DISTRIBUTED RANDOMIZED BLOCK STOCHASTIC GRADIENT TRACKING METHODS: RATE ANALYSIS AND NUMERICAL EXPERIMENTS

By

JAYESH VINAYAK YEVALE

Bachelor of Engineering in Mechanical Engineering
Savitribai Phule Pune University
Pune, India
2017

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2021

DISTRIBUTED RANDOMIZED BLOCK
STOCHASTIC GRADIENT TRACKING METHODS:
RATE ANALYSIS AND NUMERICAL EXPERIMENTS

Thesis Approved:

Dr. Farzad Yousefian

Thesis Advisor

Dr. Chenang Liu

Dr. Bing Yao

ACKNOWLEDGMENTS

This thesis would not have been possible without the expertise of my incredible advisor, Dr. Farzad Yousefian. I will be forever indebted and want to thank him for his unwavering support and encouragement throughout my masters' studies at OSU. He has been a very patient, selfless and ideal advisor, and I highly respect his valuable time guiding and mentoring wherever necessary. I will always be grateful to Dr. Yousefian for his genuine concern and his encouragement towards my betterment.

I want to appreciate the help and support from my committee members, Dr. Bing Yao and Dr. Chenang Liu, for reviewing my thesis and providing their valuable feedback timely. I am thankful to my wonderful supervisor, Dr. Terry Collins, for believing in me and constantly motivating me to ensure my growth along with my responsibilities.

The support provided by the School of Industrial Engineering and Management (IEM) has been ideal and very fruitful for both my academic and research work. I am grateful to Dr. Sunderesh Heragu for his valuable guidance and encouragement. I am thankful to Dr. Tieming Liu and Dr. Balabhaskar Balasundaram for their support and insightful courses. IEM staff members also have been very supportive from the very beginning.

I am thankful to my colleagues and friends from the IEM department. Mainly, I would like to mention Ujjval Patel, Ayushi Naik, and Kushal Shah for always being helpful and supportive. Thanks to Dr. Harshal Kaushik for being a fantastic colleague and a good friend.

Finally, I am deeply grateful to my loving parents Mr. Vinayak Yevale and Mrs. Vibhavari Yevale, my brother Suyash Yevale, and my beloved Reeta Patil for their immense faith, support and sacrifices all these years.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: JAYESH VINAYAK YEVALE

Date of Degree: DECEMBER, 2021

Title of Study: DISTRIBUTED RANDOMIZED BLOCK STOCHASTIC GRADIENT TRACKING METHODS: RATE ANALYSIS AND NUMERICAL EXPERIMENTS

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract:

Distributed optimization has been a trending topic of research in the past few decades. This is mainly due to the recent advancements in the technology of wireless sensors and also the emerging applications in machine learning. Traditionally, optimization problems were addressed using centralized schemes where the data is assumed to be available all in one place. However, the main reasons that motivate the need for distributed implementations include: (i) the unavailability of the collected data in a centralized location, (ii) the privacy of the data among agents should be preserved, and (iii) the memory and computational power limitations of data processors. Accordingly, to address these challenges, distributed optimization provides a framework where agents (e.g., data processor, sensor) communicate their local information with each other over a network and seek to minimize a global objective function. In some applications, the data may have a huge sample size or a large number of attributes. The problems associated with this type of data are often known as big data problems. In this thesis, our goal is to address such high dimensional distributed optimization problems, where the computation of the local gradient mappings may become expensive.

Recently, a distributed optimization algorithm has been developed for addressing possibly large-scale problems by considering stochasticity. This method is called Distributed Stochastic Gradient Tracking (DSGT). We develop a novel iterative method called Distributed Randomized Block Stochastic Gradient Tracking (DRBSGT), that is a randomized block variant of the existing DSGT method. We derive new non-asymptotic convergence rates of the order $1/k$ and $1/k^2$ in terms of an optimality metric and a consensus violation metric, respectively. Importantly, while block coordinate schemes have been studied for distributed optimization problems before, the proposed algorithm appears to be the first randomized block-coordinate gradient tracking method that is equipped with the aforementioned convergence rate statements. We validate the performance of the proposed method on the MNIST and a synthetic data set under different network settings. A potential future research direction is to extend the results of this thesis to an asynchronous variant of the proposed method. This will allow for the consideration of communication delays.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

1.1 Motivating Example 1

1.2 Existing Algorithm 2

1.3 Research Contributions 3

II. THE PROPOSED ALGORITHM 4

2.1 Problem formulation 4

2.2 Literature Review 5

2.3 Notation 6

2.4 Algorithm Outline 8

2.5 Preliminaries for convergence analysis 11

2.6 Concluding Remarks 12

III. CONVERGENCE RATE ANALYSIS 13

3.1 Recursive Error Bounds 13

3.2 Rate Analysis 21

3.3 Concluding Remarks 24

IV. NUMERICAL EXPERIMENTS 25

4.1 Simulation 25

4.2 Insights 26

4.3 Concluding Remarks 27

Chapter Page

V. CONCLUSION AND FUTURE DIRECTION 28

REFERENCES 29

LIST OF TABLES

Table		Page
1	Comparison of this work with other recent gradient tracking schemes for distributed optimization	6
2	Parameters of various settings used for implementation	26
3	Objective function value comparision of Algorithm 2 vs. DSGT vs. ATC for 90% CIs	26

LIST OF FIGURES

Figure		Page
1	Examples of doubly stochastic undirected network (ring, complete, and star graph)	4
2	Algorithm 2 vs. DSGT vs. ATC in terms of objective function value and consensus error	27

CHAPTER I

INTRODUCTION

In this chapter, we present a motivating example about big data optimization problems in distributed optimization in Section 1.1 and present a recently developed algorithm for addressing distributed stochastic optimization problems in Section 1.2. We then present the main research contributions of this thesis in Section 1.3.

1.1 Motivating Example

Let us consider the regularized logistic regression loss minimization problem for binary classification applications. In a progressive manner, we show that this problem can be reformulated as a distributed stochastic optimization problem. Consider a data set denoted as $\mathcal{D} \triangleq \{(u_\ell, v_\ell) \in \mathbb{R}^n \times \{-1, +1\} \mid \ell \in \mathcal{S}\}$ where $\mathcal{S} \triangleq \{1, \ldots, s\}$ denotes an index set where u_ℓ denotes an input vector corresponding to a binary value v_ℓ. Then, the optimization problem can be defined as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \qquad \text{where } f(x) \triangleq \sum_{\ell \in \mathcal{S}} \ln\left(1 + \exp\left(-v_\ell u_\ell^T x\right)\right) + \frac{\mu}{2}\|x\|^2,$$

where $\mu > 0$ is a regularization parameter that is employed as a hyper-parameter. For a distributed implementation, let us assume that the data set $\mathcal{S}$ is distributed among m agents. Let $\mathcal{S}_i$ denote the data locally known by agent i where $\mathcal{S} = \cup_{i=1}^m \mathcal{S}_i$. Note that the number of data points may differ among the agents. We let $|\mathcal{S}_i|$ denote the number of data points in the set $\mathcal{S}_i$. We rewrite the preceding loss minimization problem as a distributed regularized logistic regression loss minimization problem as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^m f_i(x) \qquad \text{where } f_i(x) \triangleq \sum_{\ell \in \mathcal{S}_i} \ln\left(1 + \exp\left(-v_\ell u_\ell^T x\right)\right) + \frac{\mu}{2m}\|x\|^2,$$

where ℓ denotes the index of a sample from the set $\mathcal{S}_i$. Next, for any $i \in \{1, \ldots, m\}$, we introduce a discrete uniform random variable $\xi_i \in \mathbb{R}^{n+1}$ where defined as $\xi_i \triangleq (u_i, v_i)$ where (u_i, v_i) takes values in $\mathcal{S}_i$. From this definition, we can write

$$f_i(x) = |\mathcal{S}_i| \mathbb{E}_{\xi_i}\left[\ln\left(1 + \exp\left(-v_i u_i^T x\right)\right)\right] + \frac{\mu}{2m}\|x\|^2.$$

By multiplying and dividing by the size of the original data set, that is $|\mathcal{S}|$, and rearranging the terms we obtain

$$f_i(x) = |\mathcal{S}| \mathbb{E}_{\xi_i}\left[\frac{|\mathcal{S}_i|}{|\mathcal{S}|} \ln\left(1 + \exp\left(-v_i u_i^T x\right)\right) + \frac{\mu}{2m|\mathcal{S}|}\|x\|^2\right].$$

Let us now define a stochastic function $F_i(x, \xi_i)$ as

$$F_i(x, \xi_i) \triangleq \frac{|\mathcal{S}_i|}{|\mathcal{S}|} \ln\left(1 + \exp\left(-v_i u_i^T x\right)\right) + \frac{\mu}{2m|\mathcal{S}|}\|x\|^2,$$

Using this definition for F_i we can write

$$f_i(x) = |\mathcal{S}| \mathbb{E}_{\xi_i}\left[F_i(x, \xi_i)\right].$$

Thus, the original problem can be cast as the following distributed stochastic optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x), \qquad \text{where } f_i(x) = |\mathcal{S}| \mathbb{E}_{\xi_i}\left[F_i(x, \xi_i)\right].$$

Equivalently, we consider the following distributed stochastic formulation

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{m} \mathbb{E}_{\xi_i}\left[F_i(x, \xi_i)\right]. \tag{1.1.1}$$

The formulation (1.1.1) has been addressed in a recent work [15] where it is assumed that the m agents cooperatively solve this optimization problem over an undirected communication network. The proposed algorithm is called Distributed Stochastic Gradient Tracking (DSGT) that is described in the following section.

1.2 Existing Algorithm

Algorithm 1 Distributed Stochastic Gradient Tracking Method (DSGT)

1: **Input:** Agents choose a doubly stochastic weight matrix $\mathbf{W}$ and set an initial step-size $\gamma_0 > 0$. For all $i \in [m]$, agent i chooses a random initial point $x_{i,0} \in \mathbb{R}^n$
2: For all $i \in [m]$, agent i generates a realization of the random variable ξ_i, denoted as $\xi_{i,0}$, and evaluates the initial gradient tracker $y_{i,0} := \nabla f_i\left(x_{i,0}, \xi_{i,0}\right)$
3: **for** $k = 0, 1, \ldots,$ **do**
4: For all $i \in [m]$, agent i generates a realization of the random variable ξ_i, denoted as $\xi_{i,k+1}$, and evaluates the local gradient mapping $\nabla f_i\left(x_{i,k+1}, \xi_{i,k+1}\right)$
5: For all $i \in [m]$, agent i does the following updates:
6: $x_{i,k+1} := \sum_{j=1}^{m} W_{ij}\left(x_{j,k} - \gamma_k y_{j,k}\right)$
7: $y_{i,k+1} := \sum_{j=1}^{m} W_{ij} y_{j,k} + \nabla f_i\left(x_{i,k+1}, \xi_{i,k+1}\right) - \nabla f_i\left(x_{i,k}, \xi_{i,k}\right)$
8: **end for**

The outline of the DSGT method is presented in Algorithm 1. In this method, at iteration k, agent i does two main updates that are presented in step 6 and step 7 in Algorithm 1. The vector $x_{i,k} \in \mathbb{R}^n$ denotes the local copy of the decision variable maintained by agent i at time k and $y_{i,k} \in \mathbb{R}^n$ denotes the local copy maintained by agent i at iteration k that is used to track the average of the gradient mapping of the global objective function. This technique is called

gradient tracking and has been employed in the past few years in distributed optimization methods to help with an acceleration of the underlying algorithm. The scalar W_{ij} denotes nonnegative weights that agent i uses in its communication with any neighboring agent j. The scalar γ_k denotes a diminishing step-size parameter. In step 6 of the algorithm, the vector $x_{i,k}$ is updated by agent i while in step 7, agent i communicates with its neighbours and obtains $y_{i,k+1}$ using the gradient tracking vectors $y_{j,k}$.

1.3 Research Contributions

In this graduate thesis, the main contributions are as follows:

1. Firstly, we develop an algorithm called distributed randomized block stochastic gradient tracking (DRBSGT) for addressing distributed stochastic optimization problems of the form (1.1.1) with possibly a large dimension in the solution space. We employ randomized block-coordinate technique where agents only require to compute a randomly selected blocks of their local gradient mapping.

2. Secondly, we derive a rate of $\mathcal{O}(1/k)$ on a suboptimality metric and $\mathcal{O}(1/k^2)$ on a consensus violation metric for the DRBSGT algorithm. Importantly, while DRBSGT generalizes DSGT to a randomized block variant, these rate statements are comparable with those of DSGT, indicating that there is no sacrifice in terms of the order of magnitude of the rate statements.

3. Finally, we validate the theoretical claims. We compare the performance of our scheme with that of other existing gradient tracking schemes and provide preliminary results on different data sets and under different network assumptions. We consider the Modified National Institute of Standards and Technology (MNIST) and synthetic data sets for the numerical analysis of this thesis.

CHAPTER II

THE PROPOSED ALGORITHM

In this chapter, we consider distributed optimization problems over networks where each agent is associated with a smooth and strongly convex local objective function. This mathematical formulation captures a wide range of applications in several areas, including telecommunication, information processing, and machine learning. We present the problem formulation and the assumptions in Section 2.1. Section 2.2 summarizes the existing literature about the recent advancement in block-coordinate and gradient tracking schemes. Section 2.3 provides the notation used throughout the thesis. We present the outline of the proposed algorithm in Section 2.4, and in Section 2.5 we provides some preliminary results that will pave the way for the convergence analysis in this thesis in the next chapter. Section 2.6 concludes this chapter.

2.1 Problem formulation

We consider the following distributed optimization problem in which each agent has a local smooth and strongly convex cost function $f_i : \mathbb{R}^n \to \mathbb{R}$. We minimize the average of all cost functions

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{2.1.1}$$

where each agent is associated with a local objective function $f_i(x)$ and communicate over an undirected graph denoted by $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is a set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of ordered pairs of vertices.

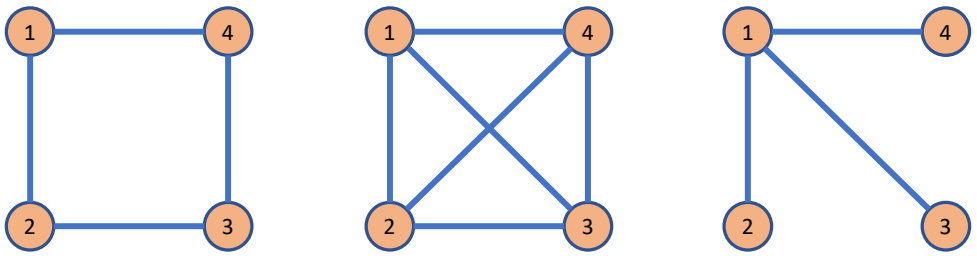

Figure 1: Examples of doubly stochastic undirected network (ring, complete, and star graph)

The contents of this thesis is submitted to Proceedings of the 2022 American Control Conference [24]

We let $\mathcal{N}(i)$ denote the set of neighbors of agent i, i.e., $\mathcal{N}(i) \triangleq \{j \mid (i,j) \in \mathcal{E}\}$. To solve the problem 2.1.1, we consider the following assumption.

Assumption 2.1.1 *For all $i \in \{1, \ldots, m\}$, function f_i is μ-strongly convex and L-smooth.*

Motivated by big data applications in machine learning and sensor networks, we are interested in addressing problem (2.1.1) in stochastic and high-dimensional settings. In particular, we assume that agents only have access to noisy local gradient mappings denoted by $\nabla f_i(\bullet, \xi_i)$ where ξ_i satisfies the following assumption.

Assumption 2.1.2 *For all $i \in \{1, \ldots, m\}$, random vectors $\xi_i \in \mathbb{R}^d$ are independent and for all $x \in \mathbb{R}^n$,*

$$\begin{aligned} &\mathbb{E}[\nabla f_i(x, \xi_i)|x] = \nabla f_i(x), \\ &\mathbb{E}\big[\| \nabla f_i(x, \xi_i) - \nabla f_i(x) \|^2 \, |x\big] \leq \nu^2 \quad \textit{for some } \nu > 0. \end{aligned} \tag{2.1.2}$$

This assumption of the above gradients holds true for many distributed learning problems. The function $f_i(x) \triangleq \mathbb{E}_{\xi_i}[F_i(x, \xi_i)]$ denotes the expected loss function of agent i.
To address high-dimensionality, we consider a block structure for x given by $x = \big[x^{(1)}; \ldots; x^{(b)}\big]$ where $x^{(\ell)} \in \mathbb{R}^{n_i}$ denotes the ℓ-th block-coordinate of $x \in \mathbb{R}^n$ and $\sum_{\ell=1}^{b} n_\ell = n$. The blocks in each iteration are selected randomly. We consider the following assumptions on the communication network.

Assumption 2.1.3 *The weight matrix $\mathbf{W}$ is double stochastic, and we have $w_{i,i} > 0$ for all $i \in [m]$.*

Assumption 2.1.4 *Let the graph $\mathcal{G}$ corresponding to the communication network be undirected and connected.*

2.2 Literature Review

Among the recent advancements in distributed optimization algorithms, gradient tracking methods have been recently studied. In these schemes, agents track the average of the global gradient mapping through communicating their estimate of the gradient locally with their neighbors in convex [12, 16, 15, 22], and nonconvex regime [8, 17, 3, 19]. In [16], Push-Pull, G-Push-Pull algorithms and their variants are developed for addressing distributed optimization over directed graphs and a linear rate of convergence was established. Recently, a stochastic variant of gradient tracking methods has been developed in [15], namely the DSGT method, where non-asymptotic convergence rates of the order $1/k$ and $1/k^2$ in terms of an optimality metric and a consensus violation metric were derived, respectively. Further, in [7], integrating the ideas from DIGing [12] and a fast incremental gradient method (SAGA) [2], S-DIGing algorithm is developed.
In the aforementioned schemes, agents have to evaluate full-dimensional gradient vectors at each iteration of the method. A popular avenue for addressing this issue is the class of block-coordinate schemes. Block-coordinate schemes, and specifically their randomized variants, have been widely studied in addressing optimization problems and games in deterministic [13,

Table 1: Comparison of this work with other recent gradient tracking schemes for distributed optimization

Ref.	Method	Problem class	Network topology	Problem formulation	Rate(s)
[15]	DSGT, GSGT	$f_i \in C^{1,1}_{\mu,L}$	Undirected	$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m f_i(x) \triangleq \mathbb{E}[F_i(x,\xi_i)]$	suboptimality: $\mathcal{O}\left(1/k\right)$ consensus: $\mathcal{O}\left(1/k^2\right)$
[16]	Push-Pull G-Push-Pull	$f_i \in C^{1,1}_{\mu,L}$	Directed	$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m f_i(x)$	linear
[14]	Block-SONATA	$f_i \in C^{1,1}$, $r_\ell \in C^{0,0}_{0,0}$	Directed	$\min_x \sum_{i=1}^m f_i(x) + \sum_{l=1}^B r_l(x_l)$ s.t. $x_l \in K_\ell,\ \ell \in \{1,\ldots,B\}$	–
[21]	S-AB	$f_i \in C^{1,1}_{\mu,L}$	Directed	$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m f_i(x) \triangleq \mathbb{E}[F_i(x,\xi_i)]$	linear
[6]	Network-DANE	$f_i \in C^{1,1}_{\mu,L}$	Undirected	$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^N \ell(x;z_i)$, where $N=$ total samples, z_i is the i^{th} sample.	linear
[7]	S-Diging	$f_i \in C^{1,1}_{\mu,L}$	Undirected	$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m f_i(x) \triangleq \mathbb{E}[F_i(x,\xi_i)]$	linear
[10]	GNSD	$f_i \in C^{1,1}$	Undirected	$\min_x \frac{1}{m}\sum_{i=1}^m f_i(x) \triangleq \mathbb{E}[F_i(x,\xi_i)]$ s.t. $x_i = x_j,\ j\in\mathcal{N}(i),\ \forall i$	$\mathcal{O}\left(1/\sqrt{k}\right)$
This work	DRBSGT	$f_i \in C^{1,1}_{\mu,L}$	Undirected	$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m f_i(x) \triangleq \mathbb{E}[f_i(x,\xi_i)]$	suboptimality: $\mathcal{O}\left(1/k\right)$ consensus: $\mathcal{O}\left(1/k^2\right)$

18, 20, 5, 4] and stochastic regimes [1, 23, 11]. In randomized block schemes, at each iteration only a randomly selected block of the gradient mapping is evaluated, requiring significantly lower computational effort per iteration than the standard schemes. Although block-coordinate schemes have been studied for distributed optimization problems before [9, 14], the convergence rate statements of randomized block gradient tracking methods are not yet established. Inspired by the DSGT method [15], our goal in this paper lies in extending DSGT to a randomized block variant that is equipped with new non-asymptotic performance guarantees.

2.3 Notation

Throughout this thesis, the vectors are default to columns and the matrices are represented in bold. Let $x_i \in \mathbb{R}^n$ holds a local copy of decision variable and an variable $y_i \in \mathbb{R}^n$ tracks the average gradient mapping. The values of these variables at the iteration k is denoted by $x_{i,k}$ and $y_{i,k}$, respectively. We let x^* to denote the unique global optimal solution of problem (2.1.1). We use $[m]$ to denote $\{1,2,\ldots,m\}$ for any integer $m \geq 1$. We let $\|\bullet\|$ denote the Euclidean norm and Frobenius norm of a vector and a matrix, respectively.
We define $\mathbf{U}_\ell \in \mathbb{R}^{n\times n_\ell}$ for $\ell \in [b]$ such that

$$[\mathbf{U}_1,\ldots,\mathbf{U}_b] = \mathbf{I}_n,$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Note that we can write,

$$\begin{aligned} x &= \sum_{\ell=1}^{b} \mathbf{U}_\ell x^{(\ell)}, \\ \left\|\mathbf{U}_\ell x^{(\ell)}\right\|^2 &= \left\|x^{(\ell)}\right\|^2 \\ \sum_{\ell=1}^{b} \left\|\mathbf{U}_\ell x^{(\ell)}\right\|^2 &= \|x\|^2. \end{aligned} \tag{2.3.1}$$

The function $f : X \to \mathbb{R}$ is said to be Lipschitz smooth with parameter L if

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \text{ and parameter } L > 0,$$

where X is a set and $x, y \in X$.
A continuous differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-strongly convex if

$$f(x) \geq f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2} \|x - y\|^2 \text{ and the parameter } \mu > .0.$$

We consider the following notation throughout the thesis. We let

$$\mathbf{x} := [x_1, x_2, \ldots, x_m]^T \in \mathbb{R}^{m \times n}, \qquad \mathbf{y} := [y_1, y_2, \ldots, y_m]^T \in \mathbb{R}^{m \times n},$$

and

$$\bar{x} := \frac{1}{m} \mathbf{1}^T \mathbf{x} \in \mathbb{R}^{1 \times n}, \qquad \bar{y} := \frac{1}{m} \mathbf{1}^T \mathbf{y} \in \mathbb{R}^{1 \times n},$$

where $\mathbf{1}$ indicates the vector for all entries as 1. We define the total objective function as

$$f(x) \triangleq \sum_{i=1}^{m} f_i(x), \qquad \mathbf{f}(\mathbf{x}) \triangleq \sum_{i=1}^{m} f_i(x_i)$$

and let

$$f_i(x) \triangleq \mathbb{E}[f_i(x, \xi_i) \mid x] .$$

In addition, we denote

$$\begin{aligned} &\boldsymbol{\xi} := [\xi_1, \xi_2, \ldots, \xi_m]^T \in \mathbb{R}^{m \times d}, \\ &\boldsymbol{\ell} := [\ell_1, \ell_2, \ldots, \ell_m]^T \in \mathbb{R}^{m \times 1}, \\ &\mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) \triangleq [\nabla f_1(x_1, \xi_1), \ldots, \nabla f_m(x_m, \xi_m)]^T, \\ &\mathbf{G}(\mathbf{x}), \triangleq \mathbb{E}[\mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) \mid \mathbf{x}] = [\nabla f_1(x_1), \ldots, \nabla f_m(x_m)]^T, \\ &G(\mathbf{x}, \boldsymbol{\xi}) \triangleq \frac{1}{m} \mathbf{1}^T \mathbf{G}(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{m} \sum_{i=1}^{m} f_i(x_i, \xi_i) \in \mathbb{R}^{m \times n}, \\ &G(\mathbf{x}) \triangleq \mathbb{E}[G(\mathbf{x}, \boldsymbol{\xi}) \mid \mathbf{x}] , \\ &\mathcal{G}(x) \triangleq G(\mathbf{1} x^T) = \frac{1}{m} \nabla f(x). \end{aligned} \tag{2.3.2}$$

Algorithm 2 Distributed Randomized Block Stochastic Gradient Tracking (DRBSGT)

1: **Input:** Agents choose $\gamma_0 > 0$ the weight matrix $\mathbf{W}$. For all $i \in [m]$, agent i chooses a random initial point $x_{i,0} \in \mathbb{R}^n$

2: For all $i \in [m]$, agent i generates realizations of the random variables $\xi_{i,0}$ and $\ell_{i,0}$ and sets $y_{i,0}^{(\ell_{i,0})} := \nabla^{\ell_{i,0}} f_i(x_{i,0}, \xi_{i,0})$ and $y_{i,0}^{(\ell)} := 0$ for all $\ell \neq \ell_{i,0}$.

3: **for** $k = 0, 1, \ldots,$ **do**

4: For all $i \in [m]$, agent i does the following update for $\ell \in [b]$:

$$x_{i,k+1} := \sum_{j=1}^{m} W_{ij} \left(x_{j,k} - \gamma_k y_{j,k} \right).$$

5: For all $i \in [m]$, agent i generates realizations of the random variables $\xi_{i,k+1}$ and $\ell_{i,k+1}$.

6: For all $i \in [m]$, agent i does the following update:

$$y_{i,k+1}^{(\ell)} := \begin{cases} \sum_{j=1}^m W_{ij} y_{j,k}^{(\ell)} + \nabla^{(\ell)} f_i(x_{i,k+1}, \xi_{i,k+1}) - \nabla^{(\ell)} f_i(x_{i,k}, \xi_{i,k}), & \text{if } \ell = \ell_{i,k+1} = \ell_{i,k} \\ \sum_{j=1}^m W_{ij} y_{j,k}^{(\ell)} + \nabla^{(\ell)} f_i(x_{i,k+1}, \xi_{i,k+1}), & \text{if } \ell = \ell_{i,k+1} \neq \ell_{i,k} \\ \sum_{j=1}^m W_{ij} y_{j,k}^{(\ell)} - \nabla^{(\ell)} f_i(x_{i,k}, \xi_{i,k}) & \text{if } \ell = \ell_{i,k} \neq \ell_{i,k+1} \\ \sum_{j=1}^m W_{ij} y_{j,k}^{(\ell)}, & \text{if } \ell \neq \ell_{i,k+1}, \ell \neq \ell_{i,k}. \end{cases}$$

7: **end for**

2.4 Algorithm Outline

The outline of the proposed algorithm is presented by Algorithm 2 (DRBSGT). This algorithm is an extension to the existing Algorithm 1 (DSGT). Additionally, in the newly proposed algorithm at every iteration in step 6, the agents only compute a randomly selected block of their local gradient mapping. This computation is under the following assumption.

Assumption 2.4.1 *For $k \geq 0$ and $i \in [m]$, let $\ell_{i,k} \in [b]$ be generated from a discrete uniform distribution, i.e., Prob$(\ell_{i,k} = \ell) = b^{-1}$ for all $\ell \in [b]$. Also, we assume these uniform distributions are independent from each other and from the random variables ξ_i.*

Algorithm 2 can be compactly written as

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{W}(\mathbf{x}_k - \gamma_k \mathbf{y}_k), \\ \mathbf{y}_{k+1} &= \mathbf{W}\mathbf{y}_k + b^{-1} \left(\mathbf{G}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{e}_{k+1} \right) - b^{-1} \left(\mathbf{G}(\mathbf{x}_k, \boldsymbol{\xi}_k) - \mathbf{e}_k \right). \end{aligned} \tag{2.4.1}$$

Throughout, we define the history of the method for $k \geq 1$ as

$$\mathcal{F}_k \triangleq \cup_{i=1}^m \{x_{i,0}, \ell_{i,0}, \xi_{i,0}, \ldots, \ell_{i,k-1}, \xi_{i,k-1}\}$$

where $\mathcal{F}_0 \triangleq \cup_{i=1}^m \{x_{i,0}, \ell_{i,0}, \xi_{i,0}\}$. We define the stochastic errors of the randomized block-coordinate scheme as

$$\begin{aligned} e_{i,k} &\triangleq \nabla f_i(x_{i,k}, \xi_{i,k}) - b \mathbf{U}_{\ell_{i,k}} \nabla^{\ell_{i,k}} f_i(x_{i,k}, \xi_{i,k}), \\ \mathbf{e}_k &\triangleq [e_{1,k}, e_{2,k}, \ldots, e_{m,k}]^T \in \mathbb{R}^{m \times n}, \\ \bar{e}_k &\triangleq \frac{1}{m} \mathbf{1}^T \mathbf{e}_k = \frac{1}{m} \sum_{i=1}^m e_{i,k}. \end{aligned} \tag{2.4.2}$$

Next, we show some key properties of the randomized errors.

Lemma 2.4.1 *We have for all $i \in [m]$ and $k \geq 0$*
(a) $\mathbb{E}[e_{i,k} \mid \mathcal{F}_k] = \mathbb{E}[\bar{e}_k \mid \mathcal{F}_k] = 0$.
(b) $\mathbb{E}[\|e_{i,k}\|^2 \mid \mathcal{F}_k] \leq (b-1)\left(\nu^2 + \|\nabla f_i(x_{i,k})\|^2\right)$.
(c) $\mathbb{E}[\|\bar{e}_k\|^2 \mid \mathcal{F}_k] \leq (b-1)\nu^2 + \frac{b-1}{m}\|\mathbf{G}(\mathbf{x}_k)\|^2$.

Proof. (a) We can write

$$\begin{aligned}
&\mathbb{E}[e_{i,k} \mid \mathcal{F}_k \cup \{\xi_{i,k}\}] \\
&= \nabla f_i(x_{i,k}, \xi_{i,k}) - b\mathbb{E}\big[\mathbf{U}_{\ell_{i,k}} \nabla^{\ell_{i,k}} f_i(x_{i,k}, \xi_{i,k}) \mid \mathcal{F}_k \cup \{\xi_{i,k}\}\big] \\
&= \nabla f_i(x_{i,k}, \xi_{i,k}) - b\sum_{\ell=1}^{b} b^{-1}\mathbf{U}_\ell \nabla^\ell f_i(x_{i,k}, \xi_{i,k}) = 0.
\end{aligned}$$

The desired result follows by taking expectations from the preceding relation with respect to $\xi_{i,k}$.
(b) Throughout the proof, we use the compact notation $\tilde{\nabla}_{i,k} \triangleq \nabla f_i(x_{i,k}, \xi_{i,k})$. We can write

$$\begin{aligned}
\|e_{i,k}\|^2 &= \left\|\left(\tilde{\nabla}_{i,k} - b\mathbf{U}_{\ell_{i,k}}\tilde{\nabla}_{i,k}^{\ell_{i,k}}\right)\right\|^2 \\
&= \left\|\tilde{\nabla}_{i,k}\right\|^2 + b^2\left\|\mathbf{U}_{\ell_{i,k}}\tilde{\nabla}_{i,k}^{\ell_{i,k}}\right\|^2 - 2b\left(\tilde{\nabla}_{i,k}\right)^T \mathbf{U}_{\ell_{i,k}}\tilde{\nabla}_{i,k}^{\ell_{i,k}}.
\end{aligned}$$

Taking conditional expectations, we have

$$\mathbb{E}\left[\|e_{i,k}\|^2 \mid \mathcal{F}_k \cup \left(\cup_{j=1}^m \{\xi_{j,k}\}\right)\right] = \left\|\tilde{\nabla}_{i,k}\right\|^2 + b\sum_{\ell=1}^{b}\left\|\mathbf{U}_\ell\tilde{\nabla}_{i,k}^{\ell}\right\|^2 - 2\tilde{\nabla}_{i,k}^T\sum_{\ell=1}^{b}\mathbf{U}_\ell\tilde{\nabla}_{i,k}^{\ell}.$$

We have $\sum_{\ell=1}^{b}\left\|\mathbf{U}_\ell\tilde{\nabla}_{i,k}^{\ell}\right\|^2 \overset{(2.3.1)}{=} \left\|\tilde{\nabla}_{i,k}\right\|^2$. From the two preceding relations, we obtain

$$\mathbb{E}\left[\|e_{i,k}\|^2 \mid \mathcal{F}_k \cup \left(\cup_{j=1}^m \{\xi_{j,k}\}\right)\right] = (b-1)\left\|\tilde{\nabla}_{i,k}\right\|^2.$$

The desired relation holds by taking expectations with respect to $\cup_{j=1}^m\{\xi_{j,k}\}$ from both sides and invoking Assumption 2.1.2.
(c) This relation follows from part (b) and by noting that we have

$$\|\bar{e}_k\|^2 \leq \frac{1}{m}\sum_{i=1}^{m}\|e_{i,k}\|^2.$$

■

The following two lemmas will be applied in the analysis and can be found in [15].

Lemma 2.4.2 *Let Assumption 2.1.3 and 2.1.4, holds true.Let ρ_W, denote the spectral norm of the matrix $\mathbf{W} - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ and $\bar{u} \triangleq \frac{1}{m}\mathbf{1}^T\mathbf{u}$. Then, $\rho_W < 1$, and $\|\mathbf{W}\mathbf{u} - \mathbf{1}\bar{u}\| \leq \rho_W\|\mathbf{u} - \mathbf{1}\bar{u}\|$ for all $\mathbf{u} \in \mathbb{R}^{m\times n}$.*

Lemma 2.4.3 *Let Assumption 2.1.1 hold. For any $\alpha \leq \frac{2}{\mu+L}$, we have*

$$\|\bar{x}_k - \alpha\mathcal{G}(\bar{x}_k) - x^*\| \leq (1-\mu\alpha)\|\bar{x}_k - x^*\|.$$

We also make use of the following result.

Lemma 2.4.4 *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m\times n}$. *Then,*
(a) $\langle \mathbf{u}, \mathbf{v}\rangle = \sum_{i=1}^{m} u_{i\bullet} v_{i\bullet}^T = \sum_{j=1}^{n} u_{\bullet j}^T v_{\bullet j}$.
(b) $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2$, *where* $\|\bullet\|$ *denotes the Frobenius norm of a matrix.*
(c) For any scalar $\lambda > 0$, *we have* $|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\| \leq \frac{1}{2}\left(\lambda\|\mathbf{u}\|^2 + \frac{1}{\lambda}\|\mathbf{v}\|^2\right)$.

Proof. (a) By the definition of $\langle \mathbf{u}, \mathbf{v}\rangle$. We have

$$\langle \mathbf{u}, \mathbf{v}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}v_{ij} = \sum_{i=1}^{m}\left(u_{i1}v_{i1} + \ldots + u_{in}v_{in}\right) = \sum_{i=1}^{m} u_{i\bullet}v_{i\bullet}^T.$$

Similarly, we also have

$$\langle \mathbf{u}, \mathbf{v}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}v_{ij} = \sum_{j=1}^{n}\left(u_{1j}v_{1j} + \ldots + u_{mj}v_{mj}\right) = \sum_{j=1}^{n} u_{\bullet j}^T v_{\bullet j}.$$

(b) By using the definition of the Frobenius norm, we have

$$\begin{aligned} \|\mathbf{u}+\mathbf{v}\|^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(u_{ij}+v_{ij}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(u_{ij}^2 + 2u_{ij}v_{ij} + v_{ij}^2\right) \\ \Rightarrow \quad \|\mathbf{u}+\mathbf{v}\|^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}\left(u_{ij}v_{ij}\right) + \sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij}^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2, \end{aligned}$$

where we used Lemma 2.4.4 (a) in the preceding inequality.
(c) Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m\times n}$ and $\lambda > 0$, and using the Cauchy-Schwarz inequality. We get

$$\begin{aligned} |\langle \mathbf{u}, \mathbf{v}\rangle| = \left|\sum_{i=1}^{m}\sum_{j=1}^{n} u_{i,j}v_{i,j}\right| &\leq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} u_{i,j}^2 v_{i,j}^2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} u_{i,j}^2}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} v_{i,j}^2} \\ \Rightarrow \quad |\langle \mathbf{u}, \mathbf{v}\rangle| &\leq \|\mathbf{u}\|\,\|\mathbf{v}\|. \end{aligned}$$

We use the properties of matrix norms to obtain the second inequality

$$\begin{aligned} & \|\lambda\mathbf{u} - \mathbf{v}\|^2 \geq 0 \\ \Rightarrow \quad & \lambda^2\|\mathbf{u}\|^2 - 2\lambda\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2 \geq 0 \\ \Rightarrow \quad & \frac{\lambda}{2}\|\mathbf{u}\|^2 - \|\mathbf{u}\|\,\|\mathbf{v}\| + \frac{1}{2\lambda}\|\mathbf{v}\|^2 \geq 0 \\ \Rightarrow \quad & \frac{1}{2}\left(\lambda\|\mathbf{u}\|^2 + \frac{1}{\lambda}\|\mathbf{v}\|^2\right) \geq \|\mathbf{u}\|\,\|\mathbf{v}\|. \end{aligned}$$

From the preceding two relations, we obtain

$$|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\| \leq \frac{1}{2}\left(\lambda\|\mathbf{u}\|^2 + \frac{1}{\lambda}\|\mathbf{v}\|^2\right).$$

■

2.5 Preliminaries for convergence analysis

We begin with presenting some important properties of the gradient mappings which are essential for the convergence analysis.

Lemma 2.5.1 *Consider Algorithm 2. Let Assumptions 2.1.1, 2.1.2, and 2.4.1 hold. Then, for all $k \geq 0$, the following results hold.*
(a) $b\bar{y}_k = G(\mathbf{x}_k, \boldsymbol{\xi}_k) - \bar{e}_k$.
(b) $\mathbb{E}[b\bar{y}_k \mid \mathcal{F}_k] = G(\mathbf{x}_k)$.
(c) $\mathbb{E}[\|b\bar{y}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] \leq \left(\frac{1}{m} + b - 1\right)\nu^2 + \frac{b-1}{m}\|\mathbf{G}(\mathbf{x}_k)\|^2$.
(d) For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m\times n}$, $\|G(\mathbf{u}) - G(\mathbf{v})\| \leq \frac{L}{\sqrt{m}}\|\mathbf{u} - \mathbf{v}\|$.
(e) $\|G(\mathbf{x}_k) - \mathcal{G}(\bar{x}_k)\| \leq \frac{L}{\sqrt{m}}\|\mathbf{x}_k - \mathbf{1}_m\bar{x}_k\|$.
(f) $\|\mathcal{G}(\bar{x}_k)\| \leq L\|\bar{x}_k - x^*\|$.
(g) $\|\mathbf{G}(\mathbf{x}_k)\|^2 \leq 2L^2\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2mL^2\|\bar{x}_k - x^*\|^2$.

Proof. (a) We use induction on k. For $k = 0$ we have

$$b\bar{y}_0 = \frac{b}{m}\mathbf{1}^T\mathbf{y}_k = \frac{b}{m}\sum_{i=1}^{m} y_{i,0} = \frac{b}{m}\sum_{i=1}^{m} b^{-1}\left(\nabla f_i(x_{i,0}, \xi_{i,0}) - e_{i,0}\right)$$
$$\Rightarrow \quad b\bar{y}_0 = G(\mathbf{x}_0, \boldsymbol{\xi}_0) - \bar{e}_0.$$

Let us assume that the relation in part (a) holds true for some k. We show that it holds true for $k+1$.

$$\begin{aligned} b\bar{y}_{k+1} &= \frac{b}{m}\mathbf{1}^T\left(\mathbf{W}\mathbf{y}_k + b^{-1}\left(\mathbf{G}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{e}_{k+1}\right) - b^{-1}\left(\mathbf{G}(\mathbf{x}_k, \boldsymbol{\xi}_k) - \mathbf{e}_k\right)\right) \\ &= \frac{b}{m}\mathbf{1}^T\mathbf{y}_k + \frac{1}{m}\mathbf{1}^T\left(\mathbf{G}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{e}_{k+1}\right) - \frac{1}{m}\mathbf{1}^T\left(\mathbf{G}(\mathbf{x}_k, \boldsymbol{\xi}_k) - \mathbf{e}_k\right) \\ &= b\bar{\mathbf{y}}_k + \frac{1}{m}\mathbf{1}^T\left(\mathbf{G}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{e}_{k+1}\right) - \frac{1}{m}\mathbf{1}^T\left(\mathbf{G}(\mathbf{x}_k, \boldsymbol{\xi}_k) - \mathbf{e}_k\right) \\ &= \frac{1}{m}\mathbf{1}^T\left(\mathbf{G}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{e}_{k+1}\right). \end{aligned}$$

Therefore, the induction hypothesis statement holds for $k+1$ and hence, the desired relation holds for all $k \geq 0$.
(b) Taking conditional expectations from the equation in part (a) and utilizing Lemma 2.4.1(a), we have

$$\begin{aligned} \mathbb{E}[b\bar{y}_k \mid \mathcal{F}_k] &= \mathbb{E}[G(\mathbf{x}_k, \boldsymbol{\xi}_k) - \bar{e}_k \mid \mathcal{F}_k] = G(\mathbf{x}_k) - \mathbb{E}[\bar{e}_k \mid \mathcal{F}_k] \\ &= G(\mathbf{x}_k) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[e_{i,k} \mid \mathcal{F}_k] = G(\mathbf{x}_k). \end{aligned}$$

(c) We can expand the following using Lemma 2.5.1 (a). We get

$$\begin{aligned} \mathbb{E}\left[\|b\bar{y}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|G(\mathbf{x}_k, \boldsymbol{\xi}_k) - \bar{e}_k - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\|G(\mathbf{x}_k, \boldsymbol{\xi}_k) - G(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\|\bar{e}_k\|^2 \mid \mathcal{F}_k\right] \\ &\quad + 2\mathbb{E}_{\boldsymbol{\xi}_k}\left[\mathbb{E}_{\boldsymbol{\ell}_k}\left[\bar{e}_k^T(G(\mathbf{x}_k, \boldsymbol{\xi}_k) - G(\mathbf{x}_k)) \mid \mathcal{F}_k \cup \left(\cup_{j=1}^{m}\{\xi_{j,k}\}\right)\right]\right] \\ &\leq \frac{\nu^2}{m} + (b-1)\nu^2 + \frac{b-1}{m}\|\mathbf{G}(\mathbf{x}_k)\|^2, \end{aligned}$$

where the last relation is obtained from Lemmas 2.4.1 and 2.1.2.
(d) For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{m\times n}$, with $u_i, v_i \in \mathbb{R}^n$ denoting the i^{th} row of $\mathbf{u}, \mathbf{v}$, respectively, we have

$$\begin{aligned}\|G(\mathbf{u}) - G(\mathbf{v})\| &= \left\|\frac{1}{m}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{u}) - \frac{1}{m}\mathbf{1}^T\nabla\mathbf{f}\left(\mathbf{v}\right)\right\| = \frac{1}{m}\left\|\sum_{i=1}^{m}\nabla f_i(u_i) - \sum_{i=1}^{m}\nabla f_i\left(v_i\right)\right\| \\ &\leq \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i(u_i) - \nabla f_i\left(v_i\right)\right\| \leq \frac{L}{\sqrt{m}}\left\|\mathbf{u} - \mathbf{v}\right\|.\end{aligned}$$

(e) By expanding the $\mathcal{G}(\bar{x}_k)$ using Equation 2.3.2. We get

$$\|G(\mathbf{x}_k) - \mathcal{G}(\bar{x}_k)\| = \|G(\mathbf{x}_k) - G(\mathbf{1}\bar{x}_k)\| \leq \frac{L}{\sqrt{m}}\|\mathbf{x}_k - \mathbf{1}_m\bar{x}_k\|.$$

(f) By Invoking $G(\mathbf{1}x^*) = \mathbf{0}$, we have

$$\|\mathcal{G}(\bar{x}_k)\| = \|G(\mathbf{1}\bar{x}_k)\| = \|G(\mathbf{1}\bar{x}_k) - G(\mathbf{1}x^*)\| \leq L\|\bar{x}_k - x^*\|.$$

(g) We can introduce the inequality and write as

$$\begin{aligned}\|\mathbf{G}(\mathbf{x}_k)\|^2 &\leq 2\|\mathbf{G}_k - \mathbf{G}(\mathbf{1}\bar{x}_k)\|^2 + 2\|\mathbf{G}(\mathbf{1}\bar{x}_k)\|^2 \\ &\leq 2L^2\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2\|\mathbf{G}(\mathbf{1}\bar{x}_k) - \mathbf{G}(\mathbf{1}x^*)\|^2 \\ &\leq 2L^2\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2mL^2\|\bar{x}_k - x^*\|^2.\end{aligned}$$

■

2.6 Concluding Remarks

The above preliminaries are essential for the convergence analysis. We will proceed to derive the rate statements on three recursive error bounds that are $\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right]$, $\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right]$, and $\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right]$ in the next chapter of this thesis. The term $\mathbb{E}[\|\bar{x}_k - x^*\|^2]$ denotes the suboptimality metric that measures the expected squared distance of the average value of solutions obtained from all agents from the optimal solution of the problem of interest. The other two metrics, i.e., $\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right]$ and $\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right]$, are called the consensus violation metrics for the vectors $x_{i,k}$ and $y_{i,k}$, respectively. We need these error metrics to be bounded to proceed with deriving their rate statements.

CHAPTER III

CONVERGENCE RATE ANALYSIS

In this chapter, we derive recursive error bounds for the error metrics $\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right]$, $\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right]$, and $\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right]$. In Section 3.1, we derive three recursive error bounds that will be used to derive rate statements. In Section 3.2, we derive the convergence and consensus rate statements. Section 3.3 concludes this chapter.

3.1 Recursive Error Bounds

Proposition 3.1.1 (Recursive error bounds) *Consider Algorithm 2. Let Assumptions 2.1.1, 2.1.2, 2.1.3, 2.1.4, and 2.4.1 hold. Then, if* $\gamma_k \leq \min\left\{\frac{2}{\mu+L}, \frac{b\mu}{4(b-1)L^2}\right\}$, *for any* $\eta > 0$ *we have*

$$
\begin{aligned}
(a)\ \mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2\right] \quad &\leq \left(1 - \frac{\mu b^{-1}\gamma_k}{2}\right)\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] \\
&\quad + \frac{b^{-1}\gamma_k L^2}{m}\left(\frac{1}{\mu} + b^{-1}(2b-1)\gamma_k\right)\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right] \\
&\quad + b^{-2}\gamma_k^2\left(\frac{1}{m} + b - 1\right)\nu^2. \\
(b)\ \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2\right] \quad &\leq \frac{1+\rho_W^2}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right] + \frac{\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right]. \\
(c)\ \mathbb{E}\left[\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}_{k+1}\|^2\right] \quad &\leq \left((1+b^{-1}\eta)\rho_W^2 + \gamma_k^2(\frac{1}{b^2} + \frac{1}{b\eta})\left(2L^2\rho_W^2\right.\right. \\
&\quad \left.\left. + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)\right)\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right] \\
&\quad + 2L^2 m\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\left(b^{-2}L^2\gamma_k^2\right. \\
&\quad \left. + (b-1)\left(3 + L^2\gamma_k^2 b^{-2}\right)\right)\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] \\
&\quad + 2L^2\left(b^{-2}L^2\gamma_k^2 + (b-1)\left(3 + L^2\gamma_k^2 b^{-2}\right.\right. \\
&\quad \left.\left. + b^{-1}\gamma_k L^2\left(\frac{1}{\mu} + b^{-1}(2b-1)\gamma_k\right)\right)\right. \\
&\quad \left. + \|\mathbf{W} - \mathbf{I}\|^2\right)\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right] \\
&\quad + \left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\nu^2\left(mL^2 b^{-2}\left(\frac{1}{m} + b - 1\right)\gamma_k^2 + 3mb\right).
\end{aligned}
$$

Proof. (a) Considering the first update rule from (2.4.1), and multiplying both sides by averaging operator $\frac{1}{m}\mathbf{1}^T$, and noting that $\mathbf{1}^T\mathbf{W} = \mathbf{1}^T$ from Assumption 2.1.3, we obtain $\bar{x}_{k+1} = \bar{x}_k - \gamma_k \bar{y}_k$. Using Lemma 2.5.1(b) and (c), we can write

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|\bar{x}_k - \gamma_k \bar{y}_k - x^*\|^2 \mid \mathcal{F}_k\right] \\
&= \|\bar{x}_k - x^*\|^2 - 2\gamma_k(\bar{x}_k - x^*)^T \mathbb{E}\left[\bar{y}_k \mid \mathcal{F}_k\right] + \gamma_k^2 \mathbb{E}\left[\|\bar{y}_k\|^2 \mid \mathcal{F}_k\right] \\
&= \|\bar{x}_k - x^*\|^2 - 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T G(\mathbf{x_k}) \\
&+ b^{-2}\gamma_k^2 \mathbb{E}\left[\|b\bar{y}_k - G(\mathbf{x_k}) + G(\mathbf{x_k})\|^2 \mid \mathcal{F}_k\right] \\
&= \|\bar{x}_k - x^*\|^2 - 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T G(\mathbf{x_k}) \\
&+ b^{-2}\gamma_k^2 \mathbb{E}\left[\|b\bar{y}_k - G(\mathbf{x_k})\|^2 \mid \mathcal{F}_k\right] + b^{-2}\gamma_k^2 \mathbb{E}\left[\|G(\mathbf{x_k})\|^2 \mid \mathcal{F}_k\right] \\
&+ 2b^{-2}\gamma_k^2 G(\mathbf{x_k})^T \mathbb{E}\left[b\bar{y}_k - G(\mathbf{x_k}) \mid \mathcal{F}_k\right] \\
&= \|\bar{x}_k - x^*\|^2 - 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T G(\mathbf{x_k}) \\
&+ b^{-2}\gamma_k^2 \mathbb{E}\left[\|b\bar{y}_k - G(\mathbf{x_k})\|^2 \mid \mathcal{F}_k\right] + b^{-2}\gamma_k^2 \|G(\mathbf{x_k})\|^2 \\
&\leq \|\bar{x}_k - x^*\|^2 - 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T G(\mathbf{x_k}) + b^{-2}\gamma_k^2 \|G(\mathbf{x_k})\|^2 \\
&+ b^{-2}\gamma_k^2 \left(\left(\frac{1}{m} + b - 1\right)\nu^2 + \frac{b-1}{m}\|\mathbf{G}(\mathbf{x}_k)\|^2\right).
\end{aligned}$$

Adding and subtracting $\mathcal{G}(\bar{x}_k)$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] &\leq \|\bar{x}_k - x^*\|^2 - 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T (G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k)) \\
&- 2b^{-1}\gamma_k(\bar{x}_k - x^*)^T \mathcal{G}(\bar{x}_k) + b^{-2}\gamma_k^2 \|G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k)\|^2 \\
&+ b^{-2}\gamma_k^2 \|\mathcal{G}(\bar{x}_k)\|^2 + 2b^{-2}\gamma_k^2 (G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k))^T \mathcal{G}(\bar{x}_k) \\
&+ b^{-2}\gamma_k^2 \left(\frac{1}{m} + b - 1\right)\nu^2 + \frac{2L^2 b^{-2}\gamma_k^2 (b-1)}{m}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \\
&+ 2(b-1)L^2 b^{-2}\gamma_k^2 \|\bar{x}_k - x^*\|^2 \\
&\leq \|\bar{x}_k - x^* - b^{-1}\gamma_k \mathcal{G}(\bar{x}_k)\|^2 \\
&- 2b^{-1}\gamma_k(\bar{x}_k - b^{-1}\gamma_k \mathcal{G}(\bar{x}_k) - x^*)^T (G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k)) \\
&+ b^{-2}\gamma_k^2 \|G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k)\|^2 + b^{-2}\gamma_k^2 \left(\frac{1}{m} + b - 1\right)\nu^2 \\
&+ \frac{2L^2 b^{-2}\gamma_k^2 (b-1)}{m}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2(b-1)L^2 b^{-2}\gamma_k^2 \|\bar{x}_k - x^*\|^2.
\end{aligned}$$

Invoking Lemmas 2.5.1 and 2.4.3 and the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k\right] &\leq (1 - b^{-1}\mu\gamma_k)^2 \|\bar{x}_k - x^*\|^2 \\
&+ 2b^{-1}\gamma_k(1 - b^{-1}\mu\gamma_k)\|\bar{x}_k - x^*\|\|G(\mathbf{x_k}) - \mathcal{G}(\bar{x}_k)\| \\
&+ \frac{b^{-2}\gamma_k^2 L^2}{m}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + b^{-2}\gamma_k^2 \left(\frac{1}{m} + b - 1\right)\nu^2 \\
&+ \frac{2L^2 b^{-2}\gamma_k^2 (b-1)}{m}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + 2(b-1)L^2 b^{-2}\gamma_k^2 \|\bar{x}_k - x^*\|^2.
\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right] \leq & \left((1-b^{-1}\mu\gamma_k)^2+2b^{-2}(b-1)\gamma_k^2L^2\right)\|\bar{x}_k-x^*\|^2 \\ & + \frac{2b^{-1}\gamma_k L(1-b^{-1}\mu\gamma_k)}{\sqrt{m}}\|\bar{x}_k-x^*\|\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\| \\ & + \frac{b^{-2}(2b-1)\gamma_k^2L^2}{m}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+b^{-2}\gamma_k^2\left(\frac{1}{m}+b-1\right)\nu^2.\end{aligned}$$

Note that we have

$$\begin{aligned}&\frac{2b^{-1}\gamma_k L(1-b^{-1}\mu\gamma_k)}{\sqrt{m}}\|\bar{x}_k-x^*\|\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\| \\ &= 2b^{-1}\gamma_k\left(\sqrt{\mu}(1-b^{-1}\mu\gamma_k)\|\bar{x}_k-x^*\|\right)\left(\frac{L}{\sqrt{\mu m}}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|\right) \\ &\leq b^{-1}\gamma_k\left(\mu(1-b^{-1}\mu\gamma_k)^2\|\bar{x}_k-x^*\|^2+\frac{L^2}{\mu m}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\right).\end{aligned}$$

From the preceding two relations, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right] \leq & \left((1-b^{-1}\mu\gamma_k)^2(1+\mu b^{-1}\gamma_k)+2b^{-2}(b-1)\gamma_k^2L^2\right)\|\bar{x}_k-x^*\|^2 \\ & + \frac{b^{-1}\gamma_k L^2}{m}\left(\frac{1}{\mu}+b^{-1}(2b-1)\gamma_k\right)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 \\ & + b^{-2}\gamma_k^2\left(\frac{1}{m}+b-1\right)\nu^2.\end{aligned}$$

From $\gamma_k \leq \frac{b\mu}{4(b-1)L^2}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2\right] \leq & (1-\frac{\mu b^{-1}\gamma_k}{2})\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right] \\ & + \frac{b^{-1}\gamma_k L^2}{m}\left(\frac{1}{\mu}+b^{-1}(2b-1)\gamma_k\right)\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\right] \\ & + b^{-2}\gamma_k^2\left(\frac{1}{m}+b-1\right)\nu^2.\end{aligned}$$

(b) From Equation 2.4.1 and invoking Lemma 2.4.4(b), we have

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2 &= \|\mathbf{W}\mathbf{x}_k-\gamma_k\mathbf{W}\mathbf{y}_k-\mathbf{1}(\bar{x}_k-\gamma_k\bar{y}_k)\|^2 \\ &= \|\mathbf{W}\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2-2\gamma_k\langle\mathbf{W}\mathbf{x}_k-\mathbf{1}\bar{x}_k,\mathbf{W}\mathbf{y}_k-\mathbf{1}\bar{y}_k\rangle \\ &\quad+\gamma_k^2\|\mathbf{W}\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2.\end{aligned}$$

By Invoking Lemma 2.4.2 and Lemma 2.4.4(c), we obtain

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2 &= \rho_W^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+2\gamma_k\|\mathbf{W}\mathbf{x}_k-\mathbf{1}\bar{x}_k\|\|\mathbf{W}\mathbf{y}_k-\mathbf{1}\bar{y}_k\|+\rho_W^2\gamma_k^2\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\\ &\le \rho_W^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+2\rho_W^2\gamma_k\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|+\rho_W^2\gamma_k^2\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\\ &\le \rho_W^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\rho_W^2\gamma_k\left(\frac{1-\rho_W^2}{2\gamma_k\rho_W^2}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\frac{2\gamma_k\rho_W^2}{1-\rho_W^2}\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\right)\\ &+\rho_W^2\gamma_k^2\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\\ &= \frac{1+\rho_W^2}{2}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\frac{\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2.\end{aligned}$$

We obtain

$$\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2\right]\le\frac{1+\rho_W^2}{2}\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\right]+\frac{\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\right].$$

(c) Next we obtain the third recursive relation. For the ease of presentation, we will use the following compact notation.

$$\begin{aligned}&\mathbf{G}_k\triangleq\mathbf{G}(\mathbf{x}_k),\quad \tilde{\mathbf{G}}_k\triangleq\mathbf{G}(\mathbf{x}_k,\boldsymbol{\xi}_k),\quad \tilde{\mathbf{G}}_k^e\triangleq\tilde{\mathbf{G}}_k-\mathbf{e}_k,\\ &\nabla_{i,k}\triangleq\nabla f_i(x_{i,k}),\quad \tilde{\nabla}_{i,k}\triangleq\nabla f_i(x_{i,k},\xi_{i,k}),\\ &\tilde{\nabla}_{i,k}^e\triangleq\nabla f_i(x_{i,k},\xi_{i,k})-e_{i,k}.\end{aligned}$$

Note that $\mathbb{E}\left[\tilde{\mathbf{G}}_k^e\mid\mathcal{F}_k\right]=\mathbf{G}_k,\quad \mathbb{E}\left[\tilde{\mathbf{G}}_{k+1}^e\mid\mathcal{F}_{k+1}\right]=\mathbf{G}_{k+1}$.
We can write, from Lemma 2.4.4(b),

$$\begin{aligned}\|e_{i,k}\|^2 &= \left\|\left(\nabla_{i,k}-bU_{\ell_k}\nabla_{i,k}^{\ell_k}\right)\right\|^2\\ &= \|\nabla_{i,k}\|^2+b^2\left\|\mathbf{U}_{\ell_k}\nabla_{i,k}^{\ell_k}\right\|^2-2b\left(\nabla_{i,k}\right)^T\mathbf{U}_{\ell_k}\nabla_{i,k}^{\ell_k}.\end{aligned}$$

Taking conditional expectations, we have

$$\mathbb{E}\left[\|e_{i,k}\|^2\mid\mathcal{F}_k\cup\{\xi_{i,k}\}\right]=\|\nabla_{i,k}\|^2+b\sum_{\ell=1}^{b}\left\|\mathbf{U}_\ell\nabla_{i,k}^{\ell}\right\|^2-2\left(\nabla_{i,k}\right)^T\sum_{\ell=1}^{b}\mathbf{U}_\ell\nabla_{i,k}^{\ell}.$$

We have

$$\sum_{\ell=1}^{b}\left\|\mathbf{U}_\ell\nabla_{i,k}^{\ell}\right\|^2=\sum_{\ell=1}^{b}\left\|\mathbf{U}_\ell\left(\nabla_{i,k}\right)^{(\ell)}\right\|^2\overset{(2.3.1)}{=}\|\nabla_{i,k}\|^2.$$

From the two preceding relations, we obtain

$$\mathbb{E}\left[\|e_{i,k}\|^2\mid\mathcal{F}_k\cup\{\xi_{i,k}\}\right]=(b-1)\|\nabla_{i,k}\|^2$$

From the update rules of the algorithm, we have

$$\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}_{k+1}\|^2 &\leq \|\mathbf{W}\mathbf{y}_k + b^{-1}\tilde{\mathbf{G}}^e_{k+1} - b^{-1}\tilde{\mathbf{G}}^e_k - \mathbf{1}\bar{y}_k + \mathbf{1}\bar{y}_k - \mathbf{1}\bar{y}_{k+1}\|^2 \\
&= \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + b^{-2}\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2 \\
&+ 2b^{-1}\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\rangle \\
&+ 2\langle \mathbf{y}_{k+1} - \mathbf{1}\bar{y}_k, \mathbf{1}(\bar{y}_k - \bar{y}_{k+1})\rangle + m\|\bar{y}_k - \bar{y}_{k+1}\|^2 \\
&\leq \rho_W^2\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + b^{-2}\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2 \\
&+ 2b^{-1}\langle \mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\rangle - m\|\bar{y}_k - \bar{y}_{k+1}\|^2 \\
&\leq (1 + b^{-1}\eta)\rho_W^2\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 + (\frac{1}{b^2} + \frac{1}{b\eta})\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2, \qquad (3.1.1)
\end{aligned}$$

where $\eta > 0$ is an arbitrary scalar. In the following, we present a few intermediary results that will be used to derive the third recursive inequality.
Claim 1: The following holds

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2 \mid \mathcal{F}_k\right] &\leq \mathbb{E}\left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] \\
&+ \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] + 2\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right]. \qquad (3.1.2)
\end{aligned}$$

Proof. We can write

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] \\
&+ 2\mathbb{E}\left[\langle \mathbf{G}_{k+1}, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&- 2\mathbb{E}\left[\langle \mathbf{G}_k, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&+ \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right]. \qquad (3.1.3)
\end{aligned}$$

Note that since $\mathbf{x}_{k+1}$ is characterized in terms of $\boldsymbol{\xi}_k$, using Assumption 2.1.2, Assumption 2.4.1, and Lemma 2.4.1 we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_k\right] &= \mathbb{E}_{\boldsymbol{\xi}_k, \boldsymbol{\ell}_k}\left[\mathbb{E}_{\boldsymbol{\xi}_{k+1}, \boldsymbol{\ell}_{k+1}}\left[\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}_k, \boldsymbol{\ell}_k}\left[\mathbb{E}_{\boldsymbol{\xi}_{k+1}, \boldsymbol{\ell}_{k+1}}\left[\tilde{\mathbf{G}}_{k+1} - \mathbf{e}_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}\right]\right] = 0.
\end{aligned}$$

We also have $\mathbb{E}\left[\tilde{\mathbf{G}}^e_k - \mathbf{G}_k \mid \mathcal{F}_k\right] = 0$. Thus, we obtain

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{G}_k, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] &= \langle \mathbf{G}_k, \mathbb{E}\left[\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k \mid \mathcal{F}_k\right]\rangle \\
&= \langle \mathbf{G}_k, 0\rangle = 0.
\end{aligned}$$

We can also write

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{G}_{k+1}, \tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[\langle \mathbf{G}_{k+1}, \mathbb{E}_{\boldsymbol{\xi}_{k+1},\boldsymbol{\ell}_{k+1}}\left[\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}\right]\right\rangle + \mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[\langle \mathbf{G}_{k+1}, \mathbb{E}_{\boldsymbol{\xi}_{k+1},\boldsymbol{\ell}_{k+1}}\left[\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}\right]\right\rangle + \mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right].
\end{aligned}$$

From the preceding relations, we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k\|^2 \mid \mathcal{F}_k\right] &\leq \mathbb{E}\left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] + 2\mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&\quad + \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \tilde{\mathbf{G}}^e_k - \mathbf{G}_{k+1} + \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] \\
&\leq \mathbb{E}\left[\|\mathbf{G}_{k+1} - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] + 2\mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \\
&\quad + \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] \\
&\quad + 2\mathbb{E}\left[\langle \tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}, \tilde{\mathbf{G}}^e_k - \mathbf{G}_k\rangle \mid \mathcal{F}_k\right].
\end{aligned}$$

Note that we have

$$\begin{aligned}
\mathbb{E}\left[\langle \tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}, \tilde{\mathbf{G}}^e_k - \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] &= \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[\langle \mathbb{E}_{\boldsymbol{\xi}_{k+1},\boldsymbol{\ell}_{k+1}}\left[\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1} \mid \mathcal{F}_{k+1}\right], \tilde{\mathbf{G}}^e_k - \mathbf{G}_k\rangle\right] \\
&= 0.
\end{aligned}$$

Also, we have

$$2\mathbb{E}\left[\langle \mathbf{G}_{k+1}, -\tilde{\mathbf{G}}^e_k + \mathbf{G}_k\rangle \mid \mathcal{F}_k\right] \leq \mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right].$$

From the last three relations, we obtain Claim 1. ■

Claim 2: The following relations hold.

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] \leq mb\nu^2 + (b-1)\|\mathbf{G}_k\|^2, \\
&\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1} - \mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] \leq mb\nu^2 + (b-1)\mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right].
\end{aligned} \tag{3.1.4}$$

Proof. From Assumption 2.4.1 and Lemma 2.4.1, we have

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|\tilde{\mathbf{G}}_k - \mathbf{e}_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] \\
&= \mathbb{E}\left[\|\tilde{\mathbf{G}}_k - \mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\|\mathbf{e}_k\|^2 \mid \mathcal{F}_k\right] \\
&= m\nu^2 + \sum_{i=1}^m \mathbb{E}\left[\|e_{i,k}\|^2 \mid \mathcal{F}_k\right] \\
&= m\nu^2 + \sum_{i=1}^m (b-1)\left(\nu^2 + \|\nabla_{i,k}\|^2\right) \\
&= mb\nu^2 + (b-1)\|\mathbf{G}_k\|^2.
\end{aligned}$$

Using this relation, we can also write

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\mathbf{G}}^e_{k+1}-\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[\mathbb{E}_{\boldsymbol{\xi}_{k+1},\boldsymbol{\ell}_{k+1}}\left[\|\tilde{\mathbf{G}}^e_{k+1}-\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_{k+1}\right]\right]\\ &\leq \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[mb\nu^2+(b-1)\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right]\\ &= mb\nu^2+(b-1)\mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right].\end{aligned}$$

■

Claim 3: The following inequality holds.

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{G}_{k+1}-\mathbf{G}_k\|^2 \mid \mathcal{F}_k\right] &\leq 2L^2\left(b^{-2}L^2\gamma_k^2+\|\mathbf{W}-\mathbf{I}\|^2\right)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\\ &+2L^2\rho_W^2\gamma_k^2\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2 \mid \mathcal{F}_k\right]\\ &+2b^{-2}mL^4\gamma_k^2\|\bar{x}_k-x^*\|^2+L^2\gamma_k^2b^{-2}(b-1)\|\mathbf{G}_k\|^2\\ &+mL^2b^{-2}(\frac{1}{m}+b-1)\nu^2\gamma_k^2. \qquad (3.1.5)\end{aligned}$$

Proof. From the Lipschitzian property of the local objectives we have $\|\mathbf{G}_{k+1}-\mathbf{G}_k\|^2 \leq L^2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2$. Next, we estimate the term $\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2$. We have

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 &= \|\mathbf{W}\mathbf{x}_k-\gamma_k\mathbf{W}\mathbf{y}_k-\mathbf{x}_k\|^2\\ &= \|(\mathbf{W}-\mathbf{I})(\mathbf{x}_k-\mathbf{1}\bar{x}_k)-\gamma_k\mathbf{W}\mathbf{y}_k-\mathbf{x}_k\|^2\\ &\leq \|\mathbf{W}-\mathbf{I}\|^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\gamma_k^2\|\mathbf{W}\mathbf{y}_k\|^2-2\gamma_k\langle(\mathbf{W}-\mathbf{I})(\mathbf{x}_k-\mathbf{1}\bar{x}_k),\mathbf{W}\mathbf{y}_k\rangle\\ &= \|\mathbf{W}-\mathbf{I}\|^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\gamma_k^2\|\mathbf{W}\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2+m\gamma_k^2\|\bar{y}_k\|^2\\ &-2\gamma_k\langle(\mathbf{W}-\mathbf{I})(\mathbf{x}_k-\mathbf{1}\bar{x}_k),\mathbf{W}\mathbf{y}_k-\mathbf{1}\bar{y}_k\rangle\\ &\leq \|\mathbf{W}-\mathbf{I}\|^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+\rho_W^2\gamma_k^2\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2+m\gamma_k^2\|\bar{y}_k\|^2\\ &+2\rho_W\gamma_k\|\mathbf{W}-\mathbf{I}\|\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|\\ &\leq 2\|\mathbf{W}-\mathbf{I}\|^2\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+2\rho_W^2\gamma_k^2\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2+m\gamma_k^2b^{-2}\|b\bar{y}_k\|^2.\end{aligned}$$

From Lemma 2.5.1(b) and (c) we have

$$\begin{aligned}\mathbb{E}\left[\|b\bar{y}_k\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\|b\bar{y}_k-G_k\|^2 \mid \mathcal{F}_k\right]+\|G_k\|^2\\ &\leq \left(\frac{1}{m}+b-1\right)\nu^2+\frac{b-1}{m}\|\mathbf{G}_k\|^2+\|G_k\|^2.\end{aligned}$$

Also, from Lemma 2.5.1(e) and (f) we have

$$\|G_k\|^2 \leq 2\|G_k-\mathcal{G}(\bar{x}_k)\|^2+2\|\mathcal{G}(\bar{x}_k)\|^2 \leq \frac{2L^2}{m}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2+2L^2\|\bar{x}_k-x^*\|^2.$$

From the preceding relations, we obtain Claim 3. ■

Claim 4: We have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] &\leq 2L^2m\|\bar{x}_k-x^*\|^2+\frac{2L^2\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2 \mid \mathcal{F}_k\right]\\ &+2L^2\left(1+b^{-1}\gamma_kL^2(\frac{1}{\mu}+b^{-1}(2b-1)\gamma_k)\right)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\\ &+2mL^2b^{-2}\nu^2\left(\frac{1}{m}+b-1\right)\gamma_k^2.\end{aligned}$$

Proof. From Lemma 2.5.1(g) we can write

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}_{\boldsymbol{\xi}_k,\boldsymbol{\ell}_k}\left[\mathbb{E}_{\boldsymbol{\xi}_{k+1},\boldsymbol{\ell}_{k+1}}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_{k+1}\right]\right] \\ &\leq 2L^2\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2 \mid \mathcal{F}_k\right] + 2mL^2\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right].\end{aligned}$$

Substituting $\mathbb{E}\left[\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2 \mid \mathcal{F}_k\right]$ and $\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right]$ from the first two recursive bounds, we can conclude Claim 4. Note that from the first two recursions we have

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^2 \mid \mathcal{F}_k\right] \leq &\left(1-\frac{\mu b^{-1}\gamma_k}{2}\right)\|\bar{x}_k-x^*\|^2 \\ &+\frac{b^{-1}\gamma_k L^2}{m}\left(\frac{1}{\mu}+b^{-1}(2b-1)\gamma_k\right)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 \\ &+b^{-2}\gamma_k^2\left(\frac{1}{m}+b-1\right)\nu^2.\end{aligned}$$

And also

$$\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|^2 \leq \frac{1+\rho_W^2}{2}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 + \frac{\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2.$$

From the preceding relations, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{G}_{k+1}\|^2 \mid \mathcal{F}_k\right] &\leq 2L^2\left(\frac{1+\rho_W^2}{2}\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 + \frac{\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2 \mid \mathcal{F}_k\right]\right) \\ &\quad + 2L^2m\left((1-\frac{\mu b^{-1}\gamma_k}{2})\|\bar{x}_k-x^*\|^2\right. \\ &\quad \left.+\frac{b^{-1}\gamma_k L^2}{\mu m}(1+b^{-1}\mu\gamma_k)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 + b^{-2}\gamma_k^2\frac{2(2b-1)\nu^2}{m}\right) \\ &\leq 2L^2m\|\bar{x}_k-x^*\|^2 + \frac{2L^2\gamma_k^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2 \mid \mathcal{F}_k\right] \\ &\quad + 2L^2\left(1+\mu^{-1}b^{-1}\gamma_k L^2(1+b^{-1}\mu\gamma_k)\right)\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2 + 4L^2b^{-2}\gamma_k^2(2b-1)\nu^2.\end{aligned}$$

■

By combining (3.1.1), Claims 1 to 4, and Lemma 2.5.1(g), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}_{k+1}-\mathbf{1}\bar{y}_{k+1}\|^2\right] \leq & \left((1+b^{-1}\eta)\rho_W^2+\gamma_k^2\left(\frac{1}{b^2}+\frac{1}{b\eta}\right)\left(2L^2\rho_W^2\right.\right.\\ & \left.\left.+\frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)\right)\mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\right]\\ & +2L^2m\left(\frac{1}{b^2}+\frac{1}{b\eta}\right)\left(b^{-2}L^2\gamma_k^2+(b-1)\left(3+L^2\gamma_k^2b^{-2}\right)\right)\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]\\ & +2L^2\left(b^{-2}L^2\gamma_k^2+(b-1)\left(3+L^2\gamma_k^2b^{-2}\right.\right.\\ & \left.\left.+b^{-1}\gamma_kL^2\left(\frac{1}{\mu}+b^{-1}(2b-1)\gamma_k\right)\right)\right.\\ & \left.+\|\mathbf{W}-\mathbf{I}\|^2\right)\left(\frac{1}{b^2}+\frac{1}{b\eta}\right)\mathbb{E}\left[\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\right]\\ & +\left(\frac{1}{b^2}+\frac{1}{b\eta}\right)\nu^2\left(mL^2b^{-2}\left(\frac{1}{m}+b-1\right)\gamma_k^2+3mb\right).\end{aligned}$$

■

3.2 Rate Analysis

We proceed with deriving the rate statements of Algorithm 2.

Theorem 3.2.1 (Rate statements) *Consider Algorithm 2. Let Assumptions 2.1.1, 2.1.2, 2.1.3, 2.1.4, and 2.4.1 hold. Let us define* $\mathbb{e}_{1,k} \triangleq \mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]$, $\mathbb{e}_{2,k} \triangleq \mathbb{E}\left[\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|^2\right]$, *and* $\mathbb{e}_{3,k} \triangleq \mathbb{E}\left[\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|^2\right]$ *for* $k \geq 0$. *Suppose* $\gamma_k := \frac{\gamma}{k+\Gamma}b$ *with* $\gamma > 0$, $\Gamma > \gamma$,

$$\begin{aligned}&\Gamma \geq \gamma\sqrt{\frac{3}{1-\rho_W^2}\left(\frac{1}{b^2}+\frac{1}{b\eta}\right)\left(2L^2\rho_W^2+\frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)},\\ &\text{and } \Gamma \geq \gamma\left(\min\left\{\frac{2b}{\mu+L},\frac{b\mu}{4(b-1)L^2}\right\}\right)^{-1}.\end{aligned} \tag{3.2.1}$$

(a) Then, there exist positive scalars $\theta_t > 0$ *for* $t = 1,\ldots,9$ *with* $\theta_4 < 1$ *and* $\theta_6 < 1$ *such that for all* $k \geq 0$ *we have*

$$\begin{aligned}\mathbb{e}_{1,k+1} &\leq (1-\theta_1\gamma_k)\mathbb{e}_{1,k}+\theta_2\gamma_k\mathbb{e}_{2,k}+\theta_3\gamma_k^2,\\ \mathbb{e}_{2,k+1} &\leq (1-\theta_4)\mathbb{e}_{2,k}+\theta_5\gamma_k^2\mathbb{e}_{3,k},\\ \mathbb{e}_{3,k+1} &\leq (1-\theta_6)\mathbb{e}_{3,k}+\theta_7\mathbb{e}_{1,k}+\theta_8\mathbb{e}_{2,k}+\theta_9.\end{aligned}$$

(b) Let $\gamma > \frac{1}{\theta_1}$. *Let us define* $\hat{\mathbb{e}}_1 := \Gamma\mathbb{e}_{1,0}\mathbb{n}_1$, $\hat{\mathbb{e}}_2 := \Gamma\mathbb{e}_{2,0}\mathbb{n}_2$, *and* $\hat{\mathbb{e}}_3 := \max\left\{\Gamma\frac{3\theta_9}{\theta_6}\mathbb{n}_3,\mathbb{e}_{3,0}\right\}$, *where* $\mathbb{n}_1,\mathbb{n}_2,\mathbb{n}_3 > 0$ *are given as* $\mathbb{n}_1 := \frac{2C_2}{C_3\Gamma-2C_1C_4C_5}$, $\mathbb{n}_2 := \frac{2C_4C_5}{\Gamma}\mathbb{n}_1$, *and* $\mathbb{n}_3 := \frac{C_5}{\Gamma}\mathbb{n}_1$, *where* $C_1 \triangleq \gamma\theta_2\mathbb{e}_{2,0}$, $C_2 \triangleq \theta_3\gamma^2$, $C_3 \triangleq (\gamma\theta_1-1)\,\mathbb{e}_{1,0}$, $C_4 \triangleq \frac{6\theta_9\theta_5\gamma^2}{\mathbb{e}_{2,0}\theta_4\theta_6}$, $C_5 \triangleq \frac{\theta_7\mathbb{e}_{1,0}}{\theta_9}$, *and* $C_6 \triangleq \frac{\theta_8\mathbb{e}_{2,0}}{\theta_9}$.

Then, if $\Gamma > \max\left\{\sqrt{2C_4C_6}, \frac{2C_1C_4C_5}{C_3}, \frac{4}{\theta_4} - 1\right\}$, *and for all*

$$\eta < b\left(\frac{1-\rho_W^2}{\rho_W^2}\right), \text{ and } k > \gamma\sqrt{\frac{\left(\frac{1}{b^2} + \frac{2\rho_W^2}{b^2(1-\rho_W^2)}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)}{(1+\rho_W^2)/2}} - \Gamma, \quad (3.2.2)$$

$$\boxed{\mathfrak{e}_{1,k} \leq \frac{\hat{\mathfrak{e}}_1}{k+\Gamma}, \qquad \mathfrak{e}_{2,k} \leq \frac{\hat{\mathfrak{e}}_2}{(k+\Gamma)^2}, \qquad \mathfrak{e}_{3,k} \leq \hat{\mathfrak{e}}_3.} \quad (3.2.3)$$

Proof. (a) Consider Proposition 3.1.1. It suffices to show that $0 < 1 - \theta_6 < 1$. From Proposition 3.1.1 (c), we have

$$1 - \theta_6 = (1 + b^{-1}\eta)\rho_W^2 + \gamma_k^2\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right).$$

It can be observed that $0 < 1 - \theta_6$. Remaining is to show $1 - \theta_6 < 1$, that is, we need to show

$$(1 + b^{-1}\eta)\rho_W^2 + \gamma_k^2\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right) < 1. \quad (3.2.4)$$

From the condition on k in equation (3.2.2), we have

$$\left(\frac{k+\Gamma}{\gamma}\right)^2 > \frac{\left(\frac{1}{b^2} + \frac{2\rho_W^2}{b^2(1-\rho_W^2)}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)}{(1+\rho_W^2)/2}.$$

This can also be written as

$$\left(\frac{\gamma}{k+\Gamma}\right)^2 < \frac{(1+\rho_W^2)/2}{\left(\frac{1}{b^2} + \frac{2\rho_W^2}{b^2(1-\rho_W^2)}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right)},$$

Substituting from the condition on η in equation (3.2.2), substituting $\eta = \frac{b}{2}\left(\frac{1-\rho_W^2}{\rho_W^2}\right)$ and from the definition of sequence $\gamma_k = \frac{\gamma}{k+\Gamma}$, we have

$$\gamma_k^2\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right) < \frac{(1+\rho_W^2)}{2} = 1 - \frac{(1-\rho_W^2)}{2} = 1 - \frac{\eta\rho_W^2}{b}$$

Therefore, we have

$$\frac{\eta\rho_W^2}{b} + \gamma_k^2\left(\frac{1}{b^2} + \frac{1}{b\eta}\right)\left(2L^2\rho_W^2 + \frac{2(b-1)L^2(1+\rho_W^2)\rho_W^2}{1-\rho_W^2}\right) < 1.$$

(b) We select θ_3 arbitrarily large such that $\mathbb{m}_1 \geq 1$. Consecutively, we can state that $\mathbb{m}_2 \geq 1$ and $\mathbb{m}_3 \geq 1$. We have

$$\mathfrak{e}_{1,0} \leq \mathfrak{e}_{1,0}\mathbb{m}_1 \leq \frac{\hat{\mathfrak{e}}_1}{\Gamma} \leq \frac{\hat{\mathfrak{e}}_1}{0+\Gamma}.$$

The first inequality in (3.2.3) holds true for $k = 0$. Now from the definition of $\tilde{\mathbb{e}}_2$, we have

$$\mathbb{e}_{2,0} \leq \mathbb{e}_{2,0}\mathbb{n}_2 \leq \frac{\hat{\mathbb{e}}_2}{\Gamma^2} \leq \frac{\hat{\mathbb{e}}_2}{(0+\Gamma)^2}.$$

This implies the second inequality in (3.2.3) holds true for $k = 0$. Further, for $\mathbb{e}_{3,0}$, we have $\mathbb{e}_{3,0} \leq \hat{\mathbb{e}}_3$. Therefore, the third inequality in (3.2.3) holds true for $k = 0$. Now let the induction hypothesis holds true for some $k \geq 0$. From the definition of n_1, we have $C_2 \leq (C_3\Gamma - 2C_1C_4C_5)\,\mathbb{n}_1$. Therefore, we have $C_1\mathbb{n}_2\Gamma + C_2 \leq C_3\mathbb{n}_1\Gamma$. Next, substituting the values of C_1, C_2, and C_3, in the above, we have

$$\frac{\hat{\mathbb{e}}_1}{(k+\Gamma)^2} \leq \frac{\gamma\theta_1\hat{\mathbb{e}}_1}{(k+\Gamma)^2} - \frac{\gamma\theta_2\hat{\mathbb{e}}_2}{(k+\Gamma)^2} - \frac{\theta_3\gamma^2}{(k+\Gamma)^2}.$$

We have $\gamma \geq \gamma_k$. Substituting in the above

$$\frac{\hat{\mathbb{e}}_1}{(k+\Gamma)^2} \leq \frac{\gamma\theta_1\hat{\mathbb{e}}_1}{(k+\Gamma)^2} - \frac{\gamma_k\theta_2\hat{\mathbb{e}}_2}{(k+\Gamma)^2} - \frac{\theta_3\gamma^2}{(k+\Gamma)^2}.$$

Further, by bounding $\frac{\hat{e}_1}{(k+\Gamma)(k+\Gamma+1)} \leq \frac{\hat{e}_1}{(k+\Gamma)^2}$, we obtain

$$\frac{\hat{\mathbb{e}}_1}{(k+\Gamma)} - \frac{\hat{\mathbb{e}}_1}{(k+\Gamma+1)} \leq \frac{\gamma\theta_1\hat{\mathbb{e}}_1}{(k+\Gamma)^2} - \frac{\gamma_k\theta_2\hat{\mathbb{e}}_2}{(k+\Gamma)^2} - \frac{\theta_3\gamma^2}{(k+\Gamma)^2}.$$

By induction hypothesis, and Theorem3.2.1(a), the first inequality of (3.2.3) holds for $k+1$. Next, from the definition of n_2, we have

$$\mathbb{n}_2 \geq \frac{C_4C_5}{\Gamma}\mathbb{n}_1 \geq C_4\mathbb{n}_3.$$

Substituting for C_4 and rearranging the terms, we obtain

$$\hat{\mathbb{e}}_3 \leq \frac{\theta_4}{2\theta_5\gamma^2}\,\hat{\mathbb{e}}_2 \leq \frac{1}{2\theta_5\gamma^2}\left(\theta_4 - \frac{\theta_4}{2}\right)\,\hat{\mathbb{e}}_2.$$

From $\frac{2\Gamma+1}{(\Gamma+1)^2} \leq \frac{\theta_4}{2}$, the preceding inequality becomes

$$\frac{2\Gamma+1}{(\Gamma+1)^2}\hat{\mathbb{e}}_2 \leq \theta_4\hat{\mathbb{e}}_2 - 2\theta_5\gamma^2\hat{\mathbb{e}}_3.$$

Further, from $\frac{2k+2\Gamma+1}{(k+\Gamma+1)^2} \leq \frac{2\Gamma+1}{(\Gamma+1)^2}$, we have

$$\frac{\hat{\mathbb{e}}_2}{(\Gamma+1)^2} - \frac{\hat{\mathbb{e}}_2}{(k+\Gamma+1)^2} \leq \frac{\theta_4\hat{\mathbb{e}}_2}{(\Gamma+1)^2} - \frac{2\theta_5\gamma^2\hat{\mathbb{e}}_3}{(\Gamma+1)^2}.$$

By induction hypothesis and Theorem 3.2.1 (a), we show the second inequality of (3.2.3) holds for $k+1$.

Next, from the definition of n_3, we have $\mathbb{n}_3\Gamma \geq C_5\mathbb{n}_1$. Substituting value of C_5 and rearranging the terms, we obtain

$$\frac{3\theta_7}{\Gamma}\hat{\mathbb{e}}_1 \leq \theta_6\hat{\mathbb{e}}_3.$$

Now, from the upper bound of Γ, we have $\Gamma^2 \geq 2C_4C_6$. From this, we have

$$\Gamma^2\left(\frac{C_5}{\Gamma}\mathbb{n}_1\right) \geq 2C_4C_6\left(\frac{C_5}{\Gamma}\mathbb{n}_1\right) \geq C_6\left(\frac{2C_4C_5}{\Gamma}\mathbb{n}_1\right).$$

Substituting for C_6 and rearranging terms, we have

$$\frac{3\theta_8}{\Gamma^2}\hat{\mathbb{e}}_2 \leq \theta_6\hat{\mathbb{e}}_3.$$

From the preceding two results and $3\theta_9 \leq \theta_6\hat{\mathbb{e}}_3$, we have

$$(1-\theta_6)\,\hat{\mathbb{e}}_3 + \frac{\theta_7}{k+\Gamma}\hat{\mathbb{e}}_1 + \frac{\theta_8}{(k+\Gamma)^2}\hat{\mathbb{e}}_2 + \theta_9 \leq \hat{\mathbb{e}}_3.$$

By induction hypothesis and Theorem 3.2.1 (a), we conclude that the third inequality of (3.2.3) holds for $k+1$. ■

3.3 Concluding Remarks

We derive the error bounds for all three error metrics that are $\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right]$, $\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2\right]$, and $\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\right]$. We provide non-asymptotic convergence rates of the order $\mathcal{O}\left(1/k\right)$ for an optimality metric, and $\mathcal{O}\left(1/k^2\right)$ for a consensus violation metric. The observed rates by the proposed DRBSGT algorithm match with those of the DSGT algorithm [15].

CHAPTER IV

NUMERICAL EXPERIMENTS

In this chapter, we provide numerical experiments to validate our theoretical results. We consider the logistic loss regression problem for the numerical analysis of the proposed scheme and discuss the other schemes for comparison with DRBSGT. In Section 4.1, we provide the setup for the numerical experiment, which includes the tuning and data set parameters. In Section 4.2, we observe the insights of the results of the numerical experiment and thesis. The Section 4.3 concludes the thesis.

For the experiments, we consider the regularized logistic regression loss minimization problem presented in Section 1.1. Consider a data set denoted by $\mathcal{D} \triangleq \{(u_j, v_j) \in \mathbb{R}^n \times \{-1, +1\} \mid j \in \mathcal{S}\}$ where $\mathcal{S} \triangleq \{1, \ldots, s\}$ denotes the index set. Let $\mathcal{S}_i$ denote the index set of the data locally known by agent i where $\cup_{i=1}^{m} \mathcal{S}_i = \mathcal{S}$. The problem can be formulated as $\min \sum_{i=1}^{m} f_i(x)$ where we define local functions f_i as

$$f_i(x) \triangleq \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}_i} \ln\left(1 + \exp\left(-v_j u_j^T x\right)\right) + \frac{\mu}{2m}\|x\|^2, \tag{4.0.1}$$

where $u_j \in \mathbb{R}^n$ and $v_j \in \{-1, 1\}$ for $j \in \mathcal{S}_i$ which denotes the binary value of the j^{th} data label.

We simulate the proposed distributed randomized block stochastic gradient tracking method (DRBSGT) algorithm on a network consisting of m agents. We provide a comparison of suboptimality and consensus metrics of DRBSGT with those of two existing methods namely, distributed stochastic gradient tracking (DSGT) [15] and adapt then combine (ATC), a variant of block distributed Successive cONvex Approximation algorithm over Time-varying digrAphs (block SONATA) in convex regimes [14].

4.1 Simulation

We perform simulations on two data sets with m agents. We use the complete and the ring graph structure to represent the communication among the agents. We implement the simulations on MNIST and Synthetic data set for $m = 5$ and $m = 5, 10$, respectively. The MNIST data set consists of $70,000$ labels and 784 attributes, whereas the Synthetic data set has $10,000$ labels and $10,000$ attributes with a Gaussian distribution with mean as 5 and standard deviation as 0.5. We consider different parameters for different data sets mentioned in Table 2. Further, we use $\gamma = 1e+1$, $\Gamma = 1e+4$, $\mu = 1e-1$, and the batch size for computing gradient from each agent $\epsilon = 1e+2$ for both data sets. Taking into account the stochasticity involved in DRBSGT and DSGT schemes, we have obtained different sample paths in our implementations. In Figure 2 we have compared the performance with respect

Table 2: Parameters of various settings used for implementation

Schemes \ Data sets	MNIST ($m = 5$)	Synthetic ($m = 5$)	Synthetic ($m = 10$)
DRBSGT	$n = 784$ $\|\mathcal{S}\| = 5e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$ $b = 14$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$ $b = 100$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$ $b = 100$
DSGT	$n = 784$ $\|\mathcal{S}\| = 5e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $\epsilon = 1e{+}2$
ATC	$n = 784$ $\|\mathcal{S}\| = 5e + 4$ $\mu = 1e{-}1$ $b = 14$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $b = 100$	$n = 1e + 4$ $\|\mathcal{S}\| = 1e + 4$ $\mu = 1e{-}1$ $b = 100$

Table 3: Objective function value comparision of Algorithm 2 vs. DSGT vs. ATC for 90% CIs

Schemes \ Data sets	MNIST ($m = 5$)	Synthetic ($m = 5$)	Synthetic ($m = 10$)
DRBSGT (Complete)	$[1.047e{+}1, 1.057e{+}1]$	$[9.282e{+}0, 9.315e{+}0]$	$[5.071e{+}0, 5.151e{+}0]$
DRBSGT (Ring)	$[1.047e{+}1, 1.056e{+}1]$	$[9.282e{+}0, 9.314e{+}0]$	$[5.105e{+}0, 5.117e{+}0]$
DSGT (Complete)	$[4.974e{+}0, 5.471e{+}0]$	$[6.508e{+}1, 14.491e{+}1]$	$[14.464e{+}1, 31.159e{+}1]$
DSGT (Ring)	$[5.057e{+}0, 5.222e{+}0]$	$[4.622e{+}1, 9.889e{+}1]$	$[2.218e{+}1, 6.011e{+}1]$
ATC (Complete)	$8.078e{+}3$	$2.489e{+}4$	$2.509e{+}4$
ATC (Ring)	$8.076e{+}3$	$2.489e{+}4$	$2.509e{+}4$

to the number of local gradient evaluations. These local gradient evaluations are the number of total samples used in each gradient step. The highlighted areas in the plots in Figure 2 represent the confidence intervals. We provide 90% confidence intervals on the errors of the sample paths for each setting in Table 3. We choose the total sample paths for MNIST and Synthetic data sets as 5 and 10, respectively.

4.2 Insights

In Figure 2, we notice that the proposed DRBSGT scheme converges for both MNIST and Synthetic data sets, considering both the suboptimality and consensus metrics. We observe that with the comparison to DSGT and ATC, the performance of DRBSGT ameliorates. From Table 3 and Figure 2, we notice that when the number of attributes and the number of agents m increases, the performance improves for the proposed algorithm. The second row in Figure 2 provides significant evidence that the consensus errors are bounded and reducing. We did not observe any significant difference in the performance based on the network connectivity structure.

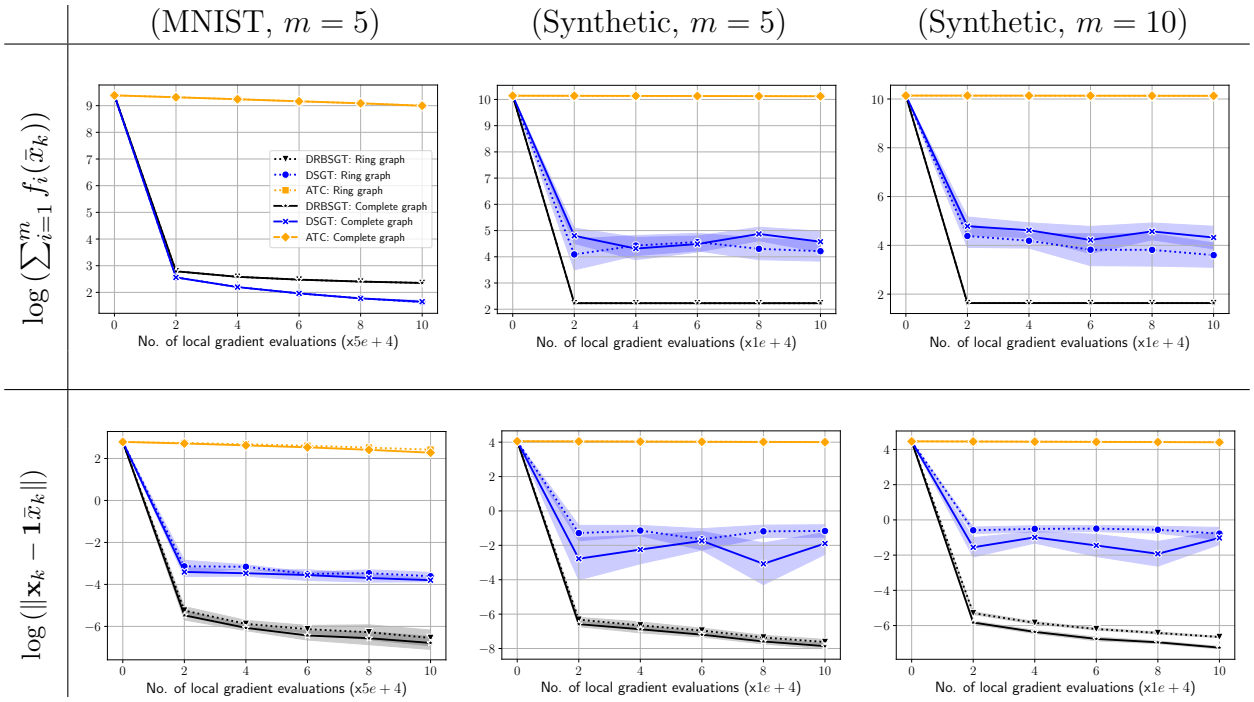

Figure 2: Algorithm 2 vs. DSGT vs. ATC in terms of objective function value and consensus error

4.3 Concluding Remarks

We validate that the theoretical claims hold after the numerical results. We compare the performance of our scheme with DSGT [15] and ATC [14] under different network assumptions on logistic loss regression minimization problems. The data sets we consider for the implementation are the Modified National Institute of Standards and Technology (MNIST) and synthetic data sets.

CHAPTER V

CONCLUSION AND FUTURE DIRECTION

In this thesis, we consider the a class of distributed stochastic optimization problems over undirected networks. Motivated by big data applications, we address this problem considering a possibility of large-dimensionality of the solution space where the computation of the local gradient mappings may become expensive. The main contributions of the thesis are as follows:

1. We develop an algorithm called distributed randomized block stochastic gradient tracking (DRBSGT) for addressing distributed stochastic optimization problems of the form (2.1.1) with possibly a large dimension in the solution space.

2. We obtain the rate of $\mathcal{O}(1/k)$ on a suboptimality and $\mathcal{O}(1/k^2)$ on a consensus violation metric for the DRBSGT algorithm, these rate statements are comparable with those of DSGT.

3. We validate the theoretical claims by performing some numerical experiments. We compare the performance of our scheme with that of DSGT [15] and ATC [14] under different network assumptions on logistic loss regression minimization problems. The data sets we consider for the implementation are the Modified National Institute of Standards and Technology (MNIST) and synthetic data sets.

In this thesis, we consider synchronous communications among the agents, i.e., at every iteration k, all the agents communicate with their neighbors synchronously and seek to minimize the average cost function. For a more practical perspective of addressing this problem, we plan to relax the assumption of synchronous communications.
As a future direction to this thesis research, we plan on considering asynchronous communications among agents. In particular:

1. We plan to develop a distributed randomized block gossip-like stochastic gradient tracking (asyn-DRBSGT) algorithm for addressing distributed stochastic optimization problems of the form (2.1.1) with possibly a large dimension in the solution space and asynchronous communication among the agents.

2. We plan to obtain reasonable rate statements for both suboptimality and consensus violation metrics for asyn-DRBSGT.

REFERENCES

[1] C. D. Dang and G. Lan, *Stochastic block mirror descent methods for nonsmooth and stochastic optimization*, SIAM Journal on Optimization **25** (2015), no. 2, 856–881.

[2] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, Advances in Neural Information Processing Systems (2014), 1646–1654.

[3] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Mathematical Programming **176** (2019), 497–544.

[4] H. D. Kaushik and F. Yousefian, *A method with convergence rates for optimization problems with variational inequality constraints*, SIAM Journal on Optimization **31** (2021), no. 3, 2171–2198.

[5] H. D. Kaushik and F. Yousefian, *A randomized block coordinate iterative regularized subgradient method for high-dimensional ill-posed convex optimization*, Proceedings of the American Control Conference, IEEE, July 2019, pp. 3420–3425.

[6] B. Li, S. Cen, Y. Chen, and Yu Chi, *Communication-efficient distributed optimization in networks with gradient tracking and variance reduction*, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (2020), 1662–1672.

[7] H. Li, L. Zheng, Z. Wang, Y. Yan, L. Feng, and J. Guo, *S-DIGing: A stochastic gradient tracking algorithm for distributed optimization*, IEEE Transactions on Emerging Topics in Computational Intelligence (2020), 1–30.

[8] P. Lorenzo and G. Scutari, *Next: In-network nonconvex optimization*, IEEE Transactions on Signal and Information Processing Over Networks **2** (2016), no. 2, 120–136.

[9] P. Di Lorenzo and G. Scutari, *Distributed nonconvex optimization over time-varying networks*, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2016), 4124–4128.

[10] S. Lu, X. Zhang, H. Sun, and M. Hong, *GNSD: a gradient-tracking based nonconvex stochastic algorithm for decentralized optimization*, 2019 IEEE Data Science Workshop (DSW) (2019), 315–321.

[11] N. Majlesinasab, F. Yousefian, and A. Pourhabib, *Self-tuned mirror descent schemes for smooth and nonsmooth high-dimensional stochastic optimization*, IEEE Transactions on Automatic Control **64** (2019), no. 10, 4377–4384.

[12] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization **27** (2017), no. 4, 2597–2633.

[13] Yu. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization **22** (2012), no. 2, 341–362.

[14] I. Notarnicola, Y. Sun, G. Scutari, and G. Notarstefano, *Distributed big-data optimization via block-wise gradient tracking*, IEEE Transactions on Automatic Control **66** (2021), no. 5, 2045–2060.

[15] S. Pu and A. Nedić, *Distributed stochastic gradient tracking methods*, Mathematical Programming **187** (2021), 409 – 457.

[16] S. Pu, W. Shi, J. Xu, and A. Nedić, *Push-pull gradient methods for distributed optimization in networks*, IEEE Transactions on Automatic Control **66** (2021), no. 1, 1–14.

[17] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Transactions on Control of Network Systems **5** (2017), no. 3, 1245–1260.

[18] P. Ricktárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming **144** (2014), 1–38.

[19] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Mathematical Programming **176** (2019), 497–544.

[20] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research **14** (2013), no. 1, 567–599.

[21] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, *Distributed stochastic optimization with gradient tracking over strongly-connected networks*, Proceedings of IEEE 58$^{\text{th}}$ Conference on Decision and Control (CDC) (2019), 8353–8358.

[22] F. Yousefian, *Bilevel distributed optimization in directed networks*, Proceedings of the American Control Conference, IEEE, 2021, pp. 2230–2235.

[23] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On stochastic mirror-prox algorithms for stochastic Cartesian variational inequalities: Randomized block coordinate and optimal averaging schemes*, Set-Valued and Variational Analysis **26** (2018), no. 4, 789–819.

[24] F. Yousefian, J. Yevale, and H. D. Kaushik, *Distributed randomized block stochastic gradient tracking method*, Proceedings of the American Control Conference(submitted), 2022, `https://arxiv.org/abs/2110.06575`.

VITA

Jayesh Vinayak Yevale

Candidate for the Degree of

Master of Science

Thesis: DISTRIBUTED RANDOMIZED BLOCK STOCHASTIC GRADIENT TRACKING METHODS: RATE ANALYSIS AND NUMERICAL EXPERIMENTS

Major Field: Industrial Engineering and Management

Biographical:

Born in Mumbai, Maharashtra, India. Eldest son of Mr. Vinayak Yevale and Mrs. Vibhavari Yevale, and brother to Suyash Yevale.

Education:

Completed the requirements for the Master of Science in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in 2021.

Completed the requirements for the Bachelor of Engineering in Mechanical Engineering at Savitribai Phule Pune University, Pune, Maharashtra, India in 2017.

Experience:

Worked as Graduate Research Assistant under Dr. Farzad Yousefian at the Department of Industrial Engineering and Management at Oklahoma State University, Stillwater, OK, from May, 2021 to December, 2021.

Worked as Graduate Teaching Assistant for Dr. Terry Collins and Dr. Farzad Yousefian at the Department of Industrial Engineering and Management at Oklahoma State University, Stillwater, OK, from August, 2019 to Dec, 2021.

Worked as an Associate Software Engineer at Accenture Solutions Pvt. Ltd., Pune, Maharastra, India, from October 2017 to March 2018.

Professional Membership:

International Affairs Officer, Graduate and Professional Student Government Association (GPSGA) at Oklahoma State University.

Member of Alpha Pi Mu, Honors Society Industrial Engineering and Management.

Member of Institute of Operations Research and Management Science (INFORMS).