0) = 0$ and $S(t) = \sum_{k=0}^{t-1} \frac{\alpha(k+1)}{H(k)}$ for $t \geqslant 1$. Then we have for each $T \geqslant 0$ and $i = 1, \cdots, m$, the following estimate

$$f(\tilde{z}_i(T+1)) - f(x^*) \leqslant \frac{me^{E_\tau}}{2\sqrt{T+1}}J_1(T) + \frac{3mDe^{E_\tau}}{\delta\sqrt{T+1}}J_2(T) + \frac{3mDe^{E_\tau}}{\delta\sqrt{T+1}}J_3(T),$$

where

$$\begin{aligned} J_1(T) &= \frac{\|\bar{x}(0) - x\|^2}{B_\zeta} + \left[2D^2\Big(1 + \ln(T+1)\Big) + 2E_{\tau,2}(T) + E_\tau(T)\Big)\right]B_\zeta \\ J_2(T) &= \left(\frac{C_0}{(1-\lambda)}\right)\|x(0)\|_1 + \frac{4mC_0E_\tau(T)}{(1-\lambda)} + \left(\frac{C_0 mD}{(1-\lambda)}\right)(1 + \ln(T)) \\ J_3(T) &= \sum_{t=0}^{T} K(t)\alpha(t+1)\Big[\|x(0)\|_1 + \sum_{s=0}^{t-1}(\alpha(s+1)D + \tau(s))\Big], \end{aligned}$$

and $x^* \in X^*$.

We mention that $J_3(t)$ is proved to be uniformly bounded for $t \geqslant 1$ in Lemma 5.5.6 under the assumption of the above theorem. Hence Theorem 5.5.2 implies that $f(\tilde{z}_i(t))$ converges to $f(x^*)$ at the rate of $O(\log(t)/\sqrt{t})$.

Lemma 5.5.3. *Suppose (A) and (B) of Assumption 5.2.1 hold. Then for any $t \geqslant 0$ and $x \in \mathbb{R}^d$ we have*

$$\begin{aligned}\sum_{i=1}^m \big(f_i(B_\zeta \bar{x}(t)) - f_i(x)\big) \leqslant\; & \frac{m}{2\alpha(t+1)B_\zeta}(\|B_\zeta \bar{x}(t) - x\|^2 - \|B_\zeta \bar{x}(t+1) - x\|^2) \\ & + \frac{B_\zeta m}{2\alpha(t+1)}\big(2\alpha(t+1)^2 D^2 + 2\tau(t)^2\big) \\ & + \frac{m}{\alpha(t+1)}\|B_\zeta \bar{x}(t) - x\|\tau(t) \\ & + 2D\sum_{i=1}^m \|\hat{z}_i(t+1) - B_\zeta \bar{x}(t)\|.\end{aligned}$$

Proof. By convexity, we have

$$\begin{aligned} f_i(\hat{z}_i(t+1)) &\leqslant f_i(x) + (\hat{z}_i(t+1) - x)\nabla f_i(\hat{z}_i(t+1)) \\ &= f_i(x) + (B_\zeta \bar{x}(t) - x)\nabla f_i(\hat{z}_i(t+1)) \\ &\quad + (\hat{z}_i(t+1) - B_\zeta \bar{x}(t))\nabla f_i(\hat{z}_i(t+1)) \\ &= f_i(x) + \frac{1}{\alpha(t+1)}(B_\zeta \bar{x}(t) - x)(w_i(t+1) - x_i(t+1) \\ &\quad + e_i(t+1)) + (\hat{z}_i(t+1) - B_\zeta \bar{x}(t))\nabla f_i(\hat{z}_i(t+1)),\end{aligned}$$

where (5.4.1) is used in the last equality. Summing up the above inequality from $i = 1$ to $i = m$, we find that

$$\begin{aligned}\sum_{i=1}^m f_i(\hat{z}_i(t+1)) - f_i(x) \leqslant\; & \underbrace{\frac{m}{\alpha(t+1)}(B_\zeta \bar{x}(t) - x)(\bar{x}(t) - \bar{x}(t+1))}_{I} \\ & + \underbrace{\frac{1}{\alpha(t+1)}(B_\zeta \bar{x}(t) - x)\sum_{i=1}^m e_i(t+1)}_{II} \\ & + \underbrace{\sum_{i=1}^m (\hat{z}_i(t+1) - B_\zeta \bar{x}(t))\nabla f_i(\hat{z}_i(t+1))}_{III}.\end{aligned}$$

Now we estimate each term in the right hand side. First using the equality $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$ for $a, b \in \mathbb{R}^d$, we have

$$I = \frac{m}{2\alpha(t+1)B_\zeta}(\|B_\zeta \bar{x}(t) - x\|^2 - \|B_\zeta \bar{x}(t+1) - x\|^2 + \|B_\zeta \bar{x}(t+1) - B_\zeta \bar{x}(t)\|^2).$$

Using (5.4.4) along with (5.4.5) and (A) of Assumption 5.2.1, we estimate the right-most term as

$$\begin{aligned}\|\bar{x}(t+1)-\bar{x}(t)\|^2 &\leqslant 2\left\|\frac{\alpha(t+1)}{m}\sum_{i=1}^{m}\nabla f_i(\hat{z}_i(t+1))\right\|^2 + 2\left\|\frac{1}{m}\sum_{i=1}^{m}e_i(t+1)\right\|^2 \\ &\leqslant 2\alpha(t+1)^2D^2 + 2\tau(t)^2.\end{aligned}$$

We apply (5.4.5) again to estimate

$$II \leqslant \frac{m}{\alpha(t+1)}\|B_\zeta\bar{x}(t)-x\|\tau(t),$$

and use (A) of Assumption 5.2.1, to deduce

$$III \leqslant D\sum_{i=1}^{m}\|\hat{z}_i(t+1)-B_\zeta\bar{x}(t)\|.$$

Combining the above estimates on I,II and III, we have

$$\begin{aligned}\sum_{i=1}^{m}\big(f_i(\hat{z}_i(t+1))-f_i(x)\big) \leqslant{}& \frac{m}{2\alpha(t+1)B_\zeta}(\|B_\zeta\bar{x}(t)-x\|^2-\|B_\zeta\bar{x}(t+1)-x\|^2) \\ &+\frac{mB_\zeta}{2\alpha(t+1)}\big(2\alpha(t+1)^2D^2+2\tau(t)^2\big) \\ &+\frac{m}{\alpha(t+1)}\|B_\zeta\bar{x}(t)-x\|\tau(t) \\ &+D\sum_{i=1}^{m}\|\hat{z}_i(t+1)-B_\zeta\bar{x}(t)\|.\end{aligned}$$

Finally we observe that gradient bounded assumption gives us the estimate

$$\sum_{i=1}^{m}\Big(f_i(B_\zeta\bar{x}(t))-f_i(\hat{z}_i(t+1))\Big) \leqslant D\sum_{i=1}^{m}\|B_\zeta\bar{x}(t)-\hat{z}_i(t+1)\|.$$

Summing up the above two inequalities, we obtain the desired estimate. □

Proof of Theorem 5.5.1

We recall the following lemma for proving Theorem 5.5.1.

Lemma 5.5.4 ([31])**.** *Consider a minimization problem*

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. Assume that the solution X^* of the problem is nonempty. Let $\{x(t)\}_{t\in\mathbb{N}}$ be a sequence such that for all $x \in X^*$ and for all $t \geqslant 0$,

$$\|x(t+1) - x\|^2 \leqslant (1 + b(t))\|x(t) - x\|^2 - a(t)(f(x(t)) - f(x)) + c(t)$$

where $b(t) \geqslant 0$, $a(t) \geqslant 0$ and $c(t) \geqslant 0$ for all $t \geqslant 0$ with $\sum_{t=0}^{\infty} b(t) < \infty$,$\sum_{t=0}^{\infty} a(t) = \infty$ and $\sum_{t=0}^{\infty} c(t) < \infty$. Then the sequence $\{x(t)\}_{t\in\mathbb{N}}$ converges to some solution $x^* \in X^*$

By manipulating the estimate in Lemma 5.5.3, we obtain the following estimate which is suitable for applying Lemma 5.5.4.

Corollary 5.5.5. *Suppose (A) and (B) of Assumption 5.2.1 hold. Then we have*

$$\begin{aligned}\|B_\zeta\bar{x}(t+1) - x\|^2 &\leqslant (1 + \tau(t))\|B_\zeta\bar{x}(t) - x\|^2 \\ &\qquad - \frac{2\alpha(t+1)B_\zeta}{m}(f(B_\zeta\bar{x}(t)) - f(x)) + c(t) + d(t),\end{aligned}$$

where

$$c(t) = \Big[2\alpha(t+1)^2 D^2 + 2\tau(t)^2 + \tau(t)\Big]B_\zeta,$$

and

$$d(t) = \frac{4\alpha(t+1)B_\zeta D}{m}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - \bar{x}(t)\|.$$

Proof. We use Young's inequality to find

$$\|B_\zeta\bar{x}(t) - x\|\tau(t) \leqslant \frac{\|B_\zeta\bar{x}(t) - x\|^2\,\tau(t)}{2B_\zeta} + \frac{\tau(t)B_\zeta}{2}.$$

Applying Lemma 5.5.3, we get

$$\begin{aligned}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big) &\leqslant \frac{m}{2\alpha(t+1)B_\zeta}(1 + \tau(t))\|B_\zeta\bar{x}(t) - x\|^2 \\ &\quad - \frac{m}{2\alpha(t+1)B_\zeta}\|B_\zeta\bar{x}(t+1) - x\|^2 \\ &\quad + \frac{mB_\zeta}{2\alpha(t+1)}\big(2\alpha(t+1)^2D^2 + 2\tau(t)^2\big) + \frac{B_\zeta m\tau(t)}{2\alpha(t+1)} \\ &\quad + 2D\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|.\end{aligned}$$

Dividing both sides by $\frac{m}{2\alpha(t+1)B_\zeta}$, it follows that

$$\begin{aligned}&\frac{2\alpha(t+1)B_\zeta}{m}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big)\\ &\leqslant (1+\tau(t))\|B_\zeta\bar{x}(t) - x\|^2 - \|B_\zeta\bar{x}(t+1) - x\|^2\\ &\quad + B_\zeta^2\Big[2\alpha(t+1)^2D^2 + 2\tau(t)^2 + \tau(t)\Big]\\ &\quad + \frac{4\alpha(t+1)DB_\zeta}{m}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|.\end{aligned}$$

Rearranging this we obtain the desired estimate. □

Now we are ready to prove Theorem 5.5.1.

Proof of Theorem 5.5.1. By Lemma 5.5.4 and Corollary 5.5.5 it is enough to prove $\sum_{t=0}^{\infty}(c(t) + d(t)) < \infty$, where

$$c(t) = 2\alpha(t+1)^2D^2 + 2\tau(t)^2 + \tau(t),$$

and

$$d(t) = \frac{4\alpha(t+1)DB_\zeta}{m}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|.$$

It follows that $\sum_{t=0}^{\infty} c(t) < \infty$ by (C) and (D) of Assumption 5.2.1.

Next we will show that $\sum_{t=0}^{\infty} d(t) < \infty$. By Proposition 5.4.1, it suffices to show that

$$\sum_{t=0}^{\infty}\alpha(t+1)\bigg(\lambda^t + \tau(t) + \sum_{s=0}^{t-1}\lambda^{t-s-1}\alpha(s+1) + \sum_{s=0}^{t-1}\lambda^{t-s-1}\tau(s)\bigg) < \infty, \qquad (5.5.1)$$

and

$$\sum_{t=0}^{\infty}\alpha(t+1)\Big[K(t) + \sum_{s=0}^{t-1}K(t)\Big(\alpha(s+1)D + \tau(s)\Big)\Big] < \infty.$$

The latter one is proved in Lemma 5.5.6 below. We proceed to prove (5.5.1). Using the Cauchy-Schwarz inequality, we have

$$\sum_{t=0}^{\infty}\alpha(t+1)\lambda^t \leqslant \frac{1}{2}\sum_{t=0}^{\infty}\alpha(t+1)^2 + \frac{1}{2}\sum_{t=0}^{\infty}\lambda^{2t} < \infty.$$

By rearranging and using the diminishing property of $\alpha(t)$ in (C) of Assumption 5.2.1, we find

$$\begin{aligned}\sum_{t=0}^{\infty}\alpha(t+1)\sum_{s=0}^{t-1}\lambda^{t-s-1}\alpha(s+1) &= \sum_{s=0}^{\infty}\sum_{t=s+1}^{\infty}\lambda^{t-s-1}\alpha(t+1)\alpha(s+1)\\ &\leqslant \sum_{s=0}^{\infty}\Big(\sum_{t=s+1}^{\infty}\lambda^{t-s-1}\Big)\alpha(s+1)^2\\ &= \frac{1}{1-\lambda}\sum_{s=0}^{\infty}\alpha(s+1)^2 < \infty.\end{aligned}$$

Similarly, due to (D) of Assumption 5.2.1, the last term is bounded as

$$\begin{aligned}\sum_{t=0}^{\infty}\alpha(t+1)\sum_{s=0}^{t-1}\lambda^{t-s-1}\tau(s) &= \sum_{s=0}^{\infty}\sum_{s<t}\lambda^{t-s-1}\alpha(t+1)\tau(s)\\ &\leqslant \sum_{s=0}^{\infty}\sum_{s<t}\lambda^{t-s-1}\alpha(s+1)\tau(s)\\ &= \frac{1}{1-\lambda}\sum_{s=0}^{\infty}\alpha(s+1)\tau(s) < \infty.\end{aligned}$$

Gathering the above estimates, we find that $\sum_{t=0}^{\infty}(c(t)+d(t)) < \infty$. Hence by Lemma 5.5.4, the sequence $\{B_{\zeta}\bar{x}(t)\}$ converges to some solution $x^* \in X^*$. Finally, we apply Corollary 5.5.5 to conclude that each sequence $\{\hat{z}_i(t)\}$, $i = 1,\cdots,n$, converges to the same solution x^*. The proof is done. □

Lemma 5.5.6. *Suppose that (D) and (E) of Assumption 5.2.1 hold. Then for the step size $\{\alpha(t)\}_{t\geqslant 0}$ satisfying (C) of Assumption 5.2.1 or $\alpha(t) = 1/\sqrt{t}$, we have*

$$\sum_{t=0}^{\infty}\alpha(t+1)\left[K(t)+\sum_{s=0}^{t-1}K(t)\left(\alpha(s+1)+\tau(s)\right)\right] < \infty.$$

Proof. From Lemma 5.3.1, we know that $\lim_{t\to\infty} t^{3/2}K(t) = 0$. Using this fact the summability of $\tau(s)$, it easily follows that

$$\sum_{t=0}^{\infty}\alpha(t+1)\Big[K(t)+\Big(\sum_{s=0}^{t-1}\tau(s)\Big)K(t)\Big] < \infty.$$

Next, for $\alpha(t)$ satisfying (C) of Assumption 5.2.1, we observe that

$$\sum_{t=0}^{\infty}\alpha(t+1)K(t)\sum_{s=0}^{t-1}\alpha(s+1) \leqslant \sum_{t=0}^{\infty}K(t)\sum_{s=0}^{t-1}\alpha(s+1)^2 < \infty.$$

For $\alpha(t) = 1/\sqrt{t}$, by using that $\sum_{s=0}^{t-1} 1/\sqrt{s+1} \leqslant 2\sqrt{t}$, we deduce

$$\sum_{t=0}^{\infty} \alpha(t+1)K(t) \sum_{s=0}^{t-1} \alpha(s+1) \leqslant 2 \sum_{t=0}^{\infty} K(t) < \infty.$$

The proof is done. □

Proof of Theorem 5.5.2

We now turn to the proof of Theorem 5.5.2. we recall that

$$H(-1) = 1 \quad \text{and} \quad H(t) = \prod_{k=0}^{t} (1 + \tau(k)) \quad \text{for} \quad t \geqslant 0$$

and

$$S(0) = 0 \quad \text{and} \quad S(t) = \sum_{s=0}^{t-1} \frac{\alpha(s+1)}{H(s)} \quad \text{for} \quad t \geqslant 1.$$

First, we find the boundedness of $H(t)$ and $S(t)$.

Lemma 5.5.7. *Let $\alpha(t) = \frac{1}{\sqrt{t}}$ and (D) and (E) of Assumption 5.2.1 hold. Then we have*

$$\sup_{t \geqslant 0} H(t) < e^{E_\tau} \quad \textit{and} \quad S(t) \geqslant e^{-E_\tau}\sqrt{t}, \tag{5.5.2}$$

where $E_\tau = \sum_{t=0}^{\infty} \tau(t) < \infty$.

Proof. By applying the inequality $1 + x \leqslant e^x$ for $x \geqslant 0$ we estimate $H(t)$ as

$$\sup_{t \geqslant 0} H(t) < \exp\left(\sum_{t=0}^{\infty} \tau(t)\right) = e^{E_\tau}.$$

Using this inequality, we deduce the following estimate

$$S(t) = \sum_{s=0}^{t-1} \frac{\alpha(s+1)}{H(s)} \geqslant e^{-E_\tau} \sum_{s=1}^{t} \frac{1}{\sqrt{s}} \geqslant e^{-E_\tau}\sqrt{t} \quad \forall\ t \geqslant 1.$$

The proof is done. □

To prove Theorem 5.5.2, we first modify Lemma 5.5.3 which states the boundedness of $\sum_{i=1}^{m} \big(f_i(\bar{x}(t)) - f_i(x)\big)$, by replacing $\bar{x}(t)$ to $\Big(\sum_{t=0}^{T} \frac{\alpha(t+1)}{H(t)} B_\zeta \bar{x}(t)\Big)/S(T+1)$.

Lemma 5.5.8. *Suppose that all the conditions are same as in Theorem 5.5.2. Then we have*

$$\begin{aligned} & f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}B_\zeta\bar{x}(t)}{S(T+1)}\right) - f(x) \\ & \leqslant \frac{me^{E_\tau}}{2\sqrt{T+1}}J_1(T) + \frac{2mDe^{E_\tau}}{\delta\sqrt{T+1}}J_2(T) + \frac{2mDe^{E_\tau}}{\delta\sqrt{T+1}}J_3(T) \end{aligned}$$

for any $T \in \mathbb{N}$, where $J_1(T), J_2(T)$, and $J_3(T)$ are defined in Theorem 5.5.2.

Proof. We recall from Lemma 5.5.3 the following inequality

$$\begin{aligned} \sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big) \leqslant & \frac{m}{2\alpha(t+1)B_\zeta}(1+\tau(t))\|B_\zeta\bar{x}(t) - x\|^2 \\ & - \frac{m}{2\alpha(t+1)B_\zeta}\|B_\zeta\bar{x}(t+1) - x\|^2 \\ & + \frac{mB_\zeta}{2\alpha(t+1)}\big(2\alpha(t+1)^2D^2 + 2\tau(t)^2\big) \\ & + \frac{B_\zeta m\tau(t)}{2\alpha(t+1)} + 2D\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|. \end{aligned} \tag{5.5.3}$$

Dividing both sides of (5.5.3) by $\frac{mH(t)}{2\alpha(t+1)}$, we get

$$\begin{aligned} \frac{2\alpha(t+1)}{mH(t)}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big) \leqslant & \frac{\|B_\zeta\bar{x}(t) - x\|^2}{B_\zeta H(t-1)} - \frac{\|B_\zeta\bar{x}(t+1) - x\|^2}{B_\zeta H(t)} \\ & + \frac{B_\zeta}{H(t)}\Bigg(2\alpha(t+1)^2D^2 + 2\tau(t)^2 + \tau(t)\Bigg) \\ & + \frac{4\alpha(t+1)D}{mH(t)}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|. \end{aligned}$$

Summing this from $t = 0$ to $t = T$ we obtain

$$\begin{aligned} & \sum_{t=0}^{T}\left[\frac{2\alpha(t+1)}{mH(t)}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big)\right] \\ & \leqslant \frac{\|\bar{x}(0) - x\|^2}{B_\zeta H(-1)} - \frac{\|\bar{x}(T+1) - x\|^2}{B_\zeta H(T)} \\ & \quad + \sum_{t=0}^{T}\frac{B_\zeta}{H(t)}\Bigg(2\alpha(t+1)^2D^2 + 2\tau(t)^2 + \tau(t)\Bigg) \\ & \quad + \sum_{t=0}^{T}\frac{4\alpha(t+1)D}{mH(t)}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|. \end{aligned}$$

This, together with the fact that $H(-1) = 1$ and $H(t) \geqslant 1$, gives

$$\begin{aligned}
&\sum_{t=0}^{T}\left[\frac{2\alpha(t+1)}{mH(t)}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big)\right] \\
&\leqslant \frac{\|\bar{x}(0)-x\|^2}{B_\zeta} + \sum_{t=0}^{T}\Big(2\alpha(t+1)^2D^2 + 2\tau(t)^2 + \tau(t)\Big)B_\zeta \qquad (5.5.4)\\
&\quad + \sum_{t=0}^{T}\left(\frac{4\alpha(t+1)D}{m}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|\right).
\end{aligned}$$

We find

$$\sum_{t=0}^{T} 2\alpha(t+1)^2D^2 = 2D^2\sum_{t=1}^{T+1}\frac{1}{t} \leqslant 2D^2\Big(1 + \ln(T+1)\Big), \tag{5.5.5}$$

and we have

$$\sum_{t=0}^{T}\Big(2\tau(t)^2 + \tau(t)\Big) = 2E_{\tau,2}(T) + E_\tau(T). \tag{5.5.6}$$

Finally, we estimate the last term of the right-hand side of (5.5.4) using Corollary 5.4.5 as follows:

$$\begin{aligned}
&\sum_{t=0}^{T}\left(\frac{4\alpha(t+1)D}{m}\sum_{i=1}^{m}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|\right) \\
&\leqslant 4D\Big(\frac{C_0}{\delta(1-\lambda)}\|x(0)\|_1 + \frac{4mC_0E_\tau(T)}{\delta(1-\lambda)} + \frac{C_0mD}{\delta(1-\lambda)}(1+\ln(T))\Big) \\
&\quad + \frac{4D}{\delta}\sum_{t=0}^{T}K(t)\alpha(t+1)\Big[\|x(0)\|_1 + \sum_{s=0}^{t-1}(\alpha(s+1)D + \tau(s))\Big].
\end{aligned}$$

Putting this estimate, (5.5.5) and (5.5.6) in (5.5.4), we achieve the following estimate

$$\begin{aligned}
&\sum_{t=0}^{T}\left[\frac{2\alpha(t+1)}{mH(t)}\sum_{i=1}^{m}\big(f_i(B_\zeta\bar{x}(t)) - f_i(x)\big)\right] \\
&\leqslant J_1(T) + \frac{4D}{\delta}J_2(T) + \frac{4D}{\delta}J_3(T).
\end{aligned}$$

Now we set $S(T) = \sum_{t=0}^{T-1}\frac{\alpha(t+1)}{H(t)}$ for $T \in \mathbb{N}$ and divide the both sides by $\frac{2S(T+1)}{m}$. Then we apply the convexity of f_i in the left hand side and use the lower bound $S(T+1) \geqslant e^{-E_\tau}\sqrt{T+1}$ to the right hand side, which leads to

$$\begin{aligned}
&f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}B_\zeta\bar{x}(t)}{S(T+1)}\right) - f(x) \\
&\leqslant \frac{me^{E_\tau}}{2\sqrt{T+1}}J_1(T) + \frac{2mDe^{E_\tau}}{\delta\sqrt{T+1}}J_2(T) + \frac{2mDe^{E_\tau}}{\delta\sqrt{T+1}}J_3(T).
\end{aligned}$$

The proof is finished. □

Now we are ready to give the proof of Theorem 5.5.2.

Proof of Theorem 5.5.2. Using the definition of $\tilde{z}_i$ and the gradient bounded assumption, we find

$$\begin{aligned}
&f(\tilde{z}_i(T+1)) - f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}B_\zeta\bar{x}(t)}{S(T+1)}\right)\\
&= f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}\hat{z}_i(t+1)}{S(T+1)}\right) - f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}B_\zeta\bar{x}(t)}{S(T+1)}\right)\\
&\leqslant \frac{mD}{S(T+1)}\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}\|\hat{z}_i(t+1) - B_\zeta\bar{x}(t)\|.
\end{aligned}$$

Then by Corollary 5.4.5 with the fact that $H(t) \geqslant 1$ and (5.5.2), we have

$$f(\tilde{z}_i(T+1)) - f\left(\frac{\sum_{t=0}^{T}\frac{\alpha(t+1)}{H(t)}B_\zeta\bar{x}(t)}{S(T+1)}\right) \leqslant \frac{mDe^{E_\tau}}{\delta\sqrt{T+1}}J_2(T) + \frac{mDe^{E_\tau}}{\delta\sqrt{T+1}}J_3(T).$$

Combining this inequality with Lemma 5.5.8, we obtain

$$\begin{aligned}
&f(\tilde{z}_i(T+1)) - f(x^*)\\
&\leqslant \frac{me^{E_\tau}}{2\sqrt{T+1}}J_1(T) + \frac{3mDe^{E_\tau}}{\delta\sqrt{T+1}}J_2(T) + \frac{3mDe^{E_\tau}}{\delta\sqrt{T+1}}J_3(T).
\end{aligned}$$

which is the desired estimate. Moreover, we see that the right hand side is bounded by $O(\log(T+1)/\sqrt{T+1})$ using Lemma 5.5.6. □

5.6 Numerical Experiment

In this section, we present simulation results of the proposed event-triggered gradient-push method to demonstrate that the theoretical results can be realized in practice.

Least Squares Solution

We consider the decentralized least squares problem:

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{m} f_i(x) \quad \text{with} \quad f_i(x) = \|q_i - p_i^T x\|^2,$$

where, each agent i in $\mathcal{V} = \{1, \cdots, m\}$ is given the local cost function f_i. The variable $p_i \in \mathbb{R}^{d \times p}$ is the input data and the variable $q_i \in \mathbb{R}^p$ is the output data. This type of problem is common in various fields, including machine learning and signal processing. Data is generated using a linear regression model: $q_i = p_i^T \tilde{x} + \varepsilon_i$ where $\tilde{x} \in \mathbb{R}^d$ is the true weight vector and $\varepsilon_i \in \mathbb{R}^p$ is the noise. The values of $\tilde{x}$ and p_i are randomly chosen from the uniform distribution in the range [0, 1] for each component. Additionally, the components of the noise vector $\varepsilon_i \in \mathbb{R}^p$ are jointly Gaussian with zero mean and variance 1. The initial points $x_i(0)$ are independent random variables, generated by a standard Gaussian distribution. In this simulation, we set the problem dimensions and the number of agents as $d = 5$, $p = 1$, and $m = 50$. We use a connected directed graph where every node has four out-neighbors.

Test 1. Here we fix $\alpha(t) = 1/t^{0.52}$ which satisfies (C) of Assumption 5.2.1 and $\zeta(t) = 1/(3t^3)$ which satisfies (E) of Assumption 5.2.1, and consider various choices of $\tau(t)$. We measure the relative distance between the variable $z_i(t)$ and the optimal point x^* as follows:

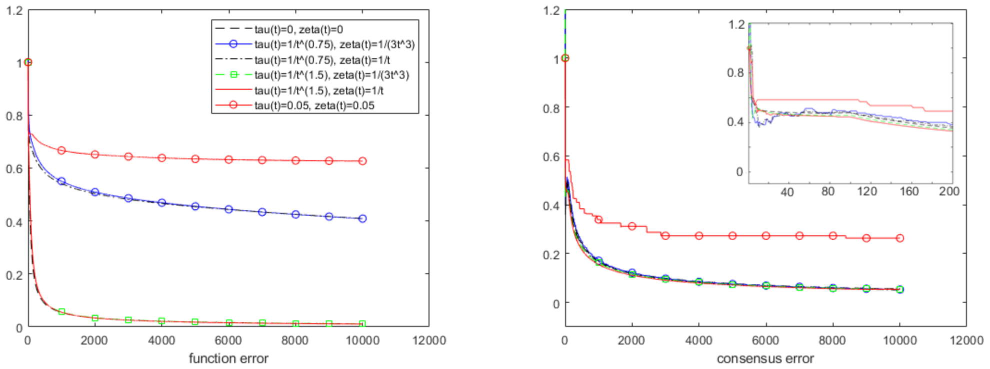
$$R_d(t) = \frac{\sum_{i=1}^m \|z_i(t) - x^*\|}{\sum_{i=1}^m \|z_i(0) - x^*\|}. \tag{5.6.1}$$

We set the termination time k_f as the first time $k \in \mathbb{N}$ when $R_d(k) < 10^{-2}$. And we let N_x and N_y be the average of total number of triggers for all agents until the termination time associated with $\tau(t)$ and $\zeta(t)$, respectively. Table 5.1 indicates the average of those values depending on $\tau(t)$ and $\zeta(t)$ in 100 trials.

We first look at the effect of $\zeta(t)$, the threshold for variables $y_i(t)$. Table 5.1 shows that the existence of the threshold ($\zeta \neq 0$) does not bring a big difference in the termination time if we compare the cases $\zeta(t) = 0$ and $\zeta(t) = 1/(3t^3)$ with same $\tau(t)$, but there is a big improvement in the number of triggers for $y_i(t)$. Next, we discuss the values N_x and k_f of Table 5.1 in terms of $\tau(t)$, the threshold for variables $x_i(t)$. As in Table 5.1, some cases give us similar or worse results compared to the cases $\tau(t) = 0$. For $\tau(t) = 1/t^{1.1}$, the number of triggers is decreased by more than 70%, and the termination time increased by more than 530%. For $\tau(t) = 1/t^{1.7}$, both the termination time and the number of triggers are increased by almost 36% when $\zeta(t) = 0$ and remain similar when $\zeta(t) = 1/(3t^3)$ compared to the cases $\tau(t) = 0$. For $\tau(t) = 1/t^{1.5}$, the termination time is almost the same, and the number of triggers is decreased by more than 20%. These results show that the proposed gradient-push with event-triggered communication with proper $\tau(t)$ and $\zeta(t)$ can diminish the number of communications to achieve convergence compared to the gradient-push algorithm without triggering. The threshold functions $\tau(t)$

$\tau(t)$	0	0	$1/t^{1.1}$	$1/t^{1.1}$	$1/t^{1.3}$	$1/t^{1.3}$	$1/t^{1.5}$	$1/t^{1.5}$	$1/t^{1.7}$	$1/t^{1.7}$
$\zeta(t)$	0	$1/(3t^3)$	0	$1/(3t^3)$	0	$1/(3t^3)$	0	$1/(3t^3)$	0	$1/(3t^3)$
N_x	11425	11305	3125	3152	3299	3288	8860	8767	15537	11177
N_y	11425	26	72705	32	15696	27	11644	26	15572	26
κ_f	11425	11305	72705	72757	15696	15572	11644	11514	15572	11207

Table 5.1: The number of triggers and termination time depending on $\tau(t)$ and $\zeta(t)$. Here the case $\tau(t) = 0$ and $\zeta(t) = 0$ corresponds to the GP algorithm [31] not involving the event-triggered communication.



(a)

(b)

Figure 5.1: (a) The values of $R_f(t)$ for Test 2 with different choices of $\tau(t)$ and $\zeta(t)$. (b) The values of $R_c(t)$ for Test 2 with different choices of $\tau(t)$ and $\zeta(t)$. The choice of $\tau(t)$ and $\zeta(t)$ for each graph of (b) corresponds to that of (a) with the same color and style.

and $\zeta(t)$ should be chosen carefully considering the characteristics of the given optimization problem.

Test 2. Here we fix $\alpha(t) = 1/\sqrt{t}$ and take several choices of $\tau(t)$ and $\zeta(t)$. We measure the relative cost error and the consensus error given by

$$R_f(t) = \frac{\sum_{i=1}^m (f(\tilde{z}_i(t)) - f^*)}{\sum_{i=1}^m (f(\tilde{z}_i(0)) - f^*)},$$

$$R_c(t) = \frac{\max_{i,j\in V} \|z_i(t) - z_j(t)\|}{\max_{i,j\in V} \|z_i(0) - z_j(t)\|}.$$

We consider two cases for $\tau(t)$: one where $\tau(t) = 1/t^{1.5}$ satisfies (D) of Assumption 5.2.1, and another where $\tau(t) = 1/t^{0.75}$ does not satisfy (D) of Assumption 5.2.1. And for $\zeta(t)$, we also consider two cases where $\zeta(t) = 1/(3t^3)$ satisfying (E) of Assumption 5.2.1 and $\zeta(t) = 1/t$ not satisfying (E) of Assumption 5.2.1. Additionally, we test two constant cases $\tau(t), \zeta(t) = 0$ and $\tau(t), \zeta(t) = 0.05$. Figure 5.1a depicts the graph of the values of $R_f(t)$ versus the iteration time. The result shows that the cost error decreases to zero when $\sum_{t=0}^{\infty} \tau(t) < \infty$ while it does not converge to zero when $\sum_{t=0}^{\infty} \tau(t) = \infty$ regardless of the choice of $\zeta(t)$. This supports the convergence result of Theorem 5.5.2. Figure 5.1b illustrates the consensus error $R_c(t)$. The result shows that the consensus error decreases to zero for any choices $\tau(t)$ and $\zeta(t)$ except the case $\tau(t), \zeta(t) = 0.05$. This numerical result supports the theoretical result obtained in Corollary 5.4.4.

Network localization

We consider the network localization problem where N free agents that only have estimates of their own positions and M anchor agents that have the information of their own exact positions in a global coordinate system. The goal of this problem is that each free agent achieves its own position in the global coordinate system only by communicating with its nearby neighbors and using the anchor agent's information. The communication pattern among agents is depicted by a directed (N+M)-node graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V}$ represents each agent, and each edge $\{i.j\} \in \mathcal{E}$ means agent i can send its information to j depending on its own sensor power. To formulate this problem, we let $x_i = (x_{i1}, x_{i2}) \in \mathbb{R}^2$ be the position of the agent i in $\mathbb{G}$. Without loss of generality, the point x_i is free agent $i \in \{1, \cdots, N\}$ and the point x_j is anchor agent for $j \in \{N+1, \cdots, N+M\}$. For each $i \in \{1, \cdots, N\}$, agent i has the set of neighboring agents, denoted by N_i, and may find the barycentric coordinates p_{ij} with respect to the neighboring agents $j \in N_i$ using their relative coordinates (coordniates with center x_i).To determine p_{ij}, each agent may solve the following problem

$$\min_{\{p_{ij}\}_{j \in N_i}} \sum_{j \in N_i} p_{ij}^2 \tag{5.6.2}$$

subject to

$$\begin{aligned} \sum_{j \in N_i} p_{ij} x_j &= x_i \\ \sum_{j \in N_i} p_{ij} &= 1. \end{aligned} \tag{5.6.3}$$

We also define p_{ij} for $j \notin N_i$ by

$$p_{ii} = -1 \quad \text{and} \quad p_{ij} = 0 \quad \text{for } j \notin N_i \cup \{i\}. \tag{5.6.4}$$

Then we have the following relation between the free agents and the anchor agents.

$$\sum_{j=1}^{N} p_{ij} x_j = q_i, \text{ for all } i \in \{1, \cdots, N\}, \tag{5.6.5}$$

where

$$q_i = (q_{i1}, q_{i2}) = \sum_{j=N+1}^{N+M} a_{ij} x_j \in \mathbb{R}^2. \tag{5.6.6}$$

Let $\overline{P} = P \otimes I_2$ and $x = (x_{11}, x_{12}, x_{21}, \cdots, x_{N1}, x_{N2})^T$. Then, $x \in \mathbb{R}^{2d}$ is the solution of the following decentralized problem:

$$\min_{s \in \mathbb{R}^{2d}} \sum_{i=1}^{N} f_i(x) := \left(|q_{i1} - \overline{P}_{2i-1} s|^2 + |q_{i2} - \overline{P}_{2i}^T s|^2 \right),$$

where $\overline{P}_i$ is the ith row of the $\overline{P}$. We remark that agent i has information of $\overline{P}_{2i-1}$, $\overline{P}_{2i}$, q_{i1}, q_{i2}.

In our experiment, we set $N = 11$ and $M = 3$. We design the network of agents so that each free agent has a sensor power to send its information to at least 4 free agents (See Figure 5.2). We choose the step size $\alpha(t) = 1/t^{(0.7)}$, which satisfies (C) of Assumption 5.2.1 and $\alpha(t) = 1/\sqrt{t}$. For each stepsize, we consider various choices of $\tau(t)$ but fix $\zeta(t) = 0$. We measure $R_d(t)$ (see Figures 5.3a, 5.4a) and the number of triggers with respect to the choice of $\tau(t)$ (see Figures 5.3b, 5.4b). Comparing the cases $\tau(t) = 0$ and $\tau(t) = 1/t^{1.3}$, we find that the number of triggering times of the case $\tau(t) = 1/t^{1.3}$ is much smaller than that of the case $\tau(t) = 0$ while they have similar decay in function errors. From this result, we see that one may reduce the communication cost for resource-aware scenarios.

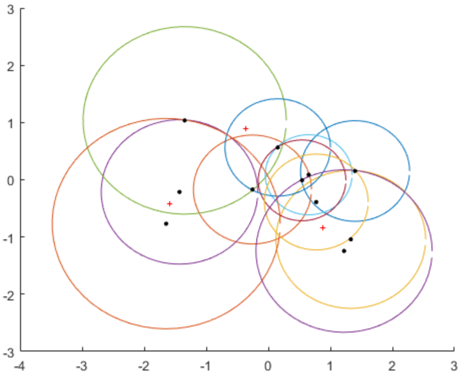
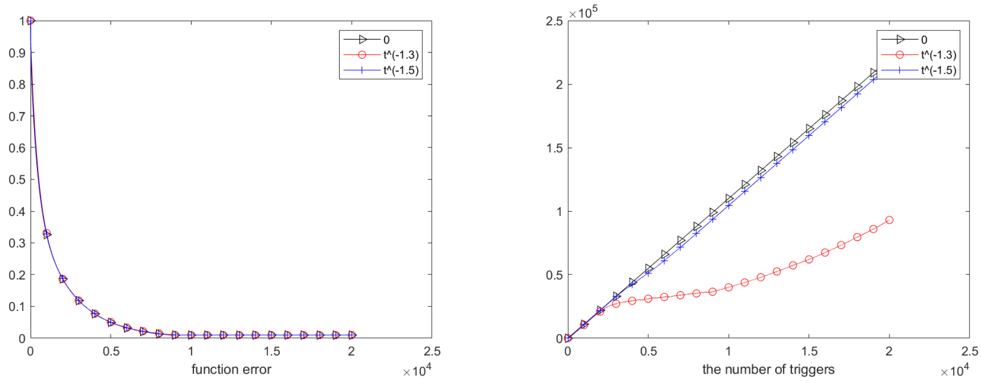


Figure 5.2: Each dot represents free agents and Each plus sigh represents known anchor agents. Each circle displays the sensor power of its corresponding free agent.

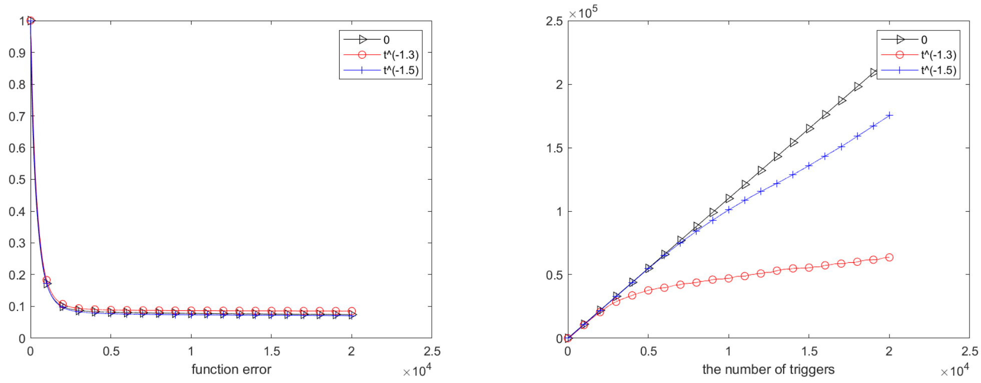
Lastly, we look at the effect of the number of anchor agents on the performance of the algorithm. Precisely, we compare the graphs of $R_d(t)$ (see (5.6.1)) and the number of triggers for various choices of M with fixing $N = 11$ (see Figures 5.5a, 5.5b). As in the previous test, we have set $\alpha(t) = 1/t^{0.7}$, $\tau(t) = 1/t^{1.5}$ and $\zeta(t) = 0$. We perform the test for $M \in \{3, 5, 7, 9, 11\}$. In figure 5.3b, we observe that for the



(a)

(b)

Figure 5.3: (a) The values of $R_d(t)$ with different choices of $\tau(t)$ for fixed $\alpha(t) = 1/t^{(0.7)}$. (b) The values of the number of triggers with different choices of $\tau(t)$ for fixed $\alpha(t) = 1/t^{(0.7)}$.

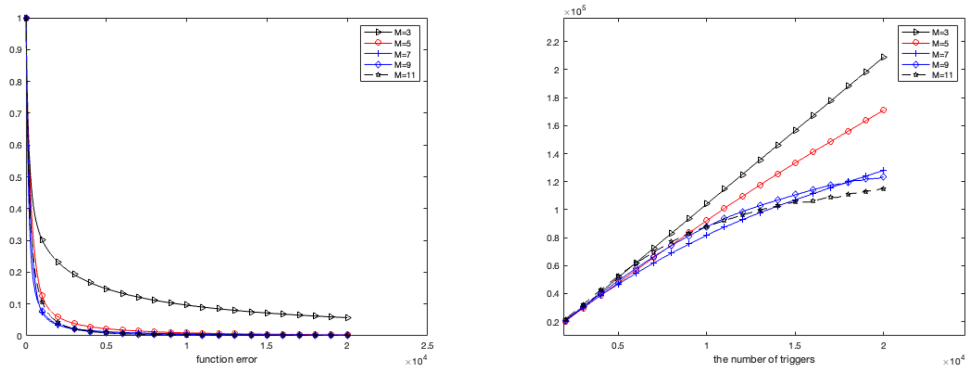


(a)

(b)

Figure 5.4: (a) The values of $R_d(t)$ with different choices of $\tau(t)$ for fixed $\alpha(t) = 1/\sqrt{t}$. (b) The values of the number of triggers with different choices of $\tau(t)$ for fixed $\alpha(t) = 1/\sqrt{t}$.

case $M = 3$ the number of triggering times is not significantly smaller than that of the time-triggered case $\tau(t) = 0$. However, if the number M is larger or equal to 7, then the number of triggers becomes notably smaller compared to the case $M = 3$. Also, the error value $R_d(t)$ decreases faster if the number of anchor agents is larger (see figure 5.5a). These results imply that if there are more information of anchors that agents can access, the event-triggering strategy of gradient-push algorithm



(a)

(b)

Figure 5.5: We fixed $\alpha(t) = 1/t^{0.7}, \tau(t) = 1/t^{1.5}$ and $\zeta(t) = 0$. (a) The values of $R_d(t)$ with different choices of the number of anchor agents M. (b) The values of the number of triggers with different choices of the number of anchor agents M.

becomes more effective by reducing the power consumption for communication.

Appendix A

Appendix

A.1 Proof of Proposition 2.4.1

We begin this section with obtaing the following lemma.

Lemma A.1.1. *1. Assume that a continuous function $f : [0, \infty) \to \mathbb{R}$ is non-increasing on $[0, \infty)$. Then for any integers $b \geqslant a \geqslant 0$, we have*

$$\sum_{s=a}^{b} f(s) \geqslant \int_{a}^{b+1} f(v)dv.$$

2. Assume that a continuous function $f : [-1, \infty) \to \mathbb{R}$ is diminishing on $[0, c)$ and increasing on $[c, \infty)$. Then for any integers $b \geqslant a \geqslant 0$, we have

$$\sum_{s=a}^{b} f(s) \leqslant \int_{a-1}^{b+1} f(v)dv.$$

Proof. (1)Since f is non-increasing, we have

$$\begin{aligned} \int_{a}^{b+1} f(s)ds &= \sum_{s=a}^{b} \int_{s}^{s+1} f(v)dv \\ &\leqslant \sum_{s=a}^{b} f(s). \end{aligned} \tag{A.1.1}$$

(2) We consider the case $c \in [a, b]$. Then

$$\begin{aligned}\sum_{s=a}^{b} f(s) &= \sum_{s=a}^{[c]} f(s) + \sum_{s=[c]+1}^{b} f(s) \\ &\leqslant \sum_{s=a}^{[c]} \int_{s-1}^{s} f(v)dv + \sum_{s=[c]+1}^{b} \int_{s}^{s+1} f(v)dv \\ &\leqslant \int_{a-1}^{b+1} f(v)dv,\end{aligned}$$

where the first inequality can be proved similarly to (1) of (A.1.1). The proof is finished. □

Now we prove Proposition 2.4.1.

Proof of Proposition 2.4.1. By (2.4.1), we have

$$A(t) \leqslant \left(1 - \frac{C_1}{(t+w-1)^p}\right) A(t-1) + \frac{QC_2}{(t+w)^{p+q}}, \quad \text{for all } t \geqslant 1, \tag{A.1.2}$$

where

$$Q = \sup_{t \geqslant 1} \frac{(t+w)^{p+q}}{(t+w-1)^{p+q}} = \left(\frac{w+1}{w}\right)^{p+q}.$$

Using (A.1.2) iteratively, we have

$$\begin{aligned}A(t) &\leqslant \prod_{s=0}^{t-1} \left(1 - \frac{C_1}{(s+w)^p}\right) A(0) + \sum_{s=1}^{t-1} \left[\frac{QC_2}{(s+w)^{p+q}} \prod_{k=s}^{t-1} \left(1 - \frac{C_1}{(k+w)^p}\right)\right] + \frac{QC_2}{(t+w)^{p+q}} \\ &\leqslant e^{-\sum_{s=0}^{t-1} \frac{C_1}{(s+w)^p}} A(0) + \sum_{s=1}^{t-1} e^{-\sum_{k=s}^{t-1} \frac{C_1}{(k+w)^p}} \frac{QC_2}{(s+w)^{p+q}} + \frac{QC_2}{(t+w)^{p+q}}.\end{aligned} \tag{A.1.3}$$

Here we used the fact $1 - x \leqslant e^{-x}$.

We first consider the case $p < 1$. Using Lemma A.1.1 we note that for any integers a and b with $b \geqslant a \geqslant 0$,

$$\sum_{s=a}^{b} \frac{C_1}{(s+w)^p} \geqslant \int_{a}^{b+1} \frac{C_1}{(s+w)^p} ds = \frac{C_1}{1-p}\Big((b+w+1)^{1-p} - (a+w)^{1-p}\Big)$$

since $s \to \frac{C_1}{(s+w)^p}$ is diminishing for $s \geqslant 0$. This gives

$$e^{-\sum_{k=s}^{t-1} \frac{C_1}{(k+w)^p}} \leqslant e^{-\frac{C_1((t+w)^{1-p} - (s+w)^{1-p})}{1-p}}.$$

Combining these estimates with (A.1.3), we have

$$\begin{aligned}A(t) &\leqslant e^{-\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p}}A(0)+\sum_{s=1}^{t}e^{-\frac{C_1((t+w)^{1-p}-(s+w)^{1-p})}{1-p}}\frac{QC_2}{(s+w)^{p+q}} \\ &\leqslant e^{-\sum_{s=0}^{t-1}\frac{C_1}{(s+w)^p}}A(0)+J_1+J_2,\end{aligned} \tag{A.1.4}$$

where

$$\begin{aligned}J_1 &:= \sum_{s=1}^{[t/2]-1}e^{-\frac{C_1((t+w)^{1-p}-(s+w)^{1-p})}{1-p}}\frac{QC_2}{(s+w)^{p+q}} \\ J_2 &:= \sum_{s=[t/2]}^{t}e^{-\frac{C_1((t+w)^{1-p}-(s+w)^{1-p})}{1-p}}\frac{QC_2}{(s+w)^{p+q}}.\end{aligned}$$

Here we regard that $J_1=0$ for $1\leqslant t\leqslant 3$. In fact, the second inequality in (A.1.4) is equality except $t=1$. We first estimate the second part as follow.

$$\begin{aligned}J_2 &= QC_2e^{-\frac{C_1}{1-p}(t+w)^{1-p}}\sum_{s=[t/2]}^{t}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\frac{1}{(s+w)^{p+q}} \\ &\leqslant QC_2e^{-\frac{C_1}{1-p}(t+w)^{1-p}}\int_{[t/2]-1}^{t+1}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\frac{1}{(s+w)^{p+q}}ds \\ &=: QC_2e^{-\frac{C_1}{1-p}(t+w)^{1-p}}\cdot I,\end{aligned}$$

where we used Lemma A.1.1 for the inequality. Notice that

$$\left(\frac{1}{C_1}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\right)'=e^{\frac{C_1}{1-p}(s+w)^{1-p}}\cdot\frac{1}{(s+w)^p}.$$

Using this and an integration by parts, we get

$$\begin{aligned}I &= \frac{1}{C_1}\left[e^{\frac{C_1}{1-p}(s+w)^{1-p}}\cdot\frac{1}{(s+w)^q}\right]_{[t/2]-1}^{t+1}+\frac{q}{C_1}\int_{[t/2]-1}^{t+1}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\frac{1}{(s+w)^{q+1}}ds \\ &= \frac{1}{C_1}\left[e^{\frac{C_1}{1-p}(t+1+w)^{1-p}}(t+1+w)^{-q}-e^{\frac{C_1}{1-p}([t/2]+w-1)^{1-p}}([t/2]+w-1)^{-q}\right] \\ &\quad+\frac{q}{C_1}\int_{[t/2]-1}^{t+1}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\frac{1}{(s+w)^{q+1}}ds.\end{aligned} \tag{A.1.5}$$

Note that

$$\frac{q}{C_1}\int_{[t/2]-1}^{t+1}e^{\frac{C_1}{1-p}(s+w)^{1-p}}\frac{1}{(s+w)^{q+1}}ds\leqslant\frac{q}{C_1}e^{\frac{C_1}{1-p}(t+1+w)^{1-p}}\int_{[t/2]-1}^{t+1}\frac{1}{(s+w)^{q+1}}ds$$

and
$$\int_{[t/2]-1}^{t+1} \frac{1}{(s+w)^{q+1}} ds = -\frac{(t+1+w)^{-q} - ([t/2]+w-1)^{-q}}{q}.$$
Using these estimates, the integration part in (A.1.5) is bounded by
$$-\frac{1}{C_1} e^{\frac{C_1}{1-p}(t+1+w)^{1-p}} \Big[(t+1+w)^{-q} - ([t/2]+w-1)^{-q}\Big].$$
Therefore we have
$$\begin{aligned}
J_2 &\leqslant \frac{QC_2}{C_1} e^{-\frac{C_1}{1-p}(t+w)^{1-p}} \\
&\quad \cdot \Bigg[\left(e^{\frac{C_1}{1-p}(t+1+w)^{1-p}} (t+1+w)^{-q} - e^{\frac{C_1}{1-p}([t/2]+w-1)^{1-p}} ([t/2]+w-1)^{-q} \right) \\
&\quad - e^{\frac{C_1}{1-p}(t+1+w)^{1-p}} \left((t+1+w)^{-q} - ([t/2]+w-1)^{-q} \right) \Bigg] \\
&\leqslant \frac{QC_2}{C_1} e^{-\frac{C_1}{1-p}\left[(t+w)^{1-p}-(t+1+w)^{1-p}\right]} ([t/2]+w-1)^{-q} \\
&\leqslant \frac{QC_2}{C_1} e^{\frac{C_1}{w^p}} ([t/2]+w-1)^{-q},
\end{aligned}$$
where we used that $(t+w+1)^{1-p} - (t+w)^{1-p} \leqslant \frac{1-p}{(t+w)^p} \leqslant \frac{1-p}{w^p}$ in the last inequality. Next, using that $s \leqslant [t/2]-1$ in the summation of J_1, we derive the following inequality:
$$\begin{aligned}
J_1 &\leqslant QC_2 e^{-\frac{C_1}{1-p}[(t+w)^{1-p} - ([t/2]-1+w)^{1-p}]} \sum_{s=1}^{[t/2]-1} \frac{1}{(s+w)^{p+q}} \\
&\leqslant QC_2 e^{-\frac{C_1 t}{2(t+w)^p}} \sum_{s=1}^{[t/2]-1} \frac{1}{(s+w)^{p+q}}.
\end{aligned}$$
Here we used $(t+w)^{1-p} - ([t/2]-1+w)^{1-p} \geqslant (1-p)\frac{t-[t/2]+1}{(t+w)^p} \geqslant \frac{(1-p)t}{2(t+w)^p}$. Combining the above two estimates,
$$A(t) \leqslant \frac{QC_2}{C_1} e^{\frac{C_1}{w^p}} ([t/2]+w-1)^{-q} + \mathcal{R}(t),$$
where
$$\mathcal{R}(t) = e^{-\sum_{s=0}^{t-1} \frac{C_1}{(s+w)^p}} A(0) + QC_2 e^{-\frac{C_1 t}{2(t+w)^p}} \sum_{s=1}^{[t/2]-1} \frac{1}{(w+s)^{p+q}}.$$

Next we consider the case $p=1$. We recall (A.1.3) as
$$A(t) \leqslant e^{-\sum_{s=0}^{t-1} \frac{C_1}{s+w}} A(0) + \sum_{s=1}^{t-1} e^{-\sum_{k=s}^{t-1} \frac{C_1}{k+w}} \frac{QC_2}{(s+w)^{1+q}} + \frac{QC_2}{(t+w)^{1+q}}.$$

Notice that for any integers a and b with $b \geqslant a \geqslant 0$, we use Lemma A.1.1 to have

$$\sum_{s=a}^{b} \frac{C_1}{s+w} \geqslant \int_a^{b+1} \frac{C_1}{s+w} ds = \log \left(\frac{b+w+1}{a+w} \right)^{C_1}.$$

Using this in (A.1.3) we get

$$\begin{aligned} A(t) &\leqslant \left(\frac{w}{t+w} \right)^{C_1} A(0) + \sum_{s=1}^{t} \left(\frac{s+w}{t+w} \right)^{C_1} \frac{QC_2}{(s+w)^{1+q}} \\ &\leqslant \left(\frac{w}{t+w} \right)^{C_1} A(0) + \left(\frac{1}{t+w} \right)^{C_1} \sum_{s=1}^{t} \frac{QC_2}{(s+w)^{1+q-C_1}}. \end{aligned} \tag{A.1.6}$$

Case 1. Suppose that $1+q-C_1 \neq 1$. Then we have

$$\begin{aligned} \sum_{s=1}^{t} \frac{1}{(s+w)^{1+q-C}} &\leqslant \int_0^{t+1} \frac{1}{(s+w)^{1+q-C_1}} ds \\ &= \frac{1}{q-C_1} \Big[w^{C_1-q} - (t+1+w)^{C_1-q} \Big], \end{aligned}$$

where we used Lemma A.1.1 for the first inequality. Hence $A(t)$ is bounded by

$$A(t) \leqslant \left(\frac{w}{t+w} \right)^{C_1} A(0) + \frac{w^{C_1-q}}{q-C_1} \frac{QC_2}{(t+w)^{C_1}} - \frac{1}{q-C_1} \frac{QC_2}{(t+w+1)^q} \frac{(t+w+1)^{C_1}}{(t+w)^{C_1}}.$$

Case 2. Suppose that $1+q-C_1 = 1$. Then we have,

$$\sum_{s=1}^{t} \frac{1}{s+w} \leqslant \int_0^t \frac{1}{s+w} ds = \log \left(\frac{t+w}{w} \right).$$

Hence $A(t)$ is bounded by

$$A(t) \leqslant \left(\frac{w}{t+w} \right)^{C_1} A(0) + \log \left(\frac{t+w}{w} \right) \frac{QC_2}{(t+w)^{C_1}}.$$

Combining the above estimates, we find

$$A(t) \leqslant \left(\frac{w}{t+w} \right)^{C_1} A(0) + \mathcal{R}'(t),$$

where

$$\mathcal{R}'(t) = \begin{cases} \frac{w^{C_1-q}}{q-C_1} \cdot \frac{QC_2}{(t+w)^{C_1}} & \text{if } q > C_1 \\ \log\left(\frac{t+w}{w}\right) \cdot \frac{QC_2}{(t+w)^{C_1}} & \text{if } q = C_1 \\ \frac{1}{C_1-q} \cdot \left(\frac{w+1}{w}\right)^{C_1} \cdot \frac{QC_2}{(t+w+1)^q} & \text{if } q < C_1. \end{cases}$$

The proof is done. □

A.2 Proof of Proposition 2.4.3

Letting $a(t) = 1 - \frac{a}{(t+w)^p}$ and $b(t) = \frac{b\beta^t}{(t+w)^p}$ for simplicity, we find that

$$\begin{aligned}
B(t+1) &\leqslant \Big(\prod_{k=0}^{t} a(k)\Big)B(0) + \sum_{j=0}^{t-1}\Big(\prod_{k=j+1}^{t} a(k)\Big)b(j) + b(t) \\
&= \sum_{j=0}^{t-1}\Big(\prod_{k=j+1}^{t} a(k)\Big)b(j) + b(t).
\end{aligned}$$

(Case $p = 1$). In this case, we have

$$B(t+1) \leqslant \sum_{j=0}^{t-1}\Big(\prod_{k=j+1}^{t}\Big(1 - \frac{a}{k+w}\Big)\Big)\frac{b\beta^j}{(j+w)} + \frac{b\beta^t}{(t+w)}.$$

Notice that

$$\begin{aligned}
\prod_{k=j+1}^{t}\Big(1 - \frac{a}{k+w}\Big) &\leqslant e^{-\sum_{k=j+1}^{t}\frac{a}{k+w}} \\
&\leqslant e^{-a\int_{j+w}^{t+w}\frac{1}{s}ds} \\
&= e^{-a\Big(\log(t+w) - \log(j+w)\Big)} \\
&= \Big(\frac{j+w}{t+w}\Big)^a.
\end{aligned}$$

Using this we get

$$\begin{aligned}
B(t+1) &\leqslant \sum_{j=0}^{t-1}\frac{(j+w)^{a-1}}{(t+w)^a}b\beta^j + \frac{b\beta^t}{(t+w)} \\
&= \frac{bw^{a-1}}{(t+w)^a}\sum_{j=0}^{t-1}\frac{(j+w)^{a-1}}{w^{a-1}}\beta^j + \frac{b\beta^t}{(t+w)} \\
&\leqslant J_{a,w,\beta}\cdot\frac{bw^{a-1}}{(t+w)^a} + \frac{b\beta^t}{(t+w)},
\end{aligned}$$

where $J_{a,w,\beta} = \sum_{j=0}^{\infty} \frac{(j+w)^{a-1}}{w^{a-1}}\beta^j$. For any $0 \leqslant l \leqslant a-1$, we may also bound it as

$$\begin{aligned}
B(t+1) &\leqslant b\sum_{j=0}^{t-1} \frac{(j+w)^l}{(t+w)^{l+1}}\beta^j + \frac{b\beta^t}{(t+w)} \\
&= \frac{bw^l}{(t+w)^{l+1}} \sum_{j=0}^{t-1} \Big(1 + \frac{j}{w}\Big)^l \beta^j + \frac{b\beta^t}{(t+w)} \\
&\leqslant \frac{bw^l}{(t+w)^{l+1}} \sum_{j=0}^{t-1} e^{jl/w}\beta^j + \frac{b\beta^t}{(t+w)} \\
&\leqslant \frac{bw^l}{(t+w)^{l+1}} \cdot \frac{1}{1 - e^{l/w}\beta} + \frac{b\beta^t}{(t+w)},
\end{aligned}$$

where we used that $1 + x \leqslant e^x$ for $x \geqslant 0$ in the second inequality.

(Case $p \in (0,1)$). We estimate further as

$$\begin{aligned}
B(t+1) &\leqslant \sum_{j=0}^{[t/2]-1} \Big(\prod_{k=j+1}^{t} a(k)\Big) b(j) + \sum_{j=[t/2]}^{t-1} \Big(\prod_{k=j+1}^{t} a(k)\Big) b(j) + b(t) \\
&\leqslant \Big(\prod_{k=[t/2]}^{t} a(k)\Big) \sum_{j=0}^{[t/2]-1} b(j) + \sum_{j=[t/2]}^{t} b(j).
\end{aligned}$$

Since $b(t) = \frac{b\beta^t}{(t+w)^p}$, we have $b(t) \leqslant \frac{b}{w^p}$. Thus,

$$\begin{aligned}
B(t+1) &\leqslant \Big(\prod_{k=[t/2]}^{t} a(k)\Big)\frac{b}{w^p} + \frac{b}{w^p}\sum_{j=[t/2]}^{t} \beta^j \\
&\leqslant \frac{b}{w^p}\Big(\prod_{k=[t/2]}^{t} a(k)\Big) + \frac{b}{w^p(1-\beta)}\beta^{[t/2]}.
\end{aligned}$$

We estimate

$$\begin{aligned}
\prod_{k=[t/2]}^{t} \Big(1 - \frac{a}{(k+w)^p}\Big) &\leqslant e^{-\sum_{k=[t/2]}^{t} \frac{a}{(k+w)^p}} \\
&\leqslant e^{-a\int_{[t/2]+w-1}^{t+w} \frac{1}{s^p} ds} \\
&= e^{-\frac{a}{(1-p)}\big((t+w)^{1-p} - ([t/2]+w-1)^{1-p}\big)}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
B(t+1) &\leqslant \frac{b}{w^p}\bigg(e^{-\frac{a}{(1-p)}((t+w)^{1-p} - ([t/2]+w-1)^{1-p})} + \frac{\beta^{[t/2]}}{(1-\beta)}\bigg) \\
&\leqslant \frac{b}{w^p}\bigg(e^{-\frac{a}{2}\cdot\frac{t}{(t+w)^p}} + \frac{\beta^{[t/2]}}{(1-\beta)}\bigg).
\end{aligned}$$

Here we used $(t+w)^{1-p} - ([t/2]+w-1)^{1-p} \geqslant (1-p)\frac{t-[t/2]+1}{(t+w)^p} \geqslant \frac{(1-p)t}{2(t+w)^p}$. The proof is done.

Bibliography

[1] Paolo Addesso, Stefano Marano, and Vincenzo Matta. Sequential sampling in sensor networks for detection with censoring nodes. *IEEE Transactions on Signal Processing*, 55(11):5497–5505, 2007.

[2] Swaroop Appadwedula, Venugopal V Veeravalli, and Douglas L Jones. Decentralized detection with censoring sensors. *IEEE Transactions on Signal Processing*, 56(4):1362–1373, 2008.

[3] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[5] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[6] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[7] Francesco Bullo, Jorge Cortés, and Sonia Martinez. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*, volume 27. Princeton University Press, 2009.

[8] Xuanyu Cao and Tamer Başar. Decentralized online convex optimization with event-triggered communications. *IEEE Transactions on Signal Processing*, 69:284–299, 2020.

[9] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics*, 9(1):427–438, 2012.

[10] Annie I-An Chen. *Fast distributed first-order methods.* PhD thesis, Massachusetts Institute of Technology, 2012.

[11] Woocheol Choi, Doheon Kim, and Seok-Bae Yun. On the convergence result of the gradient-push algorithm on directed graphs with constant stepsize. *arXiv preprint arXiv:2302.08779*, 2023.

[12] Woocheol Choi and Jimyeong Kim. On the convergence of decentralized gradient descent with diminishing stepsize, revisited. *arXiv preprint arXiv:2203.09079*, 2022.

[13] Woocheol Choi and Jimyeong Kim. On the convergence analysis of the decentralized projected gradient descent. *arXiv preprint arXiv:2303.08412*, 2023.

[14] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[15] Dimos V Dimarogonas, Emilio Frazzoli, and Karl H Johansson. Distributed event-triggered control for multi-agent systems. *IEEE Transactions on automatic control*, 57(5):1291–1297, 2011.

[16] Ziwei Dong, Shuai Mao, Wei Du, and Yang Tang. Distributed constrained optimization with linear convergence rate. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pages 937–942. IEEE, 2020.

[17] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[18] Alessandro Falsone, Kostas Margellos, Simone Garatti, and Maria Prandini. Dual decomposition for multi-agent distributed optimization with coupling constraints. *Automatica*, 84:149–158, 2017.

[19] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. Consensus-based distributed linear support vector machines. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 35–46, 2010.

[20] Xingkang He, Yu Xing, Junfeng Wu, and Karl H Johansson. Event-triggered distributed estimation with decaying communication rate. *SIAM Journal on Control and Optimization*, 60(2):992–1017, 2022.

[21] Yuichi Kajiyama, Naoki Hayashi, and Shigemasa Takai. Distributed subgradient method with edge-based event-triggered communication. *IEEE Transactions on Automatic Control*, 63(7):2248–2255, 2018.

[22] Jimyeong Kim and Woocheol Choi. Gradient-push algorithm for distributed optimization with event-triggered communications. *IEEE Access*, 2022.

[23] Ran Li and Xiaowu Mu. Distributed event-triggered subgradient method for convex optimization with general step-size. *IEEE Access*, 8:14253–14264, 2020.

[24] Qing Ling and Zhi Tian. Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing*, 58(7):3816–3827, 2010.

[25] Changxin Liu, Huiping Li, Yang Shi, and Demin Xu. Distributed event-triggered gradient method for constrained convex minimization. *IEEE Transactions on Automatic Control*, 65(2):778–785, 2019.

[26] Shuai Liu, Zhirong Qiu, and Lihua Xie. Convergence rate analysis of distributed optimization with projected subgradient algorithm. *Automatica*, 83:162–169, 2017.

[27] Qingguo Lü and Huaqing Li. Event-triggered discrete-time distributed consensus optimization over time-varying graphs. *Complexity*, 2017:1–12, 2017.

[28] Marie Maros and Joakim Jaldén. On the q-linear convergence of distributed generalized admm under non-strongly convex function components. *IEEE Transactions on Signal and Information Processing over Networks*, 5(3):442–453, 2019.

[29] Manuel Mazo and Paulo Tabuada. Decentralized event-triggered control over wireless sensor/actuator networks. *IEEE Transactions on Automatic Control*, 56(10):2456–2461, 2011.

[30] Martin Meinel, Michael Ulbrich, and Sebastian Albrecht. A class of distributed optimization methods with event-triggered communication. *Computational Optimization and Applications*, 57:517–553, 2014.

[31] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

[32] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

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

[34] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

[35] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[36] S Sundhar Ram, A Nedich, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *J Optim Theory Appl*, 147:516–545, 2010.

[37] R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

[38] Ali H Sayed. Diffusion adaptation over networks. In *Academic Press Library in Signal Processing*, volume 3, pages 323–453. Elsevier, 2014.

[39] Eugene Seneta. *Non-negative matrices and Markov chains.* Springer Science & Business Media, 2006.

[40] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[41] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[42] Victor Shnayder, Mark Hempstead, Bor-rong Chen, Geoff Werner Allen, and Matt Welsh. Simulating the power consumption of large-scale sensor network applications. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 188–200, 2004.

[43] Andrea Simonetto and Hadi Jamali-Rad. Primal recovery from consensus-based dual decomposition for distributed convex optimization. *Journal of Optimization Theory and Applications*, 168:172–197, 2016.

[44] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[45] Tianyu Wu, Kun Yuan, Qing Ling, Wotao Yin, and Ali H Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307, 2017.

[46] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

[47] Ran Xin, Shi Pu, Angelia Nedić, and Usman A Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, 2020.

[48] Menghui Xiong, Baoyong Zhang, Daniel WC Ho, Deming Yuan, and Shengyuan Xu. Event-triggered distributed stochastic mirror descent for convex optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[49] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[50] Minyi Zhong and Christos G Cassandras. Asynchronous distributed optimization with event-driven communication. *IEEE Transactions on Automatic Control*, 55(12):2735–2750, 2010.