ON DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS: ALGORITHMS, COMPLEXITY ANALYSIS, AND APPLICATIONS

By

HARSHAL D. KAUSHIK

Bachelor of Science in Mechanical Engineering
University of Pune
Pune, Maharashtra
2012

Master of Science in Applied Mechanics
Indian Institute of Technology (IIT), Madras
Chennai, Tamil Nadu
2015

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

ON DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS: ALGORITHMS, COMPLEXITY ANALYSIS, AND APPLICATIONS

Dissertation Approved:

Dr. Farzad Yousefian

Dissertation Advisor

Dr. Sunderesh S. Heragu

Dr. Austin Buchanan

Dr. Pratyaydipta Rudra

ACKNOWLEDGMENTS

This dissertation would not have been possible without the expertise of my advisor, Dr. Farzad Yousefian. I am deeply indebted and want to give my warmest thanks to Dr. Yousefian for his unwavering support and encouragement throughout my doctoral studies. He has been a very patient and an ideal advisor for me. I will be always grateful to Dr. Yousefian for his genuine concern towards my betterment and my academic success.

I want to appreciate the help and support from my committee members, Dr. Sunderesh Heragu, Dr. Austin Buchanan, Dr. Pratyaydipta Rudra, and Dr. Weiwei Hu (former member) for reviewing my dissertation and providing their valuable feedback.

The support provided by the School of Industrial Engineering and Management (IEM) and the working environment at the Advanced Technology Research Center (ATRC) have been ideal and very fruitful in my research work. I am overwhelmed by the amount of knowledge I was exposed to from all the formal as well as informal interactions with the faculty members at IEM. Specifically, I am thankful to Dr. Balabhaskar Balasundaram for his insightful courses. I am grateful to Dr. Teiming Liu and Dr. Sunderesh Heragu for their encouragement and valuable guidance. IEM staff members also have been very supportive.

I am thankful to my colleagues from IEM as well as from the School of Chemical Engineering. Long intuitive discussions were always fun and a great learning experience where we managed to discuss complicated/ challenging questions in relaxed settings. Many thanks to Bertan Özdoğru for being a very good friend.

Finally, I am deeply grateful to my loving parents, dear sisters, and all my family members for their support and sacrifices.

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

Name: HARSHAL D. KAUSHIK

Date of Degree: DECEMBER, 2021

Title of Study: ON DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS: ALGORITHMS, COMPLEXITY ANALYSIS, AND APPLICATIONS

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: Traditionally, constrained optimization models include constraints in the form of inequalities and equations. In this dissertation, we consider a unifying class of optimization problems with variational inequality (VI) constraints that allows for capturing a wide range of applications that may not be formulated by the existing standard constrained models. The main motivation arises from the notion of efficiency of equilibria in multi-agent networks. To this end, first we consider a class of optimization problems with Cartesian variational inequality (CVI) constraints, where the objective function is convex and the CVI is associated with a monotone mapping and a convex Cartesian product set. Motivated by the absence of performance guarantees for addressing this class of problems, we develop an averaged randomized block iteratively regularized gradient scheme. The main contributions include: (i) When the set of the CVI is bounded, we derive new non-asymptotic rate statements for suboptimality and infeasibility error metrics. (ii) When the set of the CVI is unbounded, we establish the global convergence in an almost sure and a mean sense. We numerically validate the proposed method on a networked Nash Cournot competition. We also implement our scheme on classical image deblurring applications and numerically demonstrate that the proposed scheme outperforms the standard sequential regularization method. In the second part, we consider a class of constrained multi-agent optimization problems where the goal is to cooperatively minimize the sum of agent-specific objectives. In this framework, the objective function and the VI mappings are locally known. We develop an iteratively regularized incremental gradient method where the agents communicate over a cycle graph. We derive new non-asymptotic agent-wise convergence rates for suboptimality and infeasibility metrics. We numerically validate the proposed scheme on a transportation network problem. We also apply the proposed scheme to address a special case of this distributed formulation, where the VI constraint characterizes a feasible set. We show the superiority of the proposed scheme to existing incremental gradient methods. A potential future direction is to extend the results of this dissertation to employ gradient tracking techniques and address multi-agent systems requiring weaker assumptions on the network topology with asynchronous communications.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

1.1 Optimization Problems with Variational Inequality Constraints 1

1.1.1 Motivation 2

1.2 Distributed Optimization Problems with Variational Inequality Constraints 5

1.2.1 Motivation 6

1.3 Summary of Contributions 8

1.4 Outline of Dissertation 10

1.5 Notation 11

II. OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS 12

2.1 Problem Formulation 12

2.2 Existing Methods and Research Gap 13

2.3 Outline of Algorithm 13

2.3.1 Preliminaries 15

2.4 Convergence Rate Analysis with Bounded Set X 16

2.5 Convergence Rate Analysis with Unbounded Set X 25

2.5.1 Preliminaries 25

2.5.2 Convergence Analysis 26

2.6 Experimental Results 31

2.7 Conclusions 32

Chapter Page

III. DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS 33

3.1 Problem Formulation 33

3.2 Existing Methods and Research Gap 34

3.3 Algorithm Outline 35

3.4 Rate and Complexity Analysis 39

3.5 Rate Analysis in the Solution Space 48

3.5.1 Preliminaries 48

3.5.2 Convergence analysis 51

3.6 Numerical Results 53

3.7 Conclusion 55

IV. ILL-POSED HIGH-DIMENSIONAL OPTIMIZATION PROBLEMS 56

4.1 Problem Formulation 56

4.2 Existing Methods and Research Gap 57

4.3 Algorithm Outline 58

4.3.1 Preliminaries 59

4.4 Convergence and Rate Analysis 60

4.5 Numerical Results 69

4.6 Concluding Remarks 70

V. LARGE-SCALE DISTRIBUTED NONLINEARLY CONSTRAINED OPTIMIZATION 71

5.1 Problem Formulation 71

5.2 Existing Methods and Research Gap 72

5.3 Algorithm Outline 73

5.4 Convergence Analysis 76

Chapter Page

5.5 Numerical Results 83
5.6 Concluding remarks 84

VI. CONCLUSIONS AND FUTURE DIRECTIONS 85
6.1 Conclusion 85
6.2 Future Directions 86

APPENDICES 92
A.1 Proof of Lemma 1.1.1 92
A.2 Proof of Lemma 1.2.1 92
A.3 Proof of Lemma 2.3.1 93
A.4 Proof of Lemma 2.3.2 93
A.5 Proof of Lemma 2.3.3 94
A.6 Proof of Corollary 2.4.1 94
A.7 Proof of Lemma 2.5.1 94
A.8 Proof of Lemma 3.3.3 96
A.9 Proof of Lemma 3.5.3 97

LIST OF TABLES

Table		Page
1	Comparison of incremental gradient schemes for solving finite-sum problems	34
2	Comparison of schemes for solving bilevel optimization problem	57
3	Comparison and memory requirements of incremental gradient schemes for addressing constrained finite sum problems.	72

LIST OF FIGURES

Figure		Page
1	The known blurred image and the unknown original image.	4
2	Image deblurring using the regularization technique with different values of η, for 10^5 iterations.	5
3	Algorithm 2 in terms of infeasibility and the objective function value . .	32
4	A transportation network with 2 nodes and 5 arcs	53
5	Performance of Algorithm 4 in terms of infeasibility and the objective function value for finding the best equilibrium in the transportation network problem	54
6	First set of images (a)-(e) are obtained using the sequential regularization at different values of η, running for 10^5 iterations. Second set of images (f)-(j) are obtained using Algorithm 5 by stopping at different iterations k.	70
7	Comparison of Algorithm 4 with standard IG methods in solving an SVM model	83

CHAPTER I

INTRODUCTION

Mathematical models and algorithms for constrained optimization often work under the premise that the functional constraints are in the form of inequalities, equations, or an easy-to-project set. A generic representation of an optimization problem in a standard form is as follows

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad \text{for all } i = 1, \ldots, m, \\ & h_j(x) = 0, \quad \text{for all } j = 1, \ldots, p, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ for all $i = 1, \ldots, m$, and $h_j : \mathbb{R}^n \to \mathbb{R}$ for all $j = 1, \ldots, p$, are functions and $x \in \mathbb{R}^n$. In a breadth of emerging applications in control theory, system constraints are too complex to be characterized as standard functional constraints. This complexity may arise in several network application domains where the optimization model is complicated by the presence of equilibrium constraints, complementarity constraints, or an inner-level large-scale optimization problem. The main goal in this dissertation lies in addressing this shortcoming by advancing the models and algorithms of constrained optimization by introducing a new unifying mathematical framework that is more powerful than the aforementioned standard optimization model in capturing a wide range of applications. In the following, we introduce this proposed mathematical formulation.

1.1 Optimization Problems with Variational Inequality Constraints

The first objective of this dissertation lies in addressing a unifying constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathrm{SOL}(X, F), \end{array} \tag{P_1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and set $X \subseteq \mathbb{R}^n$ is a Cartesian product, i.e., $X \triangleq \prod_{i=1}^{d} X_i$, where $X_i \subseteq \mathbb{R}^{n_i}$ for all $i = 1, \ldots, d$ and $\sum_{i=1}^{d} n_i = n$. Mapping $F : X \to \mathbb{R}^n$ is monotone and $\mathrm{SOL}(X, F)$ denotes the solution set of a variational inequality $\mathrm{VI}(X, F)$, defined next. A vector $x \in X$ is said to be a solution to $\mathrm{VI}(X, F)$ if for any $y \in X$, we have $F(x)^T(y - x) \geq 0$.

The variational inequality problem, first introduced in late 1950s, is an immensely powerful mathematical tool that can serve as a unifying framework for capturing a wide range of applications arising in operations research, finance, and economics (cf. [27, 30, 72, 89]). Importantly, this allows for addressing large-scale applications.

1.1.1 Motivation

In this section, to motivate the formulation (P_1), we present some examples below. Problem (P_1) can capture a variety of the standard problems in optimization and VI regimes. For example, when $F(x) = 0_n$, (P_1) is equivalent to the canonical optimization problem $\min_{x\in X} f(x)$. Also, when $f(x) = 0$, problem (P_1) is equivalent to $\mathrm{VI}(X, F)$. More detailed examples that can be reformulated as problem (P_1), are given in the following. We present four classes of optimization problems that can be reformulated as problem (P_1) as follows.

Example 1 (*Efficiency estimation of equilibria*) The main motivation for the formulation (P_1) arises from the notion of *efficiency of equilibria* in multi-agent networks including transportation networks, communication networks, and power systems. In multi-agent applications in non-cooperative regimes, the system behavior is governed by a collection of decisions (i.e., equilibrium) made by a set of independent and *self-interested* agents. As a result of this non-cooperative behavior (i.e., *game*) among the agents, the global performance of the system may become worse than the case where the agents cooperatively seek an *optimal* decision. A well-known example is the Prisoners' Dilemma where the costs of the players incurred by the Nash equilibrium are superior to their costs when they cooperate [68]. Indeed, it has been well-received in economics and computer science communities that Nash equilibria of a game may not attain full efficiency. This perception has led to a surge of research for understanding the quality of an equilibrium in non-cooperative games. In particular, addressing this question becomes imperative for network design [5] in the areas of routing [25] and load balancing [73]. In such networks, a *protocol designer* seeks *the best equilibrium* with respect to a global performance measure, i.e., function f in (P_1). To this end, the notion of "price of stability" (POS) is defined as the ratio between the best objective function value over the set of equilibria and that of an optimal outcome where there is no competition. In regard to the choice of the objective function f in (P_1), different approaches have been considered in the literature. Among popular choices are the utilitarian function and the egalitarian function. In the utilitarian approach, function f is defined as the summation of the individual objective functions of the agents, while in the egalitarian approach, the maximum of the individual cost functions is considered. In particular, in the context of network resource allocation where monetary value is measured, the utilitarian approach is also referred to as Marshallian aggregate surplus (e.g., see [44]). Below, we describe the details of the best equilibrium problem in the context of Nash games where we employ the utilitarian approach. Consider a canonical Nash game among d players where the i^{th} player is associated with a strategy $x^{(i)} \in X_i \subseteq \mathbb{R}^{n_i}$ and a cost function $g_i\left(x^{(i)}; x^{(-i)}\right)$, where $x^{(-i)}$ denotes the collection of actions of other players. Non-cooperative Nash games arise in a wide range of problems including communication networks [1, 2, 93], competitive interactions in cognitive radio networks [57, 78, 88], and power markets [47, 48, 81]. The game is defined formally as the following collection of problems for all $i = 1, \ldots, d$.

$$\begin{array}{ll} \text{minimize}_{x^{(i)}} & g_i\left(x^{(i)}; x^{(-i)}\right) \\ \text{subject to} & x^{(i)} \in X_i. \end{array} \qquad \mathrm{P}_i\left(x^{(-i)}\right)$$

A Nash equilibrium (NE) is a tuple of strategies $x^* \triangleq \left(x^{*(1)}; x^{*(2)}; \ldots; x^{*(d)}\right)$ where no player can obtain a lower cost by deviating from his own strategy, given that the strategies of the other players remain unchanged. It is known that (cf. Proposition 1.4.2 [30]) when for all

i, X_i is a closed convex set and g_i is a differentiable convex function with respect to $x^{(i)}$, the resulting equilibrium conditions of the Nash game given by $(\mathrm{P}_i\left(x^{(-i)}\right))$, are compactly captured by a Cartesian VI(X, F) where $X \triangleq \prod_{i=1}^{d} X_i$ and $F(x) \triangleq (F_1(x); \ldots; F_d(x))$ with $F_i(x) \triangleq \nabla_{x^{(i)}} g_i\left(x^{(i)}; x^{(-i)}\right)$. The set SOL$(X, F)$ will then represent the set of Nash equilibria to the game $(\mathrm{P}_i\left(x^{(-i)}\right))$. The best NE problem employing the utilitarian approach is formulated as follows.

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{d} g_i\left(x^{(i)}; x^{(-i)}\right) \\ \text{subject to} \quad & x \in \mathrm{SOL}\left(\prod_{i=1}^{d} X_i, (\nabla_{x^{(1)}} g_1; \ldots; \nabla_{x^{(d)}} g_d)\right). \end{aligned} \tag{1.1.1}$$

In Chapter 2, we address the model (1.1.1) for a class of networked Nash-Cournot games.

Example 2 (*High-dimensional constrained convex optimization*) Another class of problems that can be captured by model (P_1) is the following convex optimization problem with nonlinear inequalities and linear equations

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \\ & h_j(x) \leq 0 \quad \text{for all } j = 1, \ldots, \mathcal{J} \\ & x \in X \triangleq \prod_{i=1}^{d} X_i, \end{aligned} \tag{1.1.2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $h_j : \mathbb{R}^n \to \mathbb{R}$ is convex for all j, set X_i is convex of all $i \in \{1, \ldots, d\}$. The next lemma presents the details on how problem (1.1.2) can be cast as (P_1).

Lemma 1.1.1 *Suppose problem* (1.1.2) *is feasible. Let $h_j(x)$ be a continuously differentiable and convex function for all j. Let the set $X_i \in \mathbb{R}^{n_i}$ be nonempty, closed, and convex for all i. Then, problem* (1.1.2) *is equivalent to the problem* (P_1) *where the mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ be defined as follows*

$$F(x) \triangleq A^T(Ax - b) + \sum_{j=1}^{\mathcal{J}} \max\{0, h_j(x)\} \nabla h_j(x).$$

Proof. See Appendix A.1. ■

Example 3 (*Ill-posed high-dimensional optimization problem*) Linear inverse problems arising in image deblurring can be cast as,

$$\begin{aligned} &\text{minimize } \| Ax - b\|^2 \\ &\text{s.t.} \quad x \in \mathbb{R}^n, \end{aligned} \tag{1.1.3}$$

where $A \in \mathbb{R}^{m \times n}$ is a blurring operator, $b \in \mathbb{R}^m$ is the given blurred image in Figure 1(a), and $x \in \mathbb{R}^n$ is a deblurred image in Figure 1(b). This is an ill-posed problem in the sense

that there may be multiple solutions or the optimal solution x may be very sensitive to the perturbations in the input b. To address the ill-posedness, and to induce sparsity and stability, problem (1.1.3) can be reformulated in a bilevel structure as following (e.g., see [33]),

$$\begin{aligned} &\text{minimize } \|x\|^2 \\ &\text{s.t.} \quad x \in \operatorname{argmin}\left\{ \|Ax-b\|^2 : x \in \mathbb{R}^n \right\}. \end{aligned} \tag{1.1.4}$$

Problem (1.1.4) can be formulated as follows

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \operatorname{argmin}\{g(x) : x \in X\}, \end{aligned} \tag{1.1.5}$$

where functions f and g are defined as $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. In particular, here we consider the case where the set X has a block structure, i.e., $X = \prod_{i=1}^{d} X_i$, where $X_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^{d} n_i = n$. Under the convexity of f, g, and set X_i, problem (1.1.5) can be captured by (P_1) such that F is a gradient map, given as $F(x) \triangleq (\nabla_{x^{(1)}} g(x); \ldots; \nabla_{x^{(d)}} g(x))$ with set $X = \prod_{i=1}^{d} X_i$.

(a) Blurred image

(b) Original image

Figure 1: The known blurred image and the unknown original image.

Note that formulation (1.1.5) captures problem (1.1.4) for $f(x) \triangleq \frac{1}{2}\|x\|_2^2$, $g(x) \triangleq \frac{1}{2}\|Ax-b\|^2$ and $X \triangleq \mathbb{R}^n$. One of the popular ways to address problem (1.1.4) is by regularizing it. Consider the regularized problem (1.1.6) as the following

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|^2 + \frac{\eta\|x\|^2}{2}, \tag{1.1.6}$$

where η is the regularization parameter.

As the value of η is a priori unknown, to obtain a suitable choice for η, problem (1.1.6) must be solved multiple times for different values of η until a satisfactory deblurring is achieved. Here, $\eta \in (0, +\infty)$ governs the way by which solution of problem (1.1.4) is approximated through solving the model (1.1.6). This framework is called the sequential regularization scheme that is presented in Algorithm 1. Algorithm 1 is a two-loop framework where at each iteration, given a fixed parameter η_t, $\operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{\|Ax-b\|^2}{2} + \eta_t \frac{\|x\|^2}{2} \right\}$ needs to be solved. In the special case where $f(x) := \frac{1}{2}\|x\|^2$, it can be shown when $\eta_t \to 0$, under the

Algorithm 1 The SR scheme for solving problem (P_1) when $f := \frac{1}{2}\|\cdot\|_2^2$

1: **Input:** Initial regularization parameter $\eta_0 > 0$;
2: **for** t = 0, 1, ... **do**
3: Compute $x^*_{\eta_t} := \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\|Ax-b\|^2}{2} + \eta_t \frac{\|x\|^2}{2} \right\}$;
4: Update η_t to η_{t+1} such that $\eta_{t+1} < \eta_t$;
5: **end for**

convexity of function $\frac{\|Ax-b\|^2}{2} + \eta_t \frac{\|x\|^2}{2}$, and closedness and convexity of the set $\mathbb{R}^n$, any limit point of the *Tikhonov trajectory*, denoted by $\{x^*_{\eta_t}\}$, where $x^*_{\eta_t} := \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\|Ax-b\|^2}{2} + \eta_t \frac{\|x\|^2}{2} \right\}$, converges to the least ℓ_2–norm solution of the problem $\underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\|Ax-b\|^2}{2} \right\}$ (cf. Chapter 12 in [30]).

(a) η=0 (b) η=0.001 (c) η=0.01 (d) η=0.1 (e) η=1

Figure 2: Image deblurring using the regularization technique with different values of η, for 10^5 iterations.

Figure 2 shows the deblurred images obtained by the conventional sequential regularization (i.e. Algorithm 1) at different values of η for 10^5 iterations. As evidenced, it is computationally inefficient to find a suitable regularization parameter η. In Chapter 4, we provide numerical experiment where we formulate problem (1.1.5) as problem (P_1) and show the effectiveness of the proposed scheme.

Example 4 (*Optimization with a system of nonlinear equation constraints*) Consider the following optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & F(x) = 0, \end{array} \tag{1.1.7}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone mapping. Defining $X \triangleq \mathbb{R}^n$, $\mathrm{SOL}(X, F)$ is equal to the feasible solution set of the problem (1.2.4). This implies that problem (1.2.4) can be captured by (P_1).

1.2 Distributed Optimization Problems with Variational Inequality Constraints

In the second part of the dissertation, we consider a decentralized structure of the optimization problem (P_2) where a collection of agents (e.g., processing units, sensors) communicate

their local information with their neighboring agents to cooperatively optimize a global objective. It is through this cooperation that learning from massive datasets can be made possible. Moreover, the decentralized storage of the data over the network may allow for preserving the privacy of the agents. Distributed optimization has found a wide range of applications in wireless sensor networks, machine learning, and signal processing [64]. Despite the significant advances in the design and analysis of the optimization methods over networks, the existing models and algorithms are still less satisfactory in some regimes than those of centralized optimization. For example, there is still much left to be understood about how to tackle the presence of nonlinearity and uncertainty in the functional constraints, while requiring a low number of communications and enforcing weak assumptions on the network topology. The goal in this part of dissertation is to tackle some of the shortcomings in distributed constrained optimization through considering a new unifying mathematical framework described as follows. Consider a system with m agents where the i^{th} agent is associated with a component function $f_i : \mathbb{R}^n \to \mathbb{R}$ and a mapping $F_i : \mathbb{R}^n \to \mathbb{R}^n$. Our goal is to solve the following distributed constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m f_i(x) \\ \text{subject to} & x \in \text{SOL}\left(X, \sum_{i=1}^m F_i\right), \end{array} \tag{P_2}$$

where $X \subseteq \mathbb{R}^n$ is a set and $\text{SOL}\left(X, \sum_{i=1}^m F_i\right)$ denotes the solution set of the variational inequality $\text{VI}\left(X, \sum_{i=1}^m F_i\right)$ defined as follows: $x \in X$ solves $\text{VI}\left(X, \sum_{i=1}^m F_i\right)$ if we have $(y - x)^T \sum_{i=1}^m F_i(x) \geq 0$ for all $y \in X$. Problem (P_2) represents a distributed optimization framework in a sense that the information about f_i and F_i is locally known to the i^{th} agent, while the set X is globally known to all the agents.

1.2.1 Motivation

In this section, we provide motivating examples for problem (P_2). By choosing $F_i(x) := 0_n$ for all i, model (P_2) captures the canonical formulation of distributed optimization

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m f_i(x) \\ \text{subject to} & x \in X, \end{array} \tag{1.2.1}$$

that has been extensively studied in the literature. Next, we show how the proposed model (P_2) is employed to capture more challenging distributed constrained optimization problems.

Example 5 (*Distributed optimization problems with complementarity constraints*) Nonlinear complementarity problems (NCP) have been employed to formulate diverse applications in engineering and economics. The celebrated Wardrop's principle of equilibrium in traffic networks and also, the Walras's law of competitive equilibrium in economics are among important examples that can be represented using NCP (cf. [31]). Formally, NCP is defined as follows. Given a mapping $F : \mathbb{R}^n_+ \to \mathbb{R}^n$, $x \in \mathbb{R}^n$ solves $\text{NCP}(F)$ if $0 \leq x \perp F(x) \geq 0$, where $\perp$ denotes the perpendicularity operator between two vectors. It is known that $\text{NCP}(F)$ can be cast as $\text{VI}(\mathbb{R}^n_+, F)$ (see Proposition 1.1.3 in [30]). In many applications where F is merely monotone, $\text{NCP}(F)$ may admit multiple equilibria. In such cases,

one may consider finding the best equilibrium with respect to a global metric $f : \mathbb{R}^n \to \mathbb{R}$. For example in traffic networks, the total travel time of the network users can be considered as the objective f. In fact, the problem of computing the best equilibrium of an NCP is important to be addressed particularly in the design of transportation networks where there is a need to estimate the efficiency of the equilibrium [44, 67]. In this regime, the goal is to minimize $f(x)$ where x solves NCP(F). Motivated by applications in traffic equilibrium problems, stochastic variants of NCP have been considered more recently [23, 99]. Consider a stochastic NCP, given by

$$x \geq 0, \quad \mathbb{E}\left[F(x, \xi(\omega))\right] \geq 0, \quad x^T \mathbb{E}\left[F(x, \xi(\omega))\right] = 0,$$

where $\xi : \Omega \to \mathbb{R}^d$ is a random variable associated with the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $F : \mathbb{R}^n_+ \times \Omega \to \mathbb{R}^n$ is a stochastic single-valued mapping. Let $\mathcal{S}_i$ denote a local index set of independent and identically distributed samples from the random variable ξ. Employing a sample average approximation scheme, one can consider a distributed NCP, given by

$$x \geq 0, \quad \sum_{i=1}^m \sum_{\ell \in \mathcal{S}_i} F(x, \xi_\ell) \geq 0, \quad x^T \left(\sum_{i=1}^m \sum_{\ell \in \mathcal{S}_i} F(x, \xi_\ell)\right) = 0.$$

Let $f : \mathbb{R}^n_+ \times \Omega \to \mathbb{R}^n$ denote a stochastic objective function that measures the performance of a given equilibrium at a realization of ξ. Then, the problem of distributed computation of the best equilibrium of the preceding NCP is formulated as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \sum_{\ell \in \mathcal{S}_i} f(x, \xi_\ell) \\ \text{subject to} & x \in \mathrm{SOL}\left(\mathbb{R}^n_+, \sum_{i=1}^m \sum_{\ell \in \mathcal{S}_i} F(\bullet, \xi_\ell)\right). \end{array} \tag{1.2.2}$$

Importantly, the proposed model (P_2) captures problem (1.2.2) by defining $X \triangleq \mathbb{R}^n_+$, $f_i(x) \triangleq \sum_{\ell \in \mathcal{S}_i} f(x, \xi_\ell)$, and $F_i(x) \triangleq \sum_{\ell \in \mathcal{S}_i} F(x, \xi_\ell)$. In Chapter 3, we present preliminary numerical experiments where we solve problem (1.2.2) for a given transportation network.

Example 6 (*Distributed optimization problems with local nonlinear inequality and linear equality constraints*) Another class of problems that can be captured by problem (P_2) is given as

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^m f_i(x) & \\ \text{subject to} & g_{i,1}(x) \leq 0\ ,\ \ldots\ ,\ g_{i,n_i}(x) \leq 0, & \text{for } i \in \{1, \ldots, m\}, \\ & A_i x = b_i, & \text{for } i \in \{1, \ldots, m\}, \\ & x \in X, & \end{array} \tag{1.2.3}$$

where agent i is associated with function $f_i : \mathbb{R}^n \to \mathbb{R}$, functions $g_{i,j} : \mathbb{R}^n \to \mathbb{R}$ for $j \in [n_i]$, and parameters $A_i \in \mathbb{R}^{d_i \times n}$ and $b_i \in \mathbb{R}^{d_i}$. The notation $[m]$ is used to abbreviate $\{1, \ldots, m\}$. The set X is globally known to all the agents while $f_i(x)$, $g_{i,j}(x)$, A_i, and b_i are locally known to agent i. In the following, we show that problem (1.2.3) can be represented as model (P_2).

Lemma 1.2.1 *Consider problem* (1.2.3). *Let functions $g_{i,j}(x)$ be continuously differentiable and convex for all $i \in [m]$ and $j \in [n_i]$. Assume that the feasible region of problem* (1.2.3) *is nonempty and the set X is closed and convex. Then, problem* (1.2.3) *is equivalent to* (P_2) *where we define $F_i : \mathbb{R}^n \to \mathbb{R}^n$ as $F_i(x) \triangleq A_i^T(A_i x - b_i) + \sum_{j=1}^{n_i} \max\{0, g_{i,j}(x)\} \nabla g_{i,j}(x)$.*

Proof. See Appendix A.2. ■

Example 7 (*Distributed optimization coupling nonlinear equality constraints*) Another class of problems that can be reformulated as (P_2) is given as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} f_i(x) \\ \text{subject to} & \sum_{i=1}^{m} F_i(x) = 0, \end{array} \tag{1.2.4}$$

where agent i is associated with a local mapping $F_i : \mathbb{R}^n \to \mathbb{R}^n$ and a local objective $f_i : \mathbb{R}^n \to \mathbb{R}$. This model finds relevance to the multi-user optimization problems in network resource allocation applications where the objective and constraints are not separable by each user [22, 56, 86]. Note that the feasible solution set of problem (1.2.4) is equal to $\mathrm{SOL}(\mathbb{R}^n, \sum_{i=1}^{m} F_i)$, implying that (1.2.4) is captured by (P_2).

1.3 Summary of Contributions

The main contributions of this dissertation are summarized as follows. In addressing problem (P_1), we make the following main contributions:
(i) *Development of a single-timescale method equipped with complexity for* (P_1): In addressing (P_1), we develop an efficient first-order method called averaging randomized block iteratively regularized gradient (aRB-IRG). The proposed method is single-timescale in the sense that, unlike the SR approach, it does not require solving a VI at each iteration. Instead, it only uses evaluations of the mapping F and the subgradient of the objective function f at each iteration. In the first part of this chapter, we consider the case where the set X is bounded. We let f be a subdifferentiable merely convex function and F be a monotone mapping. In Theorem 2.4.1, we derive a suboptimality convergence rate in terms of the expected value of the objective function. We also derive a convergence rate for the infeasibility that is characterized by the expected value of a dual gap function. We also derive deterministic variants of the aforementioned convergence rates when we suppress the randomized block-coordinate scheme. In the second part of this chapter, we consider the case where the set X is unbounded and f is smooth and strongly convex. Utilizing the properties of the Tikhonov trajectory, we establish the global convergence of the scheme in an almost sure and a mean sense. To the best of our knowledge, this work appears to be the first one that provides the two rate statements for problems of the form (P_1). In particular, the complexity analysis in this work contributes to the existing convergence theory in several previous papers including [46, 49, 51, 57, 91, 95, 98]. Moreover, in the special case where the VI constraint represents the optimal solution set of an optimization problem, (P_1) captures a class of bilevel optimization problems. This class of problems has been studied in a number of recent papers in deterministic [10, 33, 75, 82], stochastic [4], and distributed regimes [94]. However, the complexity analysis in the aforementioned papers lacks a suboptimality rate, or lacks an infeasibility rate, or requires much stronger assumptions such as strong convexity and smoothness of f. In Chapter 4, to address ill-posed optimization problem in an image deblurring application, we consider a bilevel optimization problem where we seek among the optimal solutions of the inner level problem, a solution that minimizes a secondary metric. Minimum norm gradient, sequential averaging, and iterative regularization are among the known schemes developed for addressing this class of problems. However, to address the

nondifferentiability of the objective function and high-dimensionality of the solution space, we develop a single-loop scheme called randomized block iteratively regularized subgradient (RB-IRG). Under a uniform selection of the block and a careful choice of the parameter sequences, we establish the almost sure convergence and derive a non-asymptotic rate of convergence with respect to the inner level objective function.

(ii) *Advancing the convergence rate properties of the randomized block-coordinate schemes*: Block-coordinate schemes, and specifically their randomized variants, have been widely studied in addressing the standard optimization problems (e.g., see [28, 66, 71, 80, 97]). However, in addressing VI problems, there are only a handful of recent papers, including [58, 97], that employ this technique and are equipped with rate guarantees. The aforementioned papers address standard VI problems that can be viewed a special case of the model (P_1) where $f(x) := 0$. In this work, we extend the convergence and rate analysis of the randomized block-coordinate schemes to the much broader regime of optimization problems with CVI constraints.

(iii) *Addressing high-dimensionality employing a randomized block-coordinate scheme.* The proposed Algorithms 2 and 5 employ a randomized block-coordinate protocol for updating the iterates. Block-coordinate schemes, and specifically their randomized variants, have been recently studied in addressing standard optimization problems (e.g., see [28, 66, 71, 80, 97]). However, in addressing VI problems there are only a handful of recent papers, including [58, 97], that employ this feature and are equipped with rate guarantees. The aforementioned papers address standard VI problems that can be viewed a special case of problem (P_1) where $f(x) = 0$. Chapter 2 of this work extends the convergence and rate analysis of the randomized block-coordinate schemes to a address problems in a broader regime that is optimization problems with CVI constraints.

In addressing problem (P_2), we make the following main contributions:

(i) *Complexity guarantees for addressing model* (P_2): We develop a distributed iterative method equipped with agent-specific iteration complexity guarantees for solving distributed optimization problems with VI constraints of the form (P_2). To this end, employing a regularization-based relaxation technique, we propose a projected averaging iteratively regularized incremental gradient method (pair-IG) presented by Algorithm 4. In Theorem 5.4.1, under merely convexity of the global objective function and merely monotonicity of the global mapping, we derive new non-asymptotic suboptimality and infeasibility convergence rates for each agent's generated iterates. This implies a total iteration complexity of $\mathcal{O}\left((C_f + C_F)^4 \epsilon^{-4}\right)$ for obtaining an ϵ-approximate solution where C_f and C_F denote the bounds on the global objective function's subgradients and the global mapping over a compact convex set X, respectively. Iterative regularization (IR) has been recently employed as a constraint-relaxation technique in a class of bilevel optimization problems [4, 94] and also in regimes where the duality theory may not be directly applied [52, 95]. Of these, in Chapter 2 we employ the IR technique to derive a provably convergent method for solving problem (P_2) in a centralized framework, where the information of the objective function is globally known by the agents. Unlike in Chapter 2, in Chapter 3 we assume that the agents have access only to local information about both the objective function and the mapping. It is worth emphasizing that this lack of centralized access to information introduces a major challenge in both the design and the complexity analysis of the new algorithmic framework in addressing the distributed model (P_2). Motivated by the need for distributed

implementations, in a preliminary version of this work, in [51], we addressed a subclass of the distributed model (P_2) formulated as (1.1.2). Extending [51], here we show that the convergence and rate analysis in [51] can be extended to the much broader class of problems of the form (P_2) without any degradation of the speed of the algorithm. In addressing hierarchical optimization problems, there have been other iterative methods proposed in works such as [10, 33, 75, 91]. We, however, note that all of these schemes can only address a subclass of (P_2) in centralized regimes and under more restrictive assumptions such as strong convexity of the global objective function. We also note that compared to the existing methods that address VI problems (e.g., see [40, 45, 89, 95]), our work provides an avenue for addressing a significantly more general class of problems where VI is employed as a tool to characterize the constraints in distributed optimization. In Chapter 5, for addressing the problem of finite sum of nondifferentiable convex function where each component function corresponds to a hard-to-project constraint set, we devise an algorithm called averaged iteratively regularized incremental gradient (aIR-IG) that does not involve any hard-to-project computations. Under mild assumptions, we derive non-asymptotic rates of convergence for both suboptimality and infeasibility metrics.

(ii) *Distributed averaging scheme*: In pair-IG, we employ a distributed averaging scheme where agents can choose their initial averaged iterate arbitrarily and independent from each other. This relaxation in the proposed IG method appears to be novel, even for the classical IG schemes in addressing (1.2.1).

(iii) *Rate analysis in the solution space*: Motivated by the recent developments of iterative methods for MPECs [26], it is important to characterize the speed of the proposed scheme in the solution space. To this end, under strong convexity of the global objective function, in Theorem 3.5.1 we derive agent-specific rate statements that compare the generated sequence of each agent with the so-called Tikhonov trajectory, that is defined as the trajectory of the unique solutions to a family of regularized optimization problems.

(iv) *Preliminary numerical results*: To validate the theoretical results, we provide preliminary numerical experiments for computing the best equilibrium in a multi-agent traffic network problem in Chapter 3. We also compare the performance of the proposed IG scheme with that of the existing IG methods in addressing constrained finite-sum problems in Chapter 5.

1.4 Outline of Dissertation

The remainder of the dissertation is organized as follows. In Chapter 2, we address problem (P_1). We provide the main assumptions, present the outline of algorithms, and perform convergence and complexity analysis. In the case where the associated set of the CVI is bounded and the objective function is nondifferentiable and convex, we derive new non-asymptotic suboptimality and infeasibility convergence rate statements in a mean sense. We also obtain deterministic variants of the convergence rates when we suppress the randomized block-coordinate scheme. In the case where the CVI set is unbounded, utilizing the strong convexity of objective function, we establish the global convergence of the proposed algorithm in an almost sure and a mean sense. We numerically validate the proposed method on a networked Cournot competition model. In Chapter 3, to address problem (P_2), we devise a distributed algorithm, provide convergence and agent-wise rate analysis. We numerically

validate the proposed scheme on a transportation network problem. In Chapter 4, we discuss a special case of formulation (P_1), where we address ill-posed high-dimensional optimization problems. We provide the outline of the proposed algorithm, discuss almost sure convergence of the iterates generated, and provide the non-asymptotic rate statements. Chapter 5 includes a special case of problem formulation (P_2), where we address large-scale constrained convex optimization problem. Further, we provide the convergence and rate analysis for the proposed scheme. We conclude this dissertation in Chapter 6 by summarizing the key findings and discussing the potential future directions.

1.5 Notation

Throughout this dissertation, a vector $x \in \mathbb{R}^n$ is assumed to be a column vector and x^T denotes the transpose of x. We use $x^{(i)} \in \mathbb{R}^{n_i}$ to denote the i^{th} block-coordinate of vector x where $x = \big(x^{(1)}; \ldots; x^{(d)}\big)$ and $\sum_{i=1}^{d} n_i = n$. For a function $f : \mathbb{R}^n \to \mathbb{R}$, vector $\tilde{\nabla} f(x) \in \mathbb{R}^n$ is called a subgradient of f at x if $f(x) + \tilde{\nabla} f(x)^T f(y-x) \leq f(y)$ for all $y \in \text{dom}(f)$. $\tilde{\nabla}_i f(x)$ is the i^{th} block of $\tilde{\nabla} f(x)$. The subdifferential set of f at x is the set of all subgradients of f at x and is denoted by $\partial f(x)$. The Euclidean norm of a vector x is denoted by $\|x\|$, i.e., $\|x\| \triangleq \sqrt{x^T x}$. For a mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, we denote the i^{th} block-coordinate of F by $F_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$, i.e., $F(x) = (F_1(x); \ldots; F_d(x))$. A mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is said to be monotone on a convex set $X \subseteq \mathbb{R}^n$ if for any $x, y \in X$, we have $(F(x) - F(y))^T (x - y) \geq 0$. The mapping F is said to be μ–strongly monotone on a convex set $X \subseteq \mathbb{R}^n$ if $\mu > 0$ and for any $x, y \in X$, we have $(F(x) - F(y))^T (x - y) \geq \mu \|x - y\|^2$. Also, F is said to be Lipschitz with parameter $L > 0$ on the set X if for any $x, y \in X$, we have $\|F(x) - F(y)\| \leq L\|x - y\|$. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called μ–strongly convex on a convex set X if $f(x) \geq f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2}\|x - y\|^2$. Function f is μ–strongly convex if and only if ∇f is μ–strongly monotone on X. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be in the class $C^{k,r}_{\mu,L}$ if f is μ-strongly convex in $\mathbb{R}^n$, k times continuously differentiable, and its r^{th} derivative is Lipschitz continuous with constant L. We use $\mathbf{I}_n$ to denote the identity matrix of size $n \times n$. The probability of an event Z is denoted by $\text{Prob}(Z)$ and the expectation of a random variable z is denoted by $\mathsf{E}[z]$. We use $\mathbb{R}^n_+$ and $\mathbb{R}^n_{++}$ to denote $\{x \in \mathbb{R}^n \mid x \geq 0\}$ and $\{x \in \mathbb{R}^n \mid x > 0\}$, respectively. a.s. used for 'almost surely'. $\mathcal{F}_k$ denotes the set of variables $\{i_0, \ldots, i_{k-1}\}$. For a random variable i_k, $\mathsf{Prob}(i_k = i)$ is p_{i_k}. For any symmetric square matrix $B \in \mathbb{R}^{n \times n}$, the spectral norm is denoted by $\|B\|$ and is defined as the maximum absolute value of eigenvalues of the matrix, i.e., we have $\|B\| \triangleq \max\{|\lambda_{\min}(B)|, |\lambda_{\max}(B)|\}$. Note that, for a positive semidefinite matrix B, we have $\|B\| = \lambda_{\max}(B)$. The Euclidean projection of vector z onto set X is denoted as $\mathcal{P}_X(x)$, where $\mathcal{P}_X(z) \triangleq \text{argmin}_{x \in X} \|x - z\|_2$. Given a set $S \subseteq \mathbb{R}^n$, we let $\text{int}(S)$ denote the interior of S. Given an integer m, we let $[m]$ abbreviate the set $\{1, \ldots, m\}$.

CHAPTER II

OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS

In this chapter, we consider a class of optimization problems with Cartesian variational inequality (CVI) constraints (P_1) with centralized communication. This mathematical formulation captures a wide range of applications including those complicated by the presence of equilibrium constraints, complementarity constraints, or an inner-level large scale optimization problem. In particular, important application arises from the notion of equilibrium in multi-agent network, e.g., communication networks and power systems. In the literature, the complexity of the existing solution methods for optimization problems with CVI constraints appears to be unknown. Motivated by this, here we propose an averaged randomized block iteratively regularized gradient scheme aRB-IRG. Section 2.1 includes the problem formulation and Section 2.2 summarizes literature for addressing problem (P_1). The proposed scheme is presented in Algorithm 2. The outline of algorithm and the required preliminaries are provided in Section 2.3. The main contributions include: (i) In the case where the associated set of the CVI is bounded and the objective function is nondifferentiable and convex, we derive new non-asymptotic suboptimality and infeasibility convergence rate statements in an ergodic sense. We also obtain deterministic variants of the convergence rates when we suppress the randomized block-coordinate scheme. Importantly, this appears to be the first work to provide these rate guarantees for this class of problems. (ii) In the case where the CVI set is unbounded and the objective function is smooth and strongly convex, utilizing the properties of the Tikhonov trajectory, we establish the global convergence of aRB-IRG in an almost sure and a mean sense. Section 2.4 includes the convergence analysis and rate results for the case of bounded set X in problem (P_1). The analysis for an unbounded set X is presented in Section 2.5. Section 2.6 gives the numerical implementation of Algorithm 2 on a Nash Cournot competition. Section 2.7 provides some concluding remarks.

2.1 Problem Formulation

Consider function $f : \mathbb{R}^n \to \mathbb{R}$ and set $X \subseteq \mathbb{R}^n$ as a Cartesian product, i.e., $X \triangleq \prod_{i=1}^d X_i$, where $X_i \subseteq \mathbb{R}^{n_i}$ for all $i = 1, \ldots, d$ and $\sum_{i=1}^d n_i = n$. We consider centralized unifying constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathrm{SOL}(X, F). \end{array} \tag{P_1}$$

The content of this chapter has been published in the SIAM Journal on Optimization [52].

Mapping $F : X \to \mathbb{R}^n$ is a monotone map and $\mathrm{SOL}(X, F)$ denotes the solution set of a variational inequality $\mathrm{VI}(X, F)$, defined next. A vector $x \in X$ is said to be a solution to $\mathrm{VI}(X, F)$ if for any $y \in X$, we have $F(x)^T(y - x) \geq 0$.

2.2 Existing Methods and Research Gap

We first begin by providing a brief overview of the solution methods for addressing a VI problem. Starting from the seminal work of Lemke and Howson [59] and Scarf [76], who developed the first solution methods for computing equilibria, in the past few decades, there has been a surge of research on the development and analysis of the computational methods for solving VIs. Perhaps this interest lies in the strong interplay between the VIs and the formulation of optimization and equilibrium problems arising in many communication and networking problems [77]. Korpelevich's celebrated extragradient method [55] and its extensions [19, 20, 39, 41, 45, 97] were developed which require weaker assumptions than their gradient counterparts. In the past decade, there has been a trending interest in addressing VIs in the stochastic regimes. Among these, Jiang and Xu [43] developed the stochastic approximation methods for solving VIs with strongly monotone and smooth mappings. This work was later extended to the case with merely monotone mappings [46, 57] and nonsmooth mappings [96].

Despite much advances in the theory and algorithms for VIs, solving the problem (P_1) has remained challenging. To the best of our knowledge, the computational complexity of the existing solution methods for addressing (P_1) is unknown. In addressing the standard constrained optimization problems, Lagrangian duality and relaxation rules have often proven to be very successful [14]. However, when it comes to solving (P_1), the duality theory cannot be practically employed. This is primarily because unlike in the standard constrained optimization problems where the objective function provides a metric for distinguishing solutions, there is no immediate analog in the VI problems. Inspired by the contributions of Andrey Tikhonov in 1980s on addressing illposed optimization problems, the existing methods for solving (P_1) share in common a sequential regularization (SR) scheme presented by Algorithm 1. The SR scheme is a two-loop framework where at each iteration, given a fixed parameter η_t, a regularized VI denoted by $\mathrm{VI}\,(X, F + \eta_t \mathbf{I}_n)$ is required to be solved. In the special case where $f(x) := \frac{1}{2}\|x\|^2$, it can be shown when $\eta_t \to 0$, under the monotonicity of the mapping F and closedness and convexity of the set X, any limit point of the *Tikhonov trajectory* denoted by $\{x^*_{\eta_t}\}$, where $x^*_{\eta_t} \in \mathrm{SOL}\,(X, F + \eta_t \mathbf{I}_n)$, converges to the least ℓ_2–norm vector in $\mathrm{SOL}(X, F)$ (cf. Chapter 12 in [30]). The SR approach is associated with two main drawbacks: (i) It is a computationally inefficient scheme, as it requires solving a series of increasingly more difficult VI problems. (ii) The iteration complexity of the SR scheme in addressing the problem (P_1) is unknown. Accordingly, the main goal in this dissertation lies in the development of an efficient scheme equipped with computational complexity analysis for solving the problem (P_1).

2.3 Outline of Algorithm

In this section, we state the main assumptions and present the proposed scheme for solving the optimization Problem (P_1).

Assumption 2.3.1 *Consider the Problem* (P_1) *under the following conditions:*
(a) The set X_i is nonempty, closed, and convex for all $i = 1, \ldots, d$.
(b) The function f is convex and has bounded subgradients over the set X.
(c) The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuous, monotone, and bounded over the set X.
(d) The optimal solution set of Problem (P_1) *in nonempty.*

Assumption 2.3.1(b) implies that f is Lipschitz continuous over the set X. Under this assumption, we address a broad class of problems of the form Problem (P_1) where the objective function is possibly nondifferentiable and nonstrongly convex. In the following, we discuss the conditions under which Assumption 2.3.1(d) is satisfied.

Remark 2.3.1 (Existence of an optimal solution) Suppose Assumption 2.3.1(a), (b), and (c) hold. The existence of an optimal solution to the Problem (P_1) can be established under different conditions. We provide two instances as follows: (i) Suppose there exists a vector $\bar{x} \in X$ such that the set $\bar{X} \triangleq \{x \in X : \ F(x)^T(x - \bar{x}) \leq 0\}$ is bounded. Then, from Proposition 2.2.3 in [30], $\mathrm{SOL}(X, F)$ is nonempty and compact. Consequently, the Weierstrass' Theorem implies the existence of an optimal solution to the Problem (P_1). (ii) Suppose the set X is compact. Then, from Corollary 2.2.5 in [30], the set $\mathrm{SOL}(X, F)$ is nonempty and compact. Again, Assumption 2.3.1(d) is guaranteed by the Weierstrass' Theorem.

Throughout, this chapter we let $C_F > 0$ denote the bound on the Euclidean norm of the mapping F, i.e., $\|F(x)\| \leq C_F$ for all $x \in X$. Also, we let $C_f > 0$ denote the bound on the norm of the subgradients of f, i.e., $\|\tilde{\nabla} f(x)\| \leq C_f$ for all $\tilde{\nabla} f(x) \in \partial f(x)$ and $x \in X$. The outline of the proposed method is presented by Algorithm 2. At iteration k, a block-

Algorithm 2 Averaged Randomized Block Iteratively Regularized Gradient

1: **Input:** A random initial point $x_0 \in X$, $\bar{x}_0 := x_0$, initial stepsize $\gamma_0 > 0$, initial regularization parameter $\eta_0 > 0$, a scalar $0 \leq r < 1$, and $S_0 := \gamma_0^r$.
2: **for** k = 0, 1, ... **do**
3: Generate a realization of random variable i_k according to Assumption 2.3.2.
4: Evaluate $F_{i_k}(x_k)$ and $\tilde{\nabla}_{i_k} f(x_k)$ where $\tilde{\nabla} f(x_k) \in \partial f(x_k)$.
5: Update x_k as follows:

$$x_{k+1}^{(i)} := \begin{cases} \mathcal{P}_{X_{i_k}}\left(x_k^{(i_k)} - \gamma_k \left(F_{i_k}\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} f\left(x_k\right)\right)\right) & \text{if } i = i_k, \\ x_k^{(i)} & \text{if } i \neq i_k. \end{cases} \tag{2.3.1}$$

6: Obtain γ_{k+1} and η_{k+1} (cf. Theorem 2.4.1 and Theorem 2.5.1 for the update rules).
7: Update the averaged iterate $\bar{x}_k$ as follows:

$$S_{k+1} := S_k + \gamma_{k+1}^r, \quad \bar{x}_{k+1} := \frac{S_k \bar{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}}. \tag{2.3.2}$$

8: **end for**

coordinate index i_k is selected randomly as follows:

Assumption 2.3.2 (Block-coordinate selection rule) *At each iteration $k \geq 0$, the random variable i_k is generated from an independent and identically distributed discrete probability distribution such that* $\text{Prob}\,(i_k = i) = \mathsf{p}_i$ *where* $\mathsf{p}_i > 0$ *for* $i \in \{1, \ldots, d\}$ *and* $\sum_{i=1}^{d} \mathsf{p}_i = 1$.

Then, the i_k^{th} block-coordinate of the iterate x_k is updated using equation (2.3.1). Here, γ_k denotes the stepsize at iteration k and η_k denotes the regularization parameter at iteration k. We note that these sequences are updated iteratively. Here, we incorporate the information of the mapping F and the subgradient mapping $\tilde{\nabla} f$ by employing an iterative regularization scheme. At each iteration, a projection operation is performed onto a randomly selected set X_{i_k}. We will show that the convergence and rate analysis of the proposed method mainly rely on the choices of $\{\gamma_k\}$ and $\{\eta_k\}$. Accordingly, one key research objective in this section is to develop suitable update rules for the two sequences so that we can establish the convergence and derive rate statements. To obtain the rate results, we employ an averaging step using the equations given by equation (2.3.2), where the sequence $\{\bar{x}_k\}$ is obtained as a weighted average of $\{x_k\}$. The averaging weights are characterized by the stepsize γ_k and a scalar $r \in \mathbb{R}$. Note that in γ_k^r, the scalar r denotes the exponent. It will be shown that the rate results can be provided when $0 \leq r < 1$ (cf. Theorem 2.4.1).

Remark 2.3.2 Importantly, unlike Algorithm 1, Algorithm 2 is a single-timescale scheme that does not require solving any inner-level VI problem. In particular, the update rule given by step equation (2.3.1) mainly requires evaluations of random blocks of the mappings F and $\tilde{\nabla} f$. For this reason, step equation (2.3.1) is computationally more efficient than step 3 in Algorithm 1.

2.3.1 Preliminaries

In the following, we provide some definitions and preliminary results that will be used to analyze the convergence of Algorithm 2.

Definition 2.3.1 (Distance function) *For any $x, y \in \mathbb{R}^n$, function $\mathcal{D}(x, y)$ is defined as $\mathcal{D}(x, y) \triangleq \sum_{i=1}^{d} \mathsf{p}_i^{-1} \left\| x^{(i)} - y^{(i)} \right\|^2$, where p_i is given by Assumption 2.3.2.*

Remark 2.3.3 Under Assumption 2.3.2, we can relate the distance function $\mathcal{D}$ with the ℓ_2–norm as follows: $\mathsf{p}_{min}\mathcal{D}(x, y) \leq \|x - y\|^2 \leq \mathsf{p}_{max}\mathcal{D}(x, y)$ for all $x, y \in \mathbb{R}^n$, where $\mathsf{p}_{min} \triangleq \min_{1 \leq i \leq d}\{\mathsf{p}_i\}$ and $\mathsf{p}_{max} \triangleq \max_{1 \leq i \leq d}\{\mathsf{p}_i\}$.

One of the main challenges in the convergence analysis of computational methods for solving VI problems lies in the lack of access to a standard metric to quantify the quality of the solution iterates. This is in contrast with solving the standard optimization problems where the objective function can serve as an immediate performance metric for the underlying algorithm. Addressing this challenge in the literature of VI problems has led to the study of so-called *gap functions* (cf. [30, 95]). Of these, in the analysis of this section, we use the dual gap function defined as follows:

Definition 2.3.2 (The dual gap function [61]**)** *Let a nonempty closed set $X \subseteq \mathbb{R}^n$ and a mapping $F : X \to \mathbb{R}^n$ be given. Then, for any $x \in X$, the dual gap function* $\text{GAP} : X \to \mathbb{R} \cup \{+\infty\}$ *is defined as* $\text{GAP}(x) \triangleq \sup_{y \in X} F(y)^T(x - y)$.

Remark 2.3.4 When $X \neq \emptyset$, Definition 2.3.2 implies that the dual gap function is nonnegative over X. It is also known that when F is continuous and monotone and the set X is closed and convex, $\mathrm{GAP}(x^*) = 0$ if and only if $x^* \in \mathrm{SOL}(X, F)$ (cf. [45]). Thus, we conclude that under Assumption 2.3.1, the dual gap function is well-defined.

Definition 2.3.3 (Regularized mapping) *Given a vector $x \in X$, a subgradient $\tilde{\nabla} f(x) \in \partial f(x)$, and an integer $k \geq 0$, the regularized mapping $G_k : X \to \mathbb{R}$ is defined as $G_k(x) \triangleq F(x) + \eta_k \tilde{\nabla} f(x)$. The i^{th} block-coordinate of G_k is denoted by $G_{k,i}$.*

Definition 2.3.4 (History of the method) *Throughout, we let the history of the algorithm to be denoted by $\mathcal{F}_k \triangleq \{x_0, i_0, i_1, \ldots, i_{k-1}\}$ for $k \geq 1$, with $\mathcal{F}_0 \triangleq \{x_0\}$.*

Next, we show that $\bar{x}_k$ generated by Algorithm 2 is a well-defined weighted average.

Lemma 2.3.1 (Weighted averaging) *Let $\{\bar{x}_k\}$ be generated by Algorithm 2. Let us define the weights $\lambda_{k,N} \triangleq \frac{\gamma_k^r}{\sum_{j=0}^N \gamma_j^r}$ for $k \in \{0, \ldots, N\}$ and $N \geq 0$. Then, for any $N \geq 0$, we have $\bar{x}_N = \sum_{k=0}^N \lambda_{k,N} x_k$. Also, when X is a convex set, we have $\bar{x}_N \in X$.*

Proof. See Appendix A.3. ■

In the following, we define two terms that characterize the error between the true maps with their randomized block variants.

Definition 2.3.5 (Randomized block error terms) *Let $\mathbf{U}_i \in \mathbb{R}^{n \times n_i}$ for $i = 1, \ldots, d$ be the collection of matrices such that $\mathbf{I}_n = [\mathbf{U}_1, \ldots, \mathbf{U}_d] \in \mathbb{R}^{n \times n}$. Consider the following definitions for $k \geq 0$*

$$\Delta_k \triangleq F(x_k) - \mathsf{p}_{i_k}^{-1} \mathbf{U}_{i_k} F_{i_k}(x_k), \qquad \delta_k \triangleq \tilde{\nabla} f(x_k) - \mathsf{p}_{i_k}^{-1} \mathbf{U}_{i_k} \tilde{\nabla}_{i_k} f(x_k). \tag{2.3.3}$$

Lemma 2.3.2 (Properties of Δ_k and δ_k) *Consider Definition 2.3.5. We have*
(a) $\mathsf{E}[\Delta_k \mid \mathcal{F}_k] = \mathsf{E}[\delta_k \mid \mathcal{F}_k] = 0$.
(b) $\mathsf{E}[\|\Delta_k\|^2 \mid \mathcal{F}_k] \leq \left(\mathsf{p}_{min}^{-1} - 1\right) C_F^2$ and $\mathsf{E}[\|\delta_k\|^2 \mid \mathcal{F}_k] \leq \left(\mathsf{p}_{min}^{-1} - 1\right) C_f^2$.

Proof. See Appendix A.4 ■

We will use the next result in deriving the suboptimality and infeasibility rate results.

Lemma 2.3.3 (Bounds on the harmonic series) *Let $0 \leq \alpha < 1$ be a given scalar. Then, for any integer $N \geq 2^{\frac{1}{1-\alpha}} - 1$, we have $\frac{(N+1)^{1-\alpha}}{2(1-\alpha)} \leq \sum_{k=0}^N \frac{1}{(k+1)^\alpha} \leq \frac{(N+1)^{1-\alpha}}{1-\alpha}$.*

Proof. See Appendix A.5 ■

2.4 Convergence Rate Analysis with Bounded Set X

In the following result, we derive an inequality that will be later used to construct bounds on the objective function value and the dual gap function at the averaged sequence generated by Algorithm 2.

Lemma 2.4.1 *Consider the sequence $\{x_k\}$ in Algorithm 2. Suppose $\{\gamma_k\}$ and $\{\eta_k\}$ are strictly positive sequences. Let Assumption 2.3.1 and Assumption 2.3.2 hold. Let the auxiliary sequence $\{u_k\} \subset X$ be defined as $u_{k+1} \triangleq \mathcal{P}_X\left(u_k - \gamma_k\left(\Delta_k + \eta_k \delta_k\right)\right)$, where $u_0 \in X$ is an arbitrary vector. Then, for all $y \in X$ and all $k \geq 0$, we have*

$$\begin{aligned}
&\gamma_k^r F(y)^T(x_k - y) + \gamma_k^r \eta_k \tilde{\nabla} f(x_k)^T(x_k - y) \leq \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_k, y\right) + \|u_k - y\|^2\right) \\
&- \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_{k+1}, y\right) + \|u_{k+1} - y\|^2\right) + \gamma_k^r (x_k - u_k)^T\left(\Delta_k + \eta_k \delta_k\right) \\
&+ \gamma_k^{r+1}\left(\|\Delta_k\|^2 + \eta_k^2 \|\delta_k\|^2\right) + 0.5 \mathsf{p}_{i_k}^{-1} \gamma_k^{r+1} \left\|G_{k, i_k}\left(x_k\right)\right\|^2 . \qquad (2.4.1)
\end{aligned}$$

Proof. Let $k \geq 1$ be fixed. From Definition 2.3.1 and equation (2.3.1), for any $y \in X$, we have

$$\mathcal{D}\left(x_{k+1}, y\right) = \mathsf{p}_{i_k}^{-1}\left\|x_{k+1}^{(i_k)} - y^{(i_k)}\right\|^2 + \sum_{i=1, i \neq i_k}^{d} \mathsf{p}_i^{-1}\left\|x_k^{(i)} - y^{(i)}\right\|^2 . \qquad (2.4.2)$$

Next, we find a bound on the term $\left\|x_{k+1}^{(i_k)} - y^{(i_k)}\right\|^2$. From the block structure of X, we have $y^{(i_k)} \in X_{i_k}$. Invoking the nonexpansiveness property of the projection mapping, the update rule equation (2.3.1), Definition 2.3.3, and the preceding relation, we obtain

$$\left\|x_{k+1}^{(i_k)} - y^{(i_k)}\right\|^2 \leq \left\|x_k^{(i_k)} - \gamma_k G_{k, i_k}\left(x_k\right) - y^{(i_k)}\right\|^2 .$$

Combining the preceding relation with equation (2.4.2), we obtain

$$\begin{aligned}
\mathcal{D}\left(x_{k+1}, y\right) \leq & \sum_{i=1, i \neq i_k}^{d} \mathsf{p}_i^{-1}\left\|x_k^{(i)} - y^{(i)}\right\|^2 + \mathsf{p}_{i_k}^{-1}\left\|x_k^{(i_k)} - y^{(i_k)}\right\|^2 \\
& - 2\,\mathsf{p}_{i_k}^{-1} \gamma_k\left(x_k^{(i_k)} - y^{(i_k)}\right)^T G_{k, i_k}\left(x_k\right) + \mathsf{p}_{i_k}^{-1} \gamma_k^2 \left\|G_{k, i_k}\left(x_k\right)\right\|^2 .
\end{aligned}$$

Invoking Definition 2.3.1 again, we obtain

$$\mathcal{D}\left(x_{k+1}, y\right) \leq \mathcal{D}\left(x_k, y\right) - 2\,\mathsf{p}_{i_k}^{-1} \gamma_k\left(x_k^{(i_k)} - y^{(i_k)}\right)^T G_{k, i_k}\left(x_k\right) + \mathsf{p}_{i_k}^{-1} \gamma_k^2 \left\|G_{k, i_k}\left(x_k\right)\right\|^2 . \qquad (2.4.3)$$

From Definition 2.3.5 and Definition 2.3.3, we can write

$$\begin{aligned}
&\mathsf{p}_{i_k}^{-1}\left(x_k^{(i_k)} - y^{(i_k)}\right)^T G_{k, i_k}\left(x_k\right) = \mathsf{p}_{i_k}^{-1}(x_k - y)^T\left(\mathbf{U}_{i_k} G_{k, i_k}\left(x_k\right)\right) \\
&= \mathsf{p}_{i_k}^{-1}(x_k - y)^T\left(\mathbf{U}_{i_k} F_{i_k}(x_k) + \eta_k \mathbf{U}_{i_k} \tilde{\nabla}_{i_k} f\left(x_k\right)\right) \\
&= (x_k - y)^T\left(F(x_k) - \Delta_k + \eta_k \tilde{\nabla} f(x_k) - \eta_k \delta_k\right) = (x_k - y)^T\left(G_k(x_k) - \Delta_k - \eta_k \delta_k\right).
\end{aligned}$$

Combining the preceding inequality and relation equation (2.4.3), we obtain

$$\mathcal{D}\left(x_{k+1}, y\right) \leq \mathcal{D}\left(x_k, y\right) - 2\gamma_k (x_k - y)^T\left(G_k(x_k) - \Delta_k - \eta_k \delta_k\right) + \mathsf{p}_{i_k}^{-1} \gamma_k^2 \left\|G_{k, i_k}\left(x_k\right)\right\|^2 . \qquad (2.4.4)$$

Consider the definition of the auxiliary sequence $\{u_k\}$ in Lemma 2.4.1. Invoking the nonexpansiveness property of the projection again, we can obtain

$$\begin{aligned}\|u_{k+1}-y\|^2 &\leq \|u_k-\gamma_k\left(\Delta_k+\eta_k\delta_k\right)-y\|^2\\ &= \|u_k-y\|^2-2\gamma_k(u_k-y)^T\left(\Delta_k+\eta_k\delta_k\right)+\gamma_k^2\|\Delta_k+\eta_k\delta_k\|^2.\end{aligned}$$

Thus, we have

$$\|u_{k+1}-y\|^2 \leq \|u_k-y\|^2-2\gamma_k(u_k-y)^T\left(\Delta_k+\eta_k\delta_k\right)+2\gamma_k^2\|\Delta_k\|^2+2\gamma_k^2\eta_k^2\|\delta_k\|^2.$$

Adding the preceding inequality and the inequality equation (2.4.4) together, we obtain

$$\begin{aligned}&2\gamma_k(x_k-y)^TG_k(x_k) \leq \left(\mathcal{D}\left(x_k,y\right)+\|u_k-y\|^2\right)-\left(\mathcal{D}\left(x_{k+1},y\right)+\|u_{k+1}-y\|^2\right)\\ &+2\gamma_k(x_k-u_k)^T\left(\Delta_k+\eta_k\delta_k\right)+2\gamma_k^2\left(\|\Delta_k\|^2+\eta_k^2\|\delta_k\|^2\right)+\mathsf{p}_{i_k}^{-1}\gamma_k^2\left\|G_{k,i_k}\left(x_k\right)\right\|^2. \qquad (2.4.5)\end{aligned}$$

From the monotonicity property of the mapping F and Definition 2.3.3, we have

$$(x_k-y)^TG_k(x_k) \geq (x_k-y)^TF(y)+\eta_k\tilde{\nabla}f(x_k)^T(x_k-y).$$

This provides a lower bound on the left-hand side of equation (2.4.5). The inequality equation (2.4.1) is obtained by substituting this bound in equation (2.4.5) and multiplying both sides by $\frac{\gamma_k^{r-1}}{2}$, where $r-1$ denotes the exponent in γ_k^{r-1}. ■

In the following, we develop upper bounds for suboptimality and infeasibility of the weighted average iterate generated by Algorithm 2. Both of these error bounds are characterized in terms of the stepsize and the regularization parameter.

Proposition 2.4.1 (Error bounds for Algorithm 2) *Let the sequence $\{\bar{x}_k\}$ be generated by Algorithm 2, where $0 \leq r < 1$. Suppose $\{\gamma_k\}$ and $\{\eta_k\}$ are strictly positive and nonincreasing sequences. Let Assumption 2.3.1 and Assumption 2.3.2 hold and assume that the set X is bounded, i.e., $\|x\| \leq M$ for all $x \in X$ and some $M > 0$.*
(a) Let x^* be an optimal solution to Problem (P_1)*. Then, for all $N \geq 1$*

$$\mathsf{E}[f\left(\bar{x}_N\right)]-f(x^*) \leq \frac{\frac{4M^2\gamma_N^{r-1}}{\eta_N}+\sum_{k=0}^N\eta_k^{-1}\gamma_k^{r+1}\left(C_F^2+\eta_k^2C_f^2\right)}{\mathsf{p}_{min}\sum_{k=0}^N\gamma_k^r}. \qquad (2.4.6)$$

(b) Consider the dual gap function in Definition 2.3.2. Then, for all $N \geq 1$

$$\mathsf{E}[\mathrm{GAP}\left(\bar{x}_N\right)] \leq \frac{4M^2\gamma_N^{r-1}+\sum_{k=0}^N\gamma_k^r\left(2\mathsf{p}_{min}\eta_kC_fM+\gamma_kC_F^2+\gamma_k\eta_k^2C_f^2\right)}{\mathsf{p}_{min}\sum_{k=0}^N\gamma_k^r}. \qquad (2.4.7)$$

Proof. We define the following terms for all $k \geq 0$, that appear in equation (2.4.1)

$$\begin{aligned}&\Theta_{k,1} \triangleq \gamma_k^r(x_k-u_k)^T\left(\Delta_k+\eta_k\delta_k\right), \quad \Theta_{k,2} \triangleq \gamma_k^{r+1}\left(\|\Delta_k\|^2+\eta_k^2\|\delta_k\|^2\right),\\ &\Theta_{k,3} \triangleq 0.5\mathsf{p}_{i_k}^{-1}\gamma_k^{r+1}\left\|G_{k,i_k}\left(x_k\right)\right\|^2. \qquad (2.4.8)\end{aligned}$$

Next, we estimate the expected values of these terms. Consider the notation of $\mathcal{F}_k$ given by Definition 2.3.4. Note that x_k is $\mathcal{F}_k$–measurable. Also, from the definition of u_k in Lemma

2.4.1, u_k is $\mathcal{F}_k$–measurable. Note, however, that $\Theta_{k,j}$ is $\mathcal{F}_{k+1}$–measurable for all $j \in \{1,2,3\}$. Taking these into account and using the total probability law, for any $k \geq 0$ and $j \in \{1,2,3\}$, we have $\mathsf{E}[\Theta_{k,j}] = \mathsf{E}_{\mathcal{F}_k}\left[\mathsf{E}_{i_k}\left[\Theta_{k,j} \mid \mathcal{F}_k\right]\right]$. From this relation and Lemma 2.3.2, we have for any $k \geq 0$

$$\mathsf{E}[\Theta_{k,1}] = 0, \qquad \mathsf{E}[\Theta_{k,2}] = \left(\mathsf{p}_{min}^{-1} - 1\right)\gamma_k^{r+1}\left(C_F^2 + \eta_k^2 C_f^2\right). \tag{2.4.9}$$

Also, using Definition 2.3.3 and the triangle inequality, we can write

$$\begin{aligned}
&\mathsf{E}_{i_k}\left[\Theta_{k,3} \mid \mathcal{F}_k\right] = \sum_{i=1}^{d} \mathsf{p}_i \left(0.5\mathsf{p}_i^{-1}\gamma_k^{r+1}\left\|G_{k,i}\left(x_k\right)\right\|^2\right) \\
&\leq \gamma_k^{r+1}\sum_{i=1}^{d}\left(\|F_i(x_k)\|^2 + \eta_k^2\|\tilde{\nabla}_i f(x_k)\|^2\right) = \gamma_k^{r+1}\|F(x_k)\|^2 + \eta_k^2\|\tilde{\nabla} f(x_k)\|^2.
\end{aligned}$$

From the preceding inequality, we obtain

$$\mathsf{E}[\Theta_{k,3}] \leq \gamma_k^{r+1}\left(C_F^2 + \eta_k^2 C_f^2\right). \tag{2.4.10}$$

We are now ready to show the inequalities equation (2.4.6) and equation (2.4.7) as follows: (a) Consider equation (2.4.1). From the definition of subgradients of the convex function f, we have that $f(x_k) - f(y) \leq \tilde{\nabla} f(x_k)^T(x_k - y)$. Thus, from equation (2.4.8) we obtain for any $y \in X$

$$\begin{aligned}
&\gamma_k^r F(y)^T(x_k - y) + \gamma_k^r \eta_k\left(f(x_k) - f(y)\right) \leq \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_k, y\right) + \|u_k - y\|^2\right) \\
&- \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_{k+1}, y\right) + \|u_{k+1} - y\|^2\right) + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}.
\end{aligned}$$

Let us substitute $y := x^*$, where x^* denotes an optimal solution to Problem (P_1). Note that x^* must be a feasible solution to Problem (P_1), i.e., $F(x^*)^T(x_k - x^*) \geq 0$. Thus, we obtain

$$\begin{aligned}
&\gamma_k^r \eta_k\left(f(x_k) - f(x^*)\right) \leq \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_k, x^*\right) + \|u_k - x^*\|^2\right) \\
&- \frac{\gamma_k^{r-1}}{2}\left(\mathcal{D}\left(x_{k+1}, x^*\right) + \|u_{k+1} - x^*\|^2\right) + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}.
\end{aligned} \tag{2.4.11}$$

Dividing both sides by η_k and adding and subtracting $\frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\left(\mathcal{D}\left(x_k, x^*\right) + \|u_k - x^*\|^2\right)$ in the right-hand side of equation (2.4.11), we obtain for $k \geq 1$

$$\begin{aligned}
&\gamma_k^r\left(f(x_k) - f(x^*)\right) \leq \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\left(\mathcal{D}\left(x_k, x^*\right) + \|u_k - x^*\|^2\right) \\
&- \frac{\gamma_k^{r-1}}{2\eta_k}\left(\mathcal{D}\left(x_{k+1}, x^*\right) + \|u_{k+1} - x^*\|^2\right) \\
&+ \frac{1}{2}\left(\frac{\gamma_k^{r-1}}{\eta_k} - \frac{\gamma_{k-1}^{r-1}}{\eta_{k-1}}\right)\left(\mathcal{D}\left(x_k, x^*\right) + \|u_k - x^*\|^2\right) + \eta_k^{-1}\left(\Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}\right).
\end{aligned} \tag{2.4.12}$$

Since $r-1<0$ and that $\{\gamma_k\}$ and $\{\eta_k\}$ are nonincreasing, we have $\frac{\gamma_k^{r-1}}{\eta_k}-\frac{\gamma_{k-1}^{r-1}}{\eta_{k-1}}\geq 0$. Also, from the boundedness of the set X, since x_k, x^*, and u_k belong to X, using Remark 2.3.3 and the triangle inequality, we have

$$\mathcal{D}\left(x_k,x^*\right)+\left\|u_k-x^*\right\|^2\leq \mathsf{p}_{min}^{-1}\left\|x_k-x^*\right\|^2+\left\|u_k-x^*\right\|^2\leq 4M^2\left(\mathsf{p}_{min}^{-1}+1\right)\leq\frac{8M^2}{\mathsf{p}_{min}}. \tag{2.4.13}$$

Summing over equation (2.4.12) from $k=1$ to N and using equation (2.4.13), we obtain

$$\begin{aligned}\sum_{k=1}^{N}\gamma_k^r\left(f(x_k)-f(x^*)\right)&\leq\frac{\gamma_0^{r-1}}{2\eta_0}\left(\mathcal{D}\left(x_1,x^*\right)+\left\|u_1-x^*\right\|^2\right)\\&+4M^2\mathsf{p}_{min}^{-1}\left(\frac{\gamma_N^{r-1}}{\eta_N}-\frac{\gamma_0^{r-1}}{\eta_0}\right)+\sum_{k=1}^{N}\eta_k^{-1}\left(\Theta_{k,1}+\Theta_{k,2}+\Theta_{k,3}\right),\end{aligned}$$

where we drop the nonpositive term. From relation equation (2.4.11) when $k=0$, we have

$$\begin{aligned}\gamma_0^r\left(f(x_0)-f(x^*)\right)&\leq\frac{\gamma_0^{r-1}}{2\eta_0}\left(\mathcal{D}\left(x_0,x^*\right)+\left\|u_0-x^*\right\|^2\right)\\&\quad-\frac{\gamma_0^{r-1}}{2\eta_0}\left(\mathcal{D}\left(x_1,x^*\right)+\left\|u_1-x^*\right\|^2\right)+\eta_0^{-1}\left(\Theta_{0,1}+\Theta_{0,2}+\Theta_{0,3}\right).\end{aligned}$$

Adding the last two inequalities, multiplying and dividing the left-hand side by $\sum_{k=0}^{N}\gamma_k^r$, and then, invoking Lemma 2.3.1 and convexity of f, we obtain

$$\begin{aligned}\left(\sum_{k=0}^{N}\gamma_k^r\right)\left(f\left(\bar{x}_N\right)-f(x^*)\right)&\leq\frac{\gamma_0^{r-1}}{2\eta_0}\left(\mathcal{D}\left(x_0,x^*\right)+\left\|u_0-x^*\right\|^2\right)\\&+4M^2\mathsf{p}_{min}^{-1}\left(\frac{\gamma_N^{r-1}}{\eta_N}-\frac{\gamma_0^{r-1}}{\eta_0}\right)+\sum_{k=0}^{N}\eta_k^{-1}\left(\Theta_{k,1}+\Theta_{k,2}+\Theta_{k,3}\right).\end{aligned}$$

Taking the expectation on both sides and invoking equation (2.4.13), we obtain

$$\mathsf{E}[f\left(\bar{x}_N\right)]-f(x^*)\leq\frac{\frac{4M^2\mathsf{p}_{min}^{-1}\gamma_N^{r-1}}{\eta_N}+\sum_{k=0}^{N}\eta_k^{-1}\mathsf{E}[\Theta_{k,1}+\Theta_{k,2}+\Theta_{k,3}]}{\sum_{k=0}^{N}\gamma_k^r}.$$

From the relations equation (2.4.9) and equation (2.4.10), we obtain

$$\mathsf{E}[f\left(\bar{x}_N\right)]-f(x^*)\leq\frac{\frac{4M^2\mathsf{p}_{min}^{-1}\gamma_N^{r-1}}{\eta_N}+\sum_{k=0}^{N}\eta_k^{-1}\left(\mathsf{p}_{min}^{-1}\gamma_k^{r+1}\left(C_F^2+\eta_k^2C_f^2\right)\right)}{\sum_{k=0}^{N}\gamma_k^r},$$

which implies the inequality equation (2.4.6).

(b) From the Cauchy-Schwarz inequality, the definitions of C_f and M, and the triangle inequality, we have $\tilde{\nabla} f(x_k)^T (y - x_k) \leq \left\| \tilde{\nabla} f(x_k) \right\| \|x_k - y\| \leq 2C_f M$. Adding the preceding inequality with the relation equation (2.4.1), from equation (2.4.8) we obtain

$$\begin{aligned}
&\gamma_k^r F(y)^T (x_k - y) \leq \frac{\gamma_k^{r-1}}{2} \left(\mathcal{D}(x_k, y) + \|u_k - y\|^2 \right) \\
&- \frac{\gamma_k^{r-1}}{2} \left(\mathcal{D}(x_{k+1}, y) + \|u_{k+1} - y\|^2 \right) + 2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}. \qquad (2.4.14)
\end{aligned}$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2} \left(\mathcal{D}(x_k, y) + \|u_k - y\|^2 \right)$, we obtain

$$\begin{aligned}
&\gamma_k^r F(y)^T (x_k - y) \leq \frac{\gamma_{k-1}^{r-1}}{2} \left(\mathcal{D}(x_k, y) + \|u_k - y\|^2 \right) \\
&- \frac{\gamma_k^{r-1}}{2} \left(\mathcal{D}(x_{k+1}, y) + \|u_{k+1} - y\|^2 \right) + \frac{1}{2} \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1} \right) \left(\mathcal{D}(x_k, y) + \|u_k - y\|^2 \right) \\
&+ 2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}.
\end{aligned}$$

Substituting the bound given by equation (2.4.13) in the preceding relation, we obtain

$$\begin{aligned}
&\gamma_k^r F(y)^T (x_k - y) \leq \frac{\gamma_{k-1}^{r-1}}{2} \left(\mathcal{D}(x_k, y) + \|u_k - y\|^2 \right) \\
&- \frac{\gamma_k^{r-1}}{2} \left(\mathcal{D}(x_{k+1}, y) + \|u_{k+1} - y\|^2 \right) + 4M^2 \mathsf{p}_{min}^{-1} \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1} \right) \\
&+ 2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}.
\end{aligned}$$

Summing both sides from $k = 1$ to N, we obtain

$$\begin{aligned}
\sum_{k=1}^{N} \gamma_k^r F(y)^T (x_k - y) &\leq \frac{\gamma_0^{r-1}}{2} \left(\mathcal{D}(x_1, y) + \|u_1 - y\|^2 \right) + 4M^2 \mathsf{p}_{min}^{-1} \left(\gamma_N^{r-1} - \gamma_0^{r-1} \right) \\
&+ \sum_{k=1}^{N} \left(2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3} \right). \qquad (2.4.15)
\end{aligned}$$

Writing the inequality equation (2.4.14) for $k = 0$, we have

$$\begin{aligned}
\gamma_0^r F(y)^T (x_0 - y) &\leq \frac{\gamma_0^{r-1}}{2} \left(\mathcal{D}(x_0, y) + \|u_0 - y\|^2 \right) - \frac{\gamma_0^{r-1}}{2} \left(\mathcal{D}(x_1, y) + \|u_1 - y\|^2 \right) \\
&+ 2\gamma_0^r \eta_0 C_f M + \Theta_{0,1} + \Theta_{0,2} + \Theta_{0,3}. \qquad (2.4.16)
\end{aligned}$$

Adding equation (2.4.15) and equation (2.4.16) together, we obtain

$$\begin{aligned}
\sum_{k=0}^{N} \gamma_k^r F(y)^T (x_k - y) &\leq \frac{\gamma_0^{r-1}}{2} \left(\mathcal{D}(x_0, y) + \|u_0 - y\|^2 \right) + 4M^2 \mathsf{p}_{min}^{-1} \left(\gamma_N^{r-1} - \gamma_0^{r-1} \right) \\
&+ \sum_{k=0}^{N} \left(2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3} \right). \qquad (2.4.17)
\end{aligned}$$

Recalling $\bar{x}_N = \sum_{k=0}^{N} \lambda_{k,N} x_k$ in Lemma 2.3.1, applying the bound given by equation (2.4.13), and using the triangle inequality, we obtain

$$\left(\sum_{k=0}^{N} \gamma_k^r\right) F(y)^T (\bar{x}_N - y) \leq 4M^2 \mathsf{p}_{min}^{-1} \gamma_N^{r-1} + \sum_{k=0}^{N} \left(2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}\right).$$

Taking the supremum with respect to y over the set X from the left-hand side, invoking Definition 2.3.2, and then dividing both sides by $\sum_{k=0}^{N} \gamma_k^r$, we obtain

$$\mathrm{GAP}\left(\bar{x}_N\right) \leq \frac{4M^2 \mathsf{p}_{min}^{-1} \gamma_N^{r-1} + \sum_{k=0}^{N} \left(2\gamma_k^r \eta_k C_f M + \Theta_{k,1} + \Theta_{k,2} + \Theta_{k,3}\right)}{\sum_{k=0}^{N} \gamma_k^r}.$$

Taking the expectation on both sides, using the relations equation (2.4.9) and equation (2.4.10), and rearranging the terms, we obtain the inequality equation (2.4.7). ■

We are now ready to present the convergence rate results of the proposed method.

Theorem 2.4.1 (Convergence rate statements for Algorithm 2) *Consider Algorithm 2. Let Assumption 2.3.1 and Assumption 2.3.2 hold and assume that the set X is bounded such that $\|x\| \leq M$ for all $x \in X$ and some $M > 0$. Suppose for all $k \geq 0$, $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, where $\gamma_0 > 0$, $\eta_0 > 0$, and $0 < b < 0.5$. Then, for any $0 \leq r < 1$, the following results hold:*
(i) Let x^* be an optimal solution to the Problem (P_1). *Then, for all $N \geq 2^{\frac{2}{1-r}} - 1$*

$$\mathsf{E}[f\left(\bar{x}_N\right)] - f(x^*) \leq \frac{2-r}{\mathsf{p}_{min}\eta_0} \left(\frac{4M^2}{\gamma_0} + \frac{\gamma_0\left(C_F^2 + \eta_0^2 C_f^2\right)}{0.5 - 0.5r + b}\right) \frac{1}{(N+1)^{0.5-b}}. \tag{2.4.18}$$

(ii) Consider the dual gap function in Definition 2.3.2. Then, for all $N \geq 2^{\frac{2}{1-r}} - 1$

$$\mathsf{E}[\mathrm{GAP}\left(\bar{x}_N\right)] \leq \frac{2-r}{\mathsf{p}_{min}} \left(\frac{4M^2}{\gamma_0} + \frac{\gamma_0\left(C_F^2 + \eta_0^2 C_f^2\right)}{0.5 - 0.5r} + \frac{2\mathsf{p}_{min} C_f M \eta_0}{1 - 0.5r - b}\right) \frac{1}{(N+1)^b}. \tag{2.4.19}$$

Proof. Let us define the following terms:

$$\Lambda_{N,1} \triangleq \mathsf{p}_{min} \sum_{k=0}^{N} \gamma_k^r, \quad \Lambda_{N,2} \triangleq \frac{4M^2 \gamma_N^{r-1}}{\eta_N}, \quad \Lambda_{N,3} \triangleq \left(C_F^2 + \eta_0^2 C_f^2\right) \sum_{k=0}^{N} \eta_k^{-1} \gamma_k^{r+1},$$

$$\Lambda_{N,4} \triangleq 4M^2 \gamma_N^{r-1}, \quad \Lambda_{N,5} \triangleq \left(C_F^2 + \eta_0^2 C_f^2\right) \sum_{k=0}^{N} \gamma_k^{r+1}, \quad \Lambda_{N,6} \triangleq 2\mathsf{p}_{min} C_f M \sum_{k=0}^{N} \eta_k \gamma_k^r.$$

Note that from equation (2.4.6) and equation (2.4.7), we have

$$\mathsf{E}[f\left(\bar{x}_N\right)] - f(x^*) \leq \frac{\Lambda_{N,2} + \Lambda_{N,3}}{\Lambda_{N,1}}, \quad \mathsf{E}[\mathrm{GAP}\left(\bar{x}_N\right)] \leq \frac{\Lambda_{N,4} + \Lambda_{N,5} + \Lambda_{N,6}}{\Lambda_{N,1}}. \tag{2.4.20}$$

Next, we apply Lemma 2.3.3 to estimate the terms $\Lambda_{N,i}$. Substituting γ_k and η_k by their update rules, we obtain

$$\Lambda_{N,1} = \mathsf{p}_{min} \sum_{k=0}^{N} \frac{\gamma_0^r}{(k+1)^{0.5r}} \geq \frac{\mathsf{p}_{min}\gamma_0^r (N+1)^{1-0.5r}}{2(1-0.5r)},$$

$$\Lambda_{N,2} = \frac{4M^2(N+1)^{0.5(1-r)+b}}{\eta_0 \gamma_0^{1-r}}, \quad \Lambda_{N,4} = \frac{4M^2(N+1)^{0.5(1-r)}}{\gamma_0^{1-r}},$$

$$\Lambda_{N,3} = \sum_{k=0}^{N} \frac{\left(C_F^2 + \eta_0^2 C_f^2\right)\gamma_0^{r+1}}{\eta_0 (k+1)^{0.5(r+1)-b}} \leq \frac{\gamma_0^{r+1}\left(C_F^2 + \eta_0^2 C_f^2\right)(N+1)^{1-0.5(r+1)+b}}{\eta_0(1-0.5(r+1)+b)},$$

$$\Lambda_{N,5} = \left(C_F^2 + \eta_0^2 C_f^2\right) \sum_{k=0}^{N} \frac{\gamma_0^{r+1}}{(k+1)^{0.5(r+1)}} \leq \frac{\left(C_F^2 + \eta_0^2 C_f^2\right)\gamma_0^{r+1}(N+1)^{1-0.5(r+1)}}{1-0.5(r+1)},$$

$$\Lambda_{N,6} = 2\mathsf{p}_{min} C_f M \eta_0 \gamma_0^r \sum_{k=0}^{N} \frac{1}{(k+1)^{0.5r+b}} \leq \frac{2\mathsf{p}_{min} C_f M \eta_0 \gamma_0^r (N+1)^{1-0.5r-b}}{1-0.5r-b}.$$

For these inequalities to hold, we need to ensure that the conditions of Lemma 2.3.3 are met. Accordingly, we must have $0 \leq 0.5r < 1$, $0 \leq 0.5(r+1) - b < 1$, $0 \leq 0.5r + b < 1$, and $0 \leq 0.5(r+1) < 1$. These relations hold because $0 \leq r < 1$ and $0 < b < 0.5$. Another set of conditions when applying Lemma 2.3.3 includes $N \geq \max\left\{2^{1/(1-0.5r)}, 2^{1/(1-0.5(r+1)+b)}, 2^{1/(1-0.5r-b)}, 2^{1/(1-0.5(r+1))}\right\} - 1$. This relation is indeed satisfied as a consequence of $N \geq 2^{\frac{2}{1-r}} - 1$, $0 < b < 0.5$, and $0 \leq r < 1$. We conclude that all the necessary conditions for applying Lemma 2.3.3 and obtaining the aforementioned bounds for the terms $\Lambda_{N,i}$ are satisfied. To show that the inequalities equation (2.4.18) and equation (2.4.19) hold, it suffices to substitute the preceding bounds on the terms $\Lambda_{N,i}$ into the two inequalities given by equation (2.4.20). The details are as follows

$$\mathsf{E}[f(\bar{x}_N)] - f(x^*) \leq \frac{\Lambda_{N,2} + \Lambda_{N,3}}{\Lambda_{N,1}} = \frac{2-r}{\mathsf{p}_{min}\gamma_0^r (N+1)^{1-0.5r}} \left(\frac{4M^2(N+1)^{0.5-0.5r+b}}{\eta_0 \gamma_0^{1-r}} \right.$$
$$\left. + \left(\frac{\gamma_0^{r+1}}{\eta_0}\right) \frac{\left(C_F^2 + \eta_0^2 C_f^2\right)(N+1)^{0.5-0.5r+b}}{0.5-0.5r+b} \right).$$

The inequality equation (2.4.18) is obtained by rearranging the terms in the preceding relation.

$$\mathsf{E}[\mathrm{GAP}(\bar{x}_N)] \leq \frac{\Lambda_{N,4} + \Lambda_{N,5} + \Lambda_{N,6}}{\Lambda_{N,1}} \leq \frac{2-r}{\mathsf{p}_{min}\gamma_0^r (N+1)^{1-0.5r}} \left(\frac{4M^2(N+1)^{0.5-0.5r}}{\gamma_0^{1-r}} \right.$$
$$\left. + \frac{\left(C_F^2 + \eta_0^2 C_f^2\right)\gamma_0^{r+1}(N+1)^{0.5-0.5r}}{0.5-0.5r} + \frac{2\mathsf{p}_{min} C_f M \eta_0 \gamma_0^r (N+1)^{1-0.5r-b}}{1-0.5r-b} \right).$$

Then, equation (2.4.19) can be obtained by rearranging the terms in the preceding inequality. ■

Remark 2.4.1 (Iteration complexity of Algorithm 2) As an immediate result from Theorem 2.4.1, choosing $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{\sqrt[4]{k+1}}$, we obtain

$$\mathsf{E}[f(\bar{x}_N) - f(x^*)] = \mathsf{E}[\text{GAP}(\bar{x}_N)] = \mathcal{O}\left(\frac{1}{\sqrt[4]{N}}\right).$$

This implies that Algorithm 2 achieves an iteration complexity of $\mathcal{O}(\epsilon^{-4})$ in solving Problem (P_1), where $\epsilon > 0$ denotes the expected tolerance in both of the suboptimality and infeasibility metrics.

The rate statements derived in Theorem 2.4.1 are in a mean sense. In the following, we consider a deterministic variant of Algorithm 2 where we suppress the randomized block-coordinate scheme. The outline of this deterministic method is presented by Algorithm 3. In Corollary 2.4.1, we show that non-asymptotic deterministic rate statements can be derived for Algorithm 3.

Algorithm 3 Averaged Iteratively Regularized Gradient

1: **Input:** An arbitrary initial point $x_0 \in X$, $\bar{x}_0 := x_0$, initial stepsize $\gamma_0 > 0$, initial regularization parameter $\eta_0 > 0$, a scalar $0 \leq r < 1$, and $S_0 := \gamma_0^r$.
2: **for** k = 0, 1, ... **do**
3: Evaluate $F(x_k)$ and $\tilde{\nabla} f(x_k)$ where $\tilde{\nabla} f(x_k) \in \partial f(x_k)$.
4: For all $i \in \{1, \ldots, d\}$, do the following updates:

$$x_{k+1}^{(i)} := \mathcal{P}_{X_i}\Big(x_k^{(i)} - \gamma_k \Big(F_i(x_k) + \eta_k \tilde{\nabla}_i f(x_k)\Big)\Big). \tag{2.4.21}$$

5: Obtain γ_{k+1} and η_{k+1} (cf. Corollary 2.4.1 for the update rules).
6: Update the averaged iterate $\bar{x}_k$ as follows:

$$S_{k+1} := S_k + \gamma_{k+1}^r, \quad \bar{x}_{k+1} := \frac{S_k \bar{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}}. \tag{2.4.22}$$

7: **end for**

Corollary 2.4.1 (Convergence rate statements for Algorithm 3) *Consider Algorithm 3. Let Assumption 2.3.1 hold and assume that the set X is bounded such that $\|x\| \leq M$ for all $x \in X$ and some $M > 0$. Suppose for $k \geq 0$, $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{(k+1)^b}$, where $\gamma_0 > 0$, $\eta_0 > 0$, and $0 < b < 0.5$. Then, for any $0 \leq r < 1$, the following results hold:*
(i) Let x^* be an optimal solution to the Problem (P_1). Then, for all $N \geq 2^{\frac{2}{1-r}} - 1$

$$f(\bar{x}_N) - f(x^*) \leq \frac{2-r}{\eta_0}\left(\frac{4M^2}{\gamma_0} + \frac{\gamma_0\left(C_F^2 + \eta_0^2 C_f^2\right)}{0.5 - 0.5r + b}\right)\frac{1}{(N+1)^{0.5-b}}. \tag{2.4.23}$$

(ii) Consider the dual gap function in Definition 2.3.2. Then, for all $N \geq 2^{\frac{2}{1-r}} - 1$

$$\text{GAP}(\bar{x}_N) \leq (2-r)\left(\frac{4M^2}{\gamma_0} + \frac{\gamma_0\left(C_F^2 + \eta_0^2 C_f^2\right)}{0.5 - 0.5r} + \frac{2C_f M \eta_0}{1 - 0.5r - b}\right)\frac{1}{(N+1)^b}. \tag{2.4.24}$$

Proof. See Appendix A.6. ■

2.5 Convergence Rate Analysis with Unbounded Set X

The convergence and rate statements provided by Theorem 2.4.1 require the set X to be bounded. We, however, note that in some applications, e.g., in the models presented in Problem (1.2.2) and Problem (1.2.4), this assumption may not hold. Accordingly, in this section, our aim is to analyze the convergence of Algorithm 2 when X is unbounded. To this end, we consider the following main assumption.

Assumption 2.5.1 *Consider Problem* (P_1) *under the following conditions:*
(a) The set X_i is nonempty, closed, and convex for all $i = 1, \ldots, d$.
(b) The function f is continuously differentiable and μ_f–strongly convex over X.
(c) The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and monotone over X.
(d) The solution set $\mathrm{SOL}(X, F)$ *is nonempty.*

Remark 2.5.1 (Existence and uniqueness of the optimal solution) Under Assumption 2.5.1, the constraint set of Problem (P_1), i.e., $\mathrm{SOL}(X, F)$, is nonempty, closed, and convex. The convexity of this set is implied by Theorem 2.3.5 in [30] and its closedness property is obtained by the continuity of the mapping F and closedness of the set X. Because in the Problem (P_1), the objective function f is strongly convex and that the constraint set is nonempty, closed, and convex, we conclude from Proposition 1.1.2 in [14] that the Problem (P_1) has a unique optimal solution. Throughout this section, we let x^* denote this unique optimal solution.

2.5.1 Preliminaries

In this part, we provide preliminary results that will be used in the convergence analysis. We begin by defining a generalized variant of the Tikhonov trajectory that is associated with the problem of interest in this chapter.

Definition 2.5.1 (Tikhonov trajectory) *Consider the Problem* (P_1) *under Assumption 2.5.1. Let $\{\eta_k\}$ be a sequence of strictly positive scalars for all $k \geq 0$, and $x^*_{\eta_k} \in X$ denote the unique solution to the regularized variational inequality problem given by* $\mathrm{VI}\,(X, F + \eta_k \nabla f)$. *Then, the sequence $\left\{x^*_{\eta_k}\right\}$ is defined as the Tikhonov trajectory associated with the Problem* (P_1).

Remark 2.5.2 The uniqueness of the solution of $\mathrm{VI}\,(X, F + \eta_k \nabla f)$ in Definition 2.5.1 is due to the strong monotonicity of the mapping $F + \eta_k \nabla f$ and closedness and convexity of the set X (see Theorem 2.3.3 in [30]). Definition 2.5.1 generalizes the notion of Tikhonov trajectory provided in [30] such that $x^*_{\eta_k}$ is a solution to the regularized problem $\mathrm{VI}\,(X, F + \eta_k \mathbf{I}_n)$. This is indeed the special case where we choose $f(x) := \frac{1}{2}\|x\|^2$ in Definition 2.5.1.

To analyze the convergence, we utilize the properties of the Tikhonov trajectory. The following result ascertains the asymptotic convergence of this trajectory to the optimal solution of the Problem (P_1). It also provides an upper bound on the error between any two successive vectors of the trajectory.

Lemma 2.5.1 *Consider Definition 2.5.1 and let Assumption 2.5.1 hold. Let* $\{\eta_k\}$ *be a sequence such that* $\lim_{k\to\infty}\eta_k = 0$ *and* $\eta_k > 0$ *for all* $k \geq 0$. *Then we have*
(a) The Tikhonov trajectory $\{x^*_{\eta_k}\}$ *converges to a unique limit point, that is* x^*.
(b) There exists $\bar{C}_f > 0$ *such that* $\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\| \leq \frac{\bar{C}_f}{\mu_f}\left|1 - \frac{\eta_{k-1}}{\eta_k}\right|$ *for all* $k \geq 1$.

Proof. See Appendix A.7. ■

The following lemmas will be employed to establish the asymptotic convergence result.

Lemma 2.5.2 (Theorem 6, page 75 in [54]**)** *Let* $\{u_t\} \subset \mathbb{R}^n$ *denote a sequence of vectors where* $\lim_{t\to\infty} u_t = \hat{u}$. *Also, let* $\{\alpha_k\}$ *denote a sequence of strictly positive scalars such that* $\sum_{k=0}^{\infty}\alpha_k = \infty$. *Suppose* $v_k \in \mathbb{R}^n$ *is defined by* $v_k \triangleq \frac{\sum_{t=0}^{k}\alpha_t u_t}{\sum_{t=0}^{k}\alpha_t}$ *for all* $k \geq 0$. *Then,* $\lim\limits_{k\to\infty} v_k = \hat{u}$.

Lemma 2.5.3 (Lemma 10, page 49 in [69]**)** *Let* $\{v_k\}$ *be a sequence of nonnegative random variables, where* $\mathsf{E}[v_0] < \infty$, *and let* $\{\alpha_k\}$ *and* $\{\beta_k\}$ *be deterministic scalar sequences such that* $\mathsf{E}[v_{k+1}|v_0,\ldots,v_k] \leq (1-\alpha_k)v_k + \beta_k$ *for all* $k \geq 0$, $0 \leq \alpha_k \leq 1$, $\beta_k \geq 0$, $\sum_{k=0}^{\infty}\alpha_k = \infty$, $\sum_{k=0}^{\infty}\beta_k < \infty$, *and* $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k} = 0$. *Then,* $v_k \to 0$ *almost surely and* $\lim\limits_{k\to\infty}\mathsf{E}[v_k] = 0$.

2.5.2 Convergence Analysis

As a key step toward performing the convergence analysis for Algorithm 2 when the set X is unbounded, next we derive a recursive inequality for the distance between the generated sequence $\{x_k\}$ by the algorithm and the Tikhonov trajectory $\{x^*_{\eta_k}\}$. To this end, we first make the following assumption.

Assumption 2.5.2 *Consider the Problem* (P_1) *under the following assumptions:*
(a) There exist nonnegative scalars L_F *and* B_F *such that for all* $x, y \in X$

$$\|F(x) - F(y)\|^2 \leq L_F^2\|x-y\|^2 + B_F.$$

(b) The gradient mapping ∇f *is Lipschitz with parameter* $L_f > 0$.

Remark 2.5.3 By allowing L_F or B_F to be zero, Assumption 2.5.2 provides a unifying structure for considering both smooth and nonsmooth cases. In particular, when $L_F = 0$, part (a) refers to a bounded, but possibly non-Lipschitzian mapping F. Also, when $B_F = 0$, part (a) refers to a Lipschitzian, but possibly unbounded mapping F.

The following recursive relation will play a key role in establishing the convergence.

Lemma 2.5.4 (A recursive error bound for Algorithm 2) *Consider the sequence* $\{x_k\}$ *in Algorithm 2. Let Assumption 2.5.1, Assumption 2.3.2, and Assumption 2.5.2 hold. Suppose* $\{\gamma_k\}$ *and* $\{\eta_k\}$ *are nonincreasing and strictly positive where* $\lim_{k\to\infty}\eta_k = 0$ *and* $\frac{\gamma_k}{\eta_k} \leq \frac{\mu_f \mathsf{p}_{min}}{2\mathsf{p}_{max}\left(L_F^2+\eta_0^2 L_f^2\right)}$ *for all* $k \geq 0$. *Then, for all* $k \geq 1$, *we have*

$$\begin{aligned}\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq & \frac{\mathsf{p}_{max}}{\mathsf{p}_{min}}\left(1 - \frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}\right)\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) \\ & + \frac{\bar{C}_f^2\left(\mu_f\gamma_0\eta_0 + 2/\mathsf{p}_{min}\right)}{\mu_f^3\,\mathsf{p}_{min}\gamma_k\eta_k}\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 + 2\gamma_k^2 B_F. \qquad (2.5.1)\end{aligned}$$

Proof. From Definition 2.3.1, we have

$$\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) = \mathsf{p}_{i_k}^{-1} \left\| x_{k+1}^{(i_k)} - x_{\eta_k}^{*(i_k)} \right\|^2 + \sum_{i=1,\, i\neq i_k}^{d} \mathsf{p}_i^{-1} \left\| x_k^{(i)} - x_{\eta_k}^{*(i)} \right\|^2. \tag{2.5.2}$$

Next, we find a bound on the term $\left\| x_{k+1}^{(i_k)} - x_{\eta_k}^{*(i_k)} \right\|^2$. From the properties of the natural map (cf. Proposition 1.5.8 in [30]), Definition 2.3.3, and that $x^*_{\eta_k} \in X$, we have $x^*_{\eta_k} = \mathcal{P}_X\left(x^*_{\eta_k} - \gamma_k G_k\left(x^*_{\eta_k}\right)\right)$. From Assumption 2.5.1(a) and that $x^*_{\eta_k} \in \mathrm{SOL}\left(X, G_k\right) \subseteq X$, we have $x_{\eta_k}^{*(i_k)} \in X_{i_k}$. Invoking the nonexpansiveness property of the projection mapping, equation (2.3.1), and the preceding relation, we obtain

$$\left\| x_{k+1}^{(i_k)} - x_{\eta_k}^{*(i_k)} \right\|^2 \leq \left\| x_k^{(i_k)} - \gamma_k G_{k,i_k}\left(x_k\right) - x_{\eta_k}^{*(i_k)} + \gamma_k G_{k,i_k}\left(x^*_{\eta_k}\right) \right\|^2.$$

Combining the preceding relation with equation (2.5.2), we obtain

$$\begin{aligned}
\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) \leq & \sum_{i=1,\, i\neq i_k}^{d} \mathsf{p}_i^{-1} \left\| x_k^{(i)} - x_{\eta_k}^{*(i)} \right\|^2 + \mathsf{p}_{i_k}^{-1} \left\| x_k^{(i_k)} - x_{\eta_k}^{*(i_k)} \right\|^2 \\
& - 2\,\mathsf{p}_{i_k}^{-1}\gamma_k \left(x_k^{(i_k)} - x_{\eta_k}^{*(i_k)} \right)^T \left(G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right) \\
& + \mathsf{p}_{i_k}^{-1}\gamma_k^2 \left\| G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right\|^2.
\end{aligned}$$

Invoking Definition 2.3.1, from the preceding relation we obtain

$$\begin{aligned}
\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) \leq & \mathcal{D}\left(x_k, x^*_{\eta_k}\right) - 2\,\mathsf{p}_{i_k}^{-1}\gamma_k \left(x_k^{(i_k)} - x_{\eta_k}^{*(i_k)} \right)^T \left(G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right) \\
& + \mathsf{p}_{i_k}^{-1}\gamma_k^2 \left\| G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right\|^2.
\end{aligned}$$

Taking the conditional expectation from the both sides of preceding relation and noting that $\mathcal{D}\left(x_k, x^*_{\eta_k}\right)$ is $\mathcal{F}_k$–measurable, we obtain the following inequality

$$\begin{aligned}
\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) | \mathcal{F}_k\right] \leq & \mathcal{D}\left(x_k, x^*_{\eta_k}\right) + \gamma_k^2 \mathsf{E}\left[\mathsf{p}_{i_k}^{-1} \left\| G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right\|^2 \right] \\
& - 2\gamma_k \mathsf{E}\left[\mathsf{p}_{i_k}^{-1} \left(x_k^{(i_k)} - x_{\eta_k}^{*(i_k)} \right)^T \left(G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right) \right].
\end{aligned} \tag{2.5.3}$$

Next, we estimate the second and third expectations in the preceding relation

$$\begin{aligned}
& \mathsf{E}\left[\mathsf{p}_{i_k}^{-1} \left(x_k^{(i_k)} - x_{\eta_k}^{*(i_k)} \right)^T \left(G_{k,i_k}\left(x_k\right) - G_{k,i_k}\left(x^*_{\eta_k}\right) \right) \right] \\
& = \sum_{i=1}^{d} \mathsf{p}_i \mathsf{p}_i^{-1} \left(x_k^{(i)} - x_{\eta_k}^{*(i)} \right)^T \left(G_{k,i}\left(x_k\right) - G_{k,i}\left(x^*_{\eta_k}\right) \right) \\
& = \left(x_k - x^*_{\eta_k} \right)^T \left(G_k(x_k) - G_k\left(x^*_{\eta_k}\right) \right).
\end{aligned} \tag{2.5.4}$$

We can also write

$$\begin{aligned}&\mathsf{E}\Big[\mathsf{p}_{i_k}^{-1}\left\|G_{k,i_k}\left(x_k\right)-G_{k,i_k}\left(x_{\eta_k}^*\right)\right\|^2\Big]\\&=\sum_{i=1}^{d}\mathsf{p}_i\mathsf{p}_i^{-1}\left\|G_{k,i}\left(x_k\right)-G_{k,i}\left(x_{\eta_k}^*\right)\right\|^2=\left\|G_k\left(x_k\right)-G_k\left(x_{\eta_k}^*\right)\right\|^2.\end{aligned}\tag{2.5.5}$$

From Assumption 2.5.2, taking into account that G_k is $(\eta_k\mu_f)$–strongly monotone, and combining equation (2.5.3), equation (2.5.4), and equation (2.5.5) we obtain

$$\begin{aligned}\mathsf{E}\big[\mathcal{D}\left(x_{k+1},x_{\eta_k}^*\right)|\mathcal{F}_k\big]\leq\mathcal{D}\left(x_k,x_{\eta_k}^*\right)&-2\mu_f\gamma_k\eta_k\left\|x_k-x_{\eta_k}^*\right\|^2\\&+2\gamma_k^2\left(\left(L_F^2+\eta_k^2L_f^2\right)\left\|x_k-x_{\eta_k}^*\right\|^2+B_F\right).\end{aligned}$$

From Remark 2.3.3 and that $\{\eta_k\}$ is a nonincreasing sequence, we obtain

$$\begin{aligned}\mathsf{E}\big[\mathcal{D}\left(x_{k+1},x_{\eta_k}^*\right)|\mathcal{F}_k\big]\leq&\left(1-2\mu_f\gamma_k\eta_k\mathsf{p}_{min}+2\gamma_k^2\mathsf{p}_{max}\left(L_F^2+{\eta_0}^2L_f^2\right)\right)\mathcal{D}\left(x_k,x_{\eta_k}^*\right)\\&+2\gamma_k^2B_F.\end{aligned}$$

From the assumption $\gamma_k\leq\frac{\mu_f\eta_k\mathsf{p}_{min}}{2\mathsf{p}_{max}\left(L_F^2+\eta_0^2L_f^2\right)}$ and the preceding inequality, we obtain

$$\mathsf{E}\big[\mathcal{D}\left(x_{k+1},x_{\eta_k}^*\right)|\mathcal{F}_k\big]\leq\left(1-\mu_f\gamma_k\eta_k\mathsf{p}_{min}\right)\mathcal{D}\left(x_k,x_{\eta_k}^*\right)+2\gamma_k^2B_F.\tag{2.5.6}$$

The preceding relation is not yet fully recursive as the term $x_{\eta_k}^*$ on the right-hand side must change to $x_{\eta_{k-1}}^*$. Next, we find an upper bound for $\mathcal{D}\left(x_k,x_{\eta_k}^*\right)$ in terms of $\mathcal{D}\left(x_k,x_{\eta_{k-1}}^*\right)$. Note that we have $\|u+v\|^2\leq(1+\theta)\|u\|^2+\left(1+\frac{1}{\theta}\right)\|v\|^2$ for any vectors $u,v\in\mathbb{R}^n$ and $\theta>0$. Utilizing this inequality, by setting $u:=x_k-x_{\eta_{k-1}}^*$, $v:=x_{\eta_{k-1}}^*-x_{\eta_k}^*$, and $\theta:=\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}$ we obtain

$$\begin{aligned}\left\|x_k-x_{\eta_k}^*\right\|^2\leq&\left(1+\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}\right)\left\|x_k-x_{\eta_{k-1}}^*\right\|^2\\&+\left(1+\frac{2}{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}\right)\left\|x_{\eta_{k-1}}^*-x_{\eta_k}^*\right\|^2.\end{aligned}$$

Together with Lemma 2.5.1(b) and Remark 2.3.3, we have

$$\begin{aligned}\mathsf{p}_{min}\,\mathcal{D}\left(x_k,x_{\eta_k}^*\right)\leq&\left(1+\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}\right)\mathsf{p}_{max}\,\mathcal{D}\left(x_k,x_{\eta_{k-1}}^*\right)\\&+\left(1+\frac{2}{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}\right)\frac{\bar{C}_f^2}{\mu_f^2}\left(1-\frac{\eta_{k-1}}{\eta_k}\right)^2.\end{aligned}$$

Dividing both sides by p_{min} and substituting this in equation (2.5.6), we obtain

$$\begin{aligned}\mathsf{E}\big[\mathcal{D}\left(x_{k+1},x_{\eta_k}^*\right)|\mathcal{F}_k\big]\leq&\frac{\mathsf{p}_{max}}{\mathsf{p}_{min}}\left(1-\gamma_k\eta_k\mu_f\mathsf{p}_{min}\right)\left(1+\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}\right)\mathcal{D}\left(x_k,x_{\eta_{k-1}}^*\right)\\&+\frac{\bar{C}_f^2}{\mu_f^2\,\mathsf{p}_{min}}\left(1+\frac{2}{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}\right)\left(1-\frac{\eta_{k-1}}{\eta_k}\right)^2+2\gamma_k^2B_F.\end{aligned}$$

equation (2.5.1) is obtained by noting that $\left(1-\gamma_k\eta_k\mu_f\mathsf{p}_{min}\right)\left(1+\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}\right)\leq1-\frac{\mathsf{p}_{min}\mu_f\gamma_k\eta_k}{2}$. ■

In the following result, we provide a class of update rules for the stepsize and the regularization sequences such that Algorithm 2 attains both an almost sure convergence and a convergence in the mean sense.

Theorem 2.5.1 (Convergence of Algorithm 2 when X is unbounded) *Consider problem* (P_1). *Let the sequence* $\{\bar{x}_k\}$ *be generated by Algorithm 2. Let Assumption 2.5.1, Assumption 2.3.2, and Assumption 2.5.2 hold. Suppose the random block-coordinate* i_k *in Assumption 2.3.2 is drawn from a uniform distribution for all* $k \geq 0$. *Let the stepsize* $\{\gamma_k\}$ *and the regularization parameter* $\{\eta_k\}$ *be given by* $\gamma_k := \gamma_0(k+1)^{-a}$ *and* $\eta_k := \eta_0(k+1)^{-b}$, *respectively, where* $\gamma_0 > 0$, $\eta_0 > 0$, $0 < b < 0.5 < a$, *and* $a + b < 1$. *Then, the following results hold for all* $0 \leq r < 1$*:*
(i) The sequence $\{\bar{x}_k\}$ *converges almost surely to the unique optimal solution of Problem* (P_1).
(ii) We have that $\lim_{k\to\infty} \mathsf{E}[\|\bar{x}_k - x^*\|] = 0$.

Proof. The proof is done in two main steps. In the first step, we show that the non-averaged sequence $\{x_k\}$ converges to x^* in an almost sure sense and that $\lim_{k\to\infty} \mathsf{E}[\|x_k - x^*\|] = 0$. In the second step, we show that these results hold for the weighted average sequence $\{\bar{x}_k\}$ as well.
Step 1: The proof of this step is done by applying Lemma 2.5.3 to the recursive inequality equation (2.5.1) with $\mathsf{p}_i := \frac{1}{d}$ for all $i \in \{1, \ldots, d\}$. The details are as follows. First, we note that from the update rules of γ_k and η_k and that $a > b$, we have $\lim_{k\to\infty} \frac{\gamma_k}{\eta_k} = 0$. Thus, there exists an integer $k_0 \geq 1$ such that for all $k \geq k_0$, we have $\frac{\gamma_k}{\eta_k} \leq \frac{\mu_f \mathsf{p}_{min}}{2\mathsf{p}_{max}\left(L_F^2 + \eta_0^2 L_f^2\right)}$. This implies that the conditions of Lemma 2.5.4 are satisfied and the inequality equation (2.5.1) holds for all $k \geq k_0$. To apply Lemma 2.5.3, we define the following terms for all $k \geq 1$:

$$v_k \triangleq \mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right), \quad \alpha_k \triangleq \frac{\mu_f \gamma_k \eta_k}{2d},$$
$$\beta_k \triangleq \left(\frac{d\bar{C}_f^2\left(\mu_f \eta_0 \gamma_0 + 2d\right)}{\mu_f^3 \gamma_k \eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 + 2\gamma_k^2 B_F.$$

Since $\gamma_k \eta_k \to 0$, there exists an integer $k_1 \geq k_0$ such that for any $k \geq k_1$ we have $0 \leq \alpha_k \leq 1$. From the assumption that $a + b < 1$, we have that $\sum_{k=k_1}^{\infty} \alpha_k = \infty$. Next, we show that $\sum_{k=k_1}^{\infty} \beta_k < \infty$. From the update rules of γ_k and η_k and invoking the Taylor series expansion, for $k \geq 2$ we can write

$$\frac{\eta_{k-1}}{\eta_k} - 1 = \left(1 + \frac{1}{k}\right)^b - 1 = \left(1 + \frac{b}{k} + \frac{b(b-1)}{2!}\frac{1}{k^2} + \frac{b(b-1)(b-2)}{3!}\frac{1}{k^3} + \ldots\right) - 1$$
$$= \frac{b}{k}\left(1 - \frac{(1-b)}{2!k} + \frac{(1-b)(2-b)}{3!k^2} - \frac{(1-b)(2-b)(3-b)}{4!k^3} + \ldots\right) \leq \frac{b}{k}\sum_{i=0}^{\infty}\frac{1}{k^{2i}},$$

where the inequality is obtained using $b < 1$ and neglecting the negative terms. This implies that $\frac{\eta_{k-1}}{\eta_k} - 1 \leq \frac{b}{k(1-k^{-2})}$ and thus $\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 \leq \left(\frac{4b}{3k}\right)^2 \leq \frac{2b^2}{k^2}$ for all $k \geq 2$. Using the preceding relation, invoking the definition of β_k, and the update formulas of γ_k and η_k, we have that $\beta_k = \mathcal{O}\left(k^{-(2-a-b)}\right) + \mathcal{O}\left(k^{-2a}\right)$. From the assumptions on a and b, we obtain that

$\sum_{k=k_1}^{\infty} \beta_k < \infty$. Also, from the assumption $a > b$, we get $\lim_{k\to\infty} \beta_k/\alpha_k = 0$. Therefore, all conditions of Lemma 2.5.3 are satisfied. As such, we have that $\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) \to 0$ almost surely and also $\lim_{k\to\infty} \mathsf{E}\left[\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)\right] = 0$. From Remark 2.3.3 and that i_k is drawn uniformly, we obtain

$$\begin{aligned}\|x_k - x^*\|^2 &\leq 2\left\|x_k - x^*_{\eta_{k-1}}\right\|^2 + 2\left\|x^*_{\eta_{k-1}} - x^*\right\|^2 \\ &= \frac{2}{d}\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) + 2\left\|x^*_{\eta_{k-1}} - x^*\right\|^2, \end{aligned} \tag{2.5.7}$$

where the first inequality is obtained from the triangle inequality. Taking the limit from both sides of the preceding relation when $k \to \infty$ and invoking Lemma 2.5.1(a), we obtain $\lim_{k\to\infty} \|x_k - x^*\|^2 \leq \frac{2}{d} \lim_{k\to\infty} \mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)$. From the almost sure convergence of $\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)$ to zero, we conclude that $\{x_k\}$ converges to x^* almost surely. To show the convergence in mean, let us take the expectation from both sides of equation (2.5.7). Noting that the Tikhonov trajectory is deterministic, we obtain that $\mathsf{E}\left[\|x_k - x^*\|^2\right] \leq \frac{2}{d}\mathsf{E}\left[\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)\right] + 2\left\|x^*_{\eta_{k-1}} - x^*\right\|^2$. Now, taking the limit from both sides of the preceding relation when $k \to \infty$, invoking Lemma 2.5.1(a), and recalling $\lim_{k\to\infty} \mathsf{E}\left[\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)\right] = 0$, we conclude that $\lim_{k\to\infty} \mathsf{E}\left[\|x_k - x^*\|^2\right] = 0$. Invoking Jensen's inequality, we can conclude that $\lim_{k\to\infty} \mathsf{E}[\|x_k - x^*\|] = 0$.

Step 2: Invoking Lemma 2.3.1 and using the triangle inequality, we have

$$\|\bar{x}_k - x^*\| = \left\|\sum_{t=0}^{k} \lambda_{t,k} x_t - x^*\right\| = \left\|\sum_{t=0}^{k} \lambda_{t,k}\left(x_t - x^*\right)\right\| \leq \sum_{t=0}^{k} \lambda_{t,k}\|x_t - x^*\|, \tag{2.5.8}$$

where $\lambda_{t,k} \triangleq \gamma_t^r / \sum_{j=0}^{k} \gamma_j^r$. In view of Lemma 2.5.2, let us define $u_t \triangleq \|x_t - x^*\|$, $v_k \triangleq \sum_{t=0}^{k} \lambda_{t,k}\|x_t - x^*\|$, and $\alpha_t \triangleq \gamma_t^r$. Note that since $ar \leq 1$, we have $\sum_{t=0}^{\infty} \alpha_t = \gamma_0^r \sum_{t=0}^{\infty}(t+1)^{-ar} = \infty$. Also, from Step 1 we have that $\hat{u} \triangleq \lim_{t\to\infty} u_t = 0$ in an almost sure sense. Thus, from Lemma 2.5.2, we conclude that $\{v_k\}$ converges to zero almost surely. Thus, equation (2.5.8) implies that $\{\bar{x}_k\}$ converges to x^* almost surely. Next, we apply Lemma 2.5.2 again, but in a slightly different fashion to show that $\lim_{k\to\infty} \mathsf{E}[\|\bar{x}_k - x^*\|] = 0$. From equation (2.5.8), we have

$$\mathsf{E}[\|\bar{x}_k - x^*\|] \leq \sum_{t=0}^{k} \lambda_{t,k}\mathsf{E}[\|x_t - x^*\|]. \tag{2.5.9}$$

In view of Lemma 2.5.2, let us define $u_t \triangleq \mathsf{E}[\|x_t - x^*\|]$, $v_k \triangleq \sum_{t=0}^{k} \lambda_{t,k}\mathsf{E}[\|x_t - x^*\|]$, and $\alpha_t \triangleq \gamma_t^r$. First, note that from Step 1, we have $\hat{u} \triangleq \lim_{t\to\infty} u_t = 0$. In view of Lemma 2.5.2, $\lim_{k\to\infty} v_k = 0$. Thus, from equation (2.5.9), we conclude that $\lim_{k\to\infty} \mathsf{E}[\|\bar{x}_k - x^*\|] = 0$. Hence, the proof is completed. ■

2.6 Experimental Results

In this section, we revisit the problem of finding the best Nash equilibrium formulated as in Problem (1.1.1). We consider a case where the Nash game is characterized as a Cournot competition over a network. Cournot game is one of the most extensively studied economic models for competition among multiple firms, including imperfectly competitive power markets as well as rate control over communication networks [30, 44, 46]. Consider a collection of d firms who compete to sell a commodity over a network with J nodes. The decision of each firm $i \in \{1, \ldots, d\}$ includes variables y_{ij} and s_{ij}, denoting the generation and sales of the firm i at the node j, respectively. Considering the definitions $y_i \triangleq (y_{i1}; \ldots; y_{iJ})$ and $s_i \triangleq (s_{i1}; \ldots; s_{iJ})$, we can compactly denote the decision variable of the i^{th} firm as $x^{(i)} \triangleq (y_i; s_i) \in \mathbb{R}^{2J}$. The goal of the i^{th} firm lies in minimizing the net cost function $g_i\left(x^{(i)}, x^{(-i)}\right)$ over the network defined as follows

$$g_i\left(x^{(i)}; x^{(-i)}\right) \triangleq \sum_{j=1}^{J} c_{ij}(y_{ij}) - \sum_{j=1}^{J} s_{ij} p_j\left(\bar{s}_j\right),$$

where $c_{ij} : \mathbb{R} \to \mathbb{R}$ denotes the production cost function of the firm i at the node j, $\bar{s}_j \triangleq \sum_{i=1}^{d} s_{ij}$ denotes the aggregate sales from all the firms at the node j, and $p_j : \mathbb{R} \to \mathbb{R}$ denotes the price function with respect to the aggregate sales $\bar{s}_j$ at the node j. We assume that the cost functions are linear and the price functions are given as $p_j\left(\bar{s}_j\right) \triangleq \alpha_j - \beta_j \left(\bar{s}_j\right)^{\sigma}$ where $\sigma \geq 1$ and α_j and β_j are positive scalars. Throughout, we assume that the transportation costs are negligible. We let the generation be capacitated as $y_{ij} \leq \mathcal{B}_{ij}$, where $\mathcal{B}_{ij}$ is a positive scalar for $i \in \{1, \ldots, d\}$ and $j \in \{1, \ldots, J\}$. Lastly, for any firm i, the total sales must match with the total generation. Consequently, the strategy set of the firm i is given as follows

$$X_i \triangleq \left\{ (y_i; s_i) \mid \sum_{j=1}^{J} y_{ij} = \sum_{j=1}^{J} s_{ij}, \quad y_{ij}, s_{ij} \geq 0, \quad y_{ij} \leq \mathcal{B}_{ij}, \quad \text{for all } j = 1, \ldots, J \right\}.$$

Following the Problem (1.1.1), we employ the Marshallian aggregate surplus function defined as $f(x) \triangleq \sum_{i=1}^{d} g_i\left(x^{(i)}; x^{(-i)}\right)$. We note that the convexity of the function f is implied by $\sigma \geq 1$ and the monotonicity of mapping F is guaranteed when either $\sigma = 1$, or when $1 < \sigma \leq 3$ and $d \leq \frac{3\sigma - 1}{\sigma - 1}$ (cf. section 4 in [46]).

The set-up. In the experiment, we consider a Cournot game among 4 firms over 3 nodes. We let the slopes of the linear cost functions take values between 10 and 50. We assume that $\alpha_j := 50$ and $\beta_j := 0.05$ for all j, $\mathcal{B}_{ij} := 120$ for all i and j, and $\sigma := 1.01$. To report the performance of Algorithm 2 in terms of the suboptimality, we plot a sample average approximation of $\mathsf{E}[f\left(\bar{x}_N\right)]$ using the sample size of 25. With regard to the infeasibility, we compute a sample average approximation of $\mathsf{E}[\mathrm{GAP}\left(\bar{x}_N\right)]$ using the same sample size. Following Remark 2.4.1, we use $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{\sqrt[4]{k+1}}$. To select the block-coordinates in Algorithm 2, we use a discrete uniform distribution.

Results and insights. Figure 3 shows the experimental results. Here, in the top three figures, we compare the performance of Algorithm 2 with that of Algorithm 1 in terms of infeasibility measured by the sample averaged gap function. Importantly, the proposed algorithm performs significantly better than the SR scheme. This claim is supported by

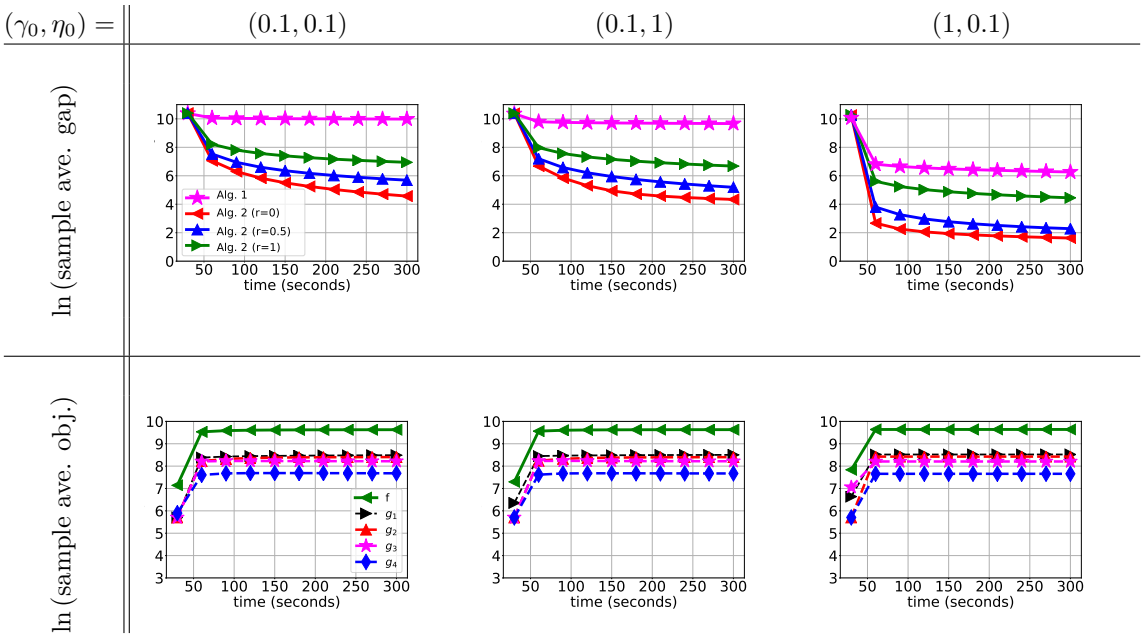

Figure 3: Algorithm 2 in terms of infeasibility and the objective function value

considering the different values of the parameter r and the initial conditions of the proposed scheme in terms of the initial stepsize γ_0 and the initial regularization parameter η_0. The three figures in the bottom row of Figure 3 demonstrate the performance of Algorithm 2 in terms of reaching a stability in the objective values. This includes the Marshallian objective function f as well as the individual objective functions g_i. Note that all the objective values in Figure 3 appear to reach to a desired level of stability after around 60 seconds. This interesting observation could be linked to the impact of the averaging scheme equation (2.3.2). Generally, it is expected that the trajectories of the objective function values in Figure 3 be noisy due to the randomness in the block-coordinate selection rule. However, the weighted averaging scheme employed in Algorithm 2 appears to induce much robustness with respect to this uncertainty, resulting in an accelerated convergence.

2.7 Conclusions

Motivated by the applications arising from noncooperative multi-agent networks, we consider a class of optimization problems with Cartesian variational inequality (CVI) constraints. The computational complexity of the solution methods for addressing this class of problems appears to be unknown. We develop a single-timescale algorithm equipped with non-asymptotic suboptimality and infeasibility convergence rates. Moreover, in the case where the set associated with the CVI is unbounded, we establish the global convergence of the sequence generated by the proposed algorithm. We apply the method in computing the best Nash equilibrium in a networked Cournot competition. Our experimental results show that the proposed method outperforms the classical sequential regularized schemes.

CHAPTER III

DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS

In this chapter, we consider a class of constrained multi-agent optimization problems where the goal is to cooperatively minimize a sum of agent-specific nondifferentiable merely convex functions. The constraint set is characterized as a variational inequality (VI) problem where each agent is associated with a local monotone mapping. Section 3.1 includes the problem formulation and Section 3.2 summarizes the literature for addressing problem (P_2). In addressing the model of interest, our contributions are as follows: (i) We develop an iteratively regularized incremental gradient method where at each iteration, agents communicate over a cycle graph to update their solution iterates using their local information about the objective and the mapping. The proposed method is single-timescale in the sense that it does not involve any excessive hard-to-project computation per iteration. (ii) We derive non-asymptotic agent-wise convergence rates for the suboptimality of the global objective function and infeasibility of the VI constraints measured by a suitably defined dual gap function. (iii) To analyze the convergence rate in the solution space, assuming the objective function is strongly convex and smooth, we derive non-asymptotic agent-wise rates on an error metric that relates the generated iterates with the Tikhonov trajectory. The proposed method in Section 3.3 appears to be the first fully iterative scheme equipped with iteration complexity that can address distributed optimization problems with VI constraints. Section 3.4 includes the non-asymptotic agent-wise convergence rate analysis for the suboptimality of the global objective function and infeasibility of the VI constraints. Section 3.5 provides the non-asymptotic agent-wise rates of the generated iterates in comparison with the Tikhonov trajectory. In Section 3.6, we provide preliminary numerical experiments for computing the best equilibrium in a transportation network problem.

3.1 Problem Formulation

The goal in this chapter is to tackle some of the shortcomings in distributed constrained optimization through considering a new unifying mathematical framework described as follows. Consider a system with m agents where the i^{th} agent is associated with a component function $f_i : \mathbb{R}^n \to \mathbb{R}$ and a mapping $F_i : \mathbb{R}^n \to \mathbb{R}^n$. Our goal is to solve the following

The content of this chapter is submitted to IEEE Transactions on Automatic Control [50].

distributed constrained optimization problem

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \mathrm{SOL}\left(X, \sum_{i=1}^{m} F_i\right), \end{aligned} \tag{P_2}$$

where $X \subseteq \mathbb{R}^n$ is a set and $\mathrm{SOL}\left(X, \sum_{i=1}^{m} F_i\right)$ denotes the solution set of the variational inequality $\mathrm{VI}\left(X, \sum_{i=1}^{m} F_i\right)$ defined as follows: $x \in X$ solves $\mathrm{VI}\left(X, \sum_{i=1}^{m} F_i\right)$ if we have $(y - x)^T \sum_{i=1}^{m} F_i(x) \geq 0$ for all $y \in X$. Problem (P_2) represents a distributed optimization framework in the sense that the information about f_i and F_i is locally known to the i^{th} agent, while the set X is globally known to all the agents. We consider the case where the local functions f_i are nondifferentiable and merely convex, and mappings F_i are single-valued, continuous, and merely monotone.

3.2 Existing Methods and Research Gap

In addressing the proposed formulation (P_2), our focus in this chapter lies in the development of an incremental gradient (IG) method. IG methods are among popular avenues for

Table 1: Comparison of incremental gradient schemes for solving finite-sum problems

Reference	Method	Problem class	Problem formulation	Convergence rate(s)
[62]	Projected IG	$f_i \in C_{0,L}^{0,0}$	$\min_{x \in X} \sum_{i=1}^{m} f_i(x)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$
[17, 34]	IAG	$f_i \in C_{\mu,L}^{1,1}$	$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x)$	$\mathcal{O}\left(\rho^k\right)$
[29]	SAGA	$f_i \in C_{0,L}^{1,1},\ C_{\mu,L}^{1,1}$	$\min_{x \in X} \sum_{i=1}^{m} f_i(x)$	$\mathcal{O}\left(\frac{1}{k}\right)$, $\mathcal{O}\left(\rho^k\right)$
[87]	Proximal IAG	$f_i \in C_{\mu,L}^{1,1}$	$\min_{x \in X} \sum_{i=1}^{m} f_i(x)$	$\mathcal{O}\left(\rho^k\right)$
[35]	IG	$f_i \in C_{\mu,L}^{2,1}$	$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x)$	$\mathcal{O}\left(\frac{1}{k}\right)$, $\mathcal{O}\left(\frac{1}{k^2}\right)$
[42]	Primal-Dual IG	$f_i \in C_{0,L}^{0,0}$	$\min_{x \in X} \sum_{i=1}^{m} f_i(x)$ $Ax - b \in -\mathcal{K}$	$\mathcal{O}\left(\frac{1}{k}\right)$
This work	pair-IG	$f_i \in C_{0,L}^{0,0}$, F_i is monotone	$\min \sum_{i=1}^{m} f_i(x)$ $x \in \mathrm{SOL}\left(X, \sum_{i=1}^{m} F_i\right)$	suboptimality: $\mathcal{O}\left(k^{b-0.5}\right)$ infeasibility: $\mathcal{O}\left(k^{-b}\right)$ where $0 < b < 0.5$

addressing the classical distributed optimization model (1.2.1) and they have received an increasing attention in recent years in addressing applications arising in sensor networks and machine learning [34, 35, 62, 89]. In these schemes, utilizing the additive structure of the problem, the algorithm cycles through the data blocks and updates the local estimates of the optimal solution in a sequential manner [14]. While the first variants of IG schemes find their roots in addressing neural networks as early as in the 1980s [15], the complexity analysis of these schemes has been a trending research topic in the fields of control and machine learning in the past two decades. In addressing the constrained problems with easy-to-project constraint sets, the projected incremental gradient (P-IG) method and its subgradient variant were developed [63]. Considering the smooth case, the P-IG scheme is described as follows. Given an initial point $x_{0,1} \in X$ where $X \subseteq \mathbb{R}^n$ denotes the constraint

set, for each $k \geq 0$, consider the update rules given by

$$x_{k,i+1} := \mathcal{P}_X\left(x_{k,i} - \gamma_k \nabla f_i\left(x_{k,i}\right)\right), \quad \text{for } i \in [m],$$
$$x_{k+1,1} := x_{k,m+1},$$

where $x_{k,i} \in \mathbb{R}^n$ denotes agent i's local copy of the decision variables at iteration k, $\mathcal{P}$ denotes the Euclidean projection operator defined as $\mathcal{P}_X(z) \triangleq \operatorname{argmin}_{x \in X} \|x - z\|_2$, and $\gamma_k > 0$ denotes the stepsize parameter. To motivate our research, we provide an overview of the different variants of existing IG schemes and then, highlight some of the shortcomings of these methods in the constrained regime, in particular, in addressing VI constraints in (P_2). Recently, under the assumption of strong convexity and twice continuous differentiability of the objective function, and also, boundedness of the generated iterates, the standard IG method was proved to converge with the rate $\mathcal{O}(1/k)$ in the unconstrained case [35]. This is an improvement to the previously known rate of $\mathcal{O}(1/\sqrt{k})$ for the merely convex case. Accelerated variants of IG schemes with provable convergence speeds were also developed, including the incremental aggregated gradient method (IAG) [17, 34], SAG [74], and SAGA [29]. While addressing the merely convex case, SAGA using averaging achieves a sublinear convergence rate, assuming strong convexity and smoothness, this is improved for non-averaging variants of SAGA and IAG to a linear rate. Table 1 presents a summary of the standard IG schemes in addressing unconstrained and constrained finite-sum problems. As evidenced, most of the past research efforts on the design and analysis of algorithms for distributed constrained optimization problems have focused on addressing easy-to-project sets or sets with linear functional inequalities. This has been done through employing duality theory, projection, or penalty methods (see [6, 21, 65, 79]). Also, a celebrated variant of the dual based schemes is the alternating direction method of multipliers (ADMM) (e.g., see [60, 84]). Other related papers that have utilized duality theory in distributed constrained regimes include [6, 13, 36]. Despite the extensive work in the area of constrained optimization, no provably convergent iterative method exits in the literature that can be employed to solve distributed optimization problems with VI constraints. In fact, we are unaware of any IG methods with complexity guarantees that can be employed for addressing any of the individual subclass problems, specifically Examples 5, 6, and 7 from Chapter 1.

3.3 Algorithm Outline

In this section we present the main assumptions on problem (P_2), the outline of the proposed algorithm, and a few preliminary results that will be applied later in the rate analysis. Throughout this chapter, we let $f(x) \triangleq \sum_{i=1}^m f_i(x)$ and $F(x) \triangleq \sum_{i=1}^m F_i(x)$ denote the global objective and global mapping in problem (P_2), respectively.

Assumption 3.3.1 (Properties of problem (P_2)) *Suppose the following conditions hold.*
(a) Function $f_i : \mathbb{R}^n \to \mathbb{R}$ is real-valued and merely convex (possibly nondifferentiable) on its domain for all $i \in [m]$.
(b) Mapping $F_i : \mathbb{R}^n \to \mathbb{R}^n$ is real-valued, continuous, and merely monotone on its domain for all $i \in [m]$.
(c) The set $X \subseteq int(dom(f) \cap dom(F))$ is nonempty, convex, and compact.

Remark 3.3.1 Under Assumption 3.3.1 we have the following immediate results. From Theorem 2.3.5 and Corollary 2.2.5 in [30], the set $\mathrm{SOL}(X, F)$ is nonempty, convex, and compact. For all i, the nonemptiness of the subdifferential set $\partial f_i(x)$ for any $x \in \mathrm{int}(\mathrm{dom}(f_i))$ is implied from Theorem 3.14 in [9]. Also, Theorem 3.16 in [9] implies that f_i has bounded subgradients over the compact set X. Further, mapping F_i is bounded over the set X.

In view of compactness of the set X and continuity of f, throughout this chapter we let positive scalars $M_X < \infty$ and $M_f < \infty$ be defined as $M_X \triangleq \sup_{x\in X} \|x\|$ and $M_f \triangleq \sup_{x\in X} |f(x)|$, respectively. We also let $f^* \in \mathbb{R}$ denote the optimal objective value of problem (P_2). In view of Remark 3.3.1, throughout we let scalars $C_F > 0$ and $C_f > 0$ be defined such that for all $i \in [m]$ and for all $x \in X$ we have $\|F_i(x)\| \leq \frac{C_F}{m}$, and $\left\|\tilde{\nabla} f_i\,(x)\right\| \leq \frac{C_f}{m}$ for all $\tilde{\nabla} f_i\,(x) \in \partial f_i(x)$. In the following, we comment on the Lipschitz continuity of the local and global objective functions.

Remark 3.3.2 Under Assumption 3.3.1 and from Theorem 3.61 in [9], function f_i is Lipschitz continuous with the parameter $\frac{C_f}{m}$ over the set X, i.e., for all $i \in [m]$ we have $|f_i(x) - f_i(y)| \leq \frac{C_f}{m}\|x - y\|$ for all $x, y \in X$. We also have $\|\tilde{\nabla} f(x)\| \leq C_f$ for all $x \in X$ and all $\tilde{\nabla} f(x) \in \partial f(x)$. This implies that $|f(x) - f(y)| \leq C_f\|x - y\|$ for all $x, y \in X$.

We now present an overview of the proposed method given by Algorithm 4. We use vector $x_{k,i}$ to denote the local copy of the global decision vector maintained by agent i at iteration k. At each iteration, agents update their iterates in a cyclic manner. Each agent $i \in [m]$ uses only its local information including the subgradient of the function f_i and mapping F_i and evaluates the regularized mapping $F_i + \eta_k\tilde{\nabla} f_i$ at $x_{k,i}$. Here, γ_k and η_k denote the stepsize and the regularization parameter at iteration k, respectively. Importantly, through employing an iterative regularization technique, we let both of these parameters be updated iteratively at suitable prescribed rates (cf. Theorem 3.4.1). Each agent computes and returns a weighted averaging iterative denoted by $\bar{x}_{k,i}$ where the weights are characterized in terms of the stepsize γ_k and an arbitrary scalar $r \in [0, 1)$. Notably, this averaging technique is carried out in a distributed fashion in the sense that agents do not require to start from the same initialized averaging iterate. This is in contrast with the standard incremental gradient schemes where the averaging scheme is limited to a centralized initialization. Next we show that for any $i \in [m]$, $\bar{x}_{N,i}$ is indeed a well-defined weighted average of $\bar{x}_{0,i}$ and the iterates $x_{k-1,i+1}$ for $k \in [N]$.

Lemma 3.3.1 *Consider the sequence $\{\bar{x}_{k,i}\}$ generated by agent $i \in [m]$ in Algorithm 4. For $k \in \{0, \ldots, N\}$, let us define the weights $\lambda_{k,N} \triangleq \frac{\gamma_k^r}{\sum_{j=0}^N \gamma_j^r}$. Then for all $i \in [m]$ we have*

$$\bar{x}_{N,i} = \lambda_{0,N}\bar{x}_{0,i} + \sum_{k=1}^{N} \lambda_{k,N} x_{k-1,i+1}.$$

Further, for a convex set X we have $\bar{x}_{N,i} \in X$.

Proof. We use induction on $N \geq 0$ to show the equation. For $N = 0$, from $\lambda_{0,0} = 1$ we have $\bar{x}_{0,i} = \lambda_{0,0}\bar{x}_{0,i}$. Now, assume that the equation holds for some $N \geq 0$. This implies

$$\bar{x}_{N,i} = \lambda_{0,N}\bar{x}_{0,i} + \sum_{k=1}^{N} \lambda_{k,N} x_{k-1,i+1} = \frac{\gamma_0^r\bar{x}_{0,i} + \sum_{k=1}^N \gamma_k^r x_{k-1,i+1}}{\sum_{j=0}^N \gamma_j^r}. \tag{3.3.3}$$

Algorithm 4 projected averaging iteratively regularized Incremental subGradient (pair-IG)

input: Agent 1 arbitrarily chooses an initial vector $x_{0,1} \in X$. Agent i arbitrarily chooses $\bar{x}_{0,i} \in X$, for all $i \in [m]$. Let $S_0 := \gamma_0^r$ with an arbitrary $0 \leq r < 1$.
for $k = 0, 1, \ldots, N-1$ **do**
  Update $S_{k+1} := S_k + \gamma_{k+1}^r$
  for $i = 1, \ldots, m$ **do**

$$x_{k,i+1} := \mathcal{P}_X \left(x_{k,i} - \gamma_k \left(F_i \left(x_{k,i} \right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i} \right) \right) \right) \tag{3.3.1}$$

$$\bar{x}_{k+1,i} := \left(\frac{S_k}{S_{k+1}} \right) \bar{x}_{k,i} + \left(\frac{\gamma_{k+1}^r}{S_{k+1}} \right) x_{k,i+1} \tag{3.3.2}$$

  end for
  Set $x_{k+1,1} := x_{k,m+1}$
end for
return: $\bar{x}_{N,i}$ for all $i \in [m]$

Using equation (3.3.3), we now show that the hypothesis statement holds for any $N+1$. From equation (3.3.2) we have $\bar{x}_{N+1,i} = \left(\frac{S_N}{S_{N+1}} \right) \bar{x}_{N,i} + \left(\frac{\gamma_{N+1}^r}{S_{N+1}} \right) x_{N,i+1}$. Note that from equation (3.3.2) in Algorithm 4 we have $S_k = \sum_{t=0}^{k} \gamma_t^r$ for all $k \geq 0$. From this and using equation (3.3.3) we obtain

$$\begin{aligned} \bar{x}_{N+1,i} &= \left(\frac{\sum_{t=0}^{N} \gamma_t^r}{\sum_{t=0}^{N+1} \gamma_t^r} \right) \bar{x}_{N,i} + \left(\frac{\gamma_{N+1}^r}{\sum_{t=0}^{N+1} \gamma_t^r} \right) x_{N,i+1} = \frac{\gamma_0^r \bar{x}_{0,i} + \sum_{k=1}^{N} \gamma_k^r x_{k-1,i+1} + \gamma_{N+1}^r x_{N,i+1}}{\sum_{t=0}^{N+1} \gamma_t^r} \\ &= \frac{\gamma_0^r \bar{x}_{0,i} + \sum_{k=1}^{N+1} \gamma_k^r x_{k-1,i+1}}{\sum_{t=0}^{N+1} \gamma_t^r}. \end{aligned}$$

From the definition of $\lambda_{k,N}$ we conclude that the hypothesis holds for $N+1$ and thus, the result holds for all $N \geq 0$. To show the second part, note that from the initialization in Algorithm 4 and the projection in equation (3.3.1), we have $\bar{x}_{0,i}, x_{k-1,i+1} \in X$ for all i and $k \geq 1$. From the first part, $\bar{x}_{N,i}$ is a convex combination of $\bar{x}_{0,i}, x_{0,i+1}, \ldots, x_{N-1,i+1}$. Therefore, from the convexity of the set X we conclude that $\bar{x}_{N,i} \in X$. ■

For the ease of presentation throughout the analysis, we define a sequence $\{x_k\}$ as follows.

Definition 3.3.1 *Consider Algorithm 4. Let the sequence $\{x_k\}$ be defined as $x_k \triangleq x_{k-1,m+1} = x_{k,1}$, for all $k \geq 1$, with $x_0 \triangleq x_{0,1}$.*

In the following result, we characterize the distance between the local variable of any arbitrary agent with that of the first and the last agent at any given iteration. This result will be utilized in the analysis.

Lemma 3.3.2 *Consider Algorithm 4. Let Assumption 3.3.1 hold. Then the following inequalities hold for all $i \in [m]$ and $k \geq 0$*
(a) $\|x_k - x_{k,i}\| \leq \frac{(i-1)\gamma_k \left(C_F + \eta_k C_f \right)}{m}$. *(b)* $\|x_{k,i+1} - x_{k+1}\| \leq \frac{(m-i)\gamma_k \left(C_F + \eta_k C_f \right)}{m}$.

Proof. (a) Let $k \geq 0$ be an arbitrary integer. We use induction on i to show this result. From Definition 3.3.1, for $i = 1$ and $k \geq 0$ we have $\|x_k - x_{k,1}\| = 0$, implying that the result holds for $i = 1$. Now suppose the hypothesis statement holds for some $i \in [m]$. We have

$$\begin{aligned}\|x_k - x_{k,i+1}\| &= \left\|\mathcal{P}_X\left(x_k\right) - \mathcal{P}_X\left(x_{k,i} - \gamma_k\left(F_i(x_{k,i}) + \eta_k\tilde{\nabla} f_i\left(x_{k,i}\right)\right)\right)\right\| \\ &\leq \|x_k - x_{k,i}\| + \gamma_k\left\|F_i(x_{k,i}) + \eta_k\tilde{\nabla} f_i\left(x_{k,i}\right)\right\| \\ &\leq \|x_k - x_{k,i}\| + \frac{\gamma_k\left(C_F + \eta_k C_f\right)}{m} \leq \frac{i\gamma_k\left(C_F + \eta_k C_f\right)}{m},\end{aligned}$$

where the first inequality is obtained from the nonexpansivity property of the projection. Therefore the hypothesis statement holds for any i and the proof of part (a) is completed. (b) To show this result, we use downward induction on $i \in [m]$. Note that the relation trivially holds for the base case $i = m$. Suppose it holds for some $i \in \{2, \ldots, m\}$. We show that it holds for $i - 1$ as well. From Definition 3.3.1 we have

$$\begin{aligned}\|x_{k,i} - x_{k+1}\| &= \|x_{k,i} - x_{k,i+1} + x_{k,i+1} - x_{k,m+1}\| \\ &\leq \|x_{k,i} - x_{k,i+1}\| + \|x_{k,i+1} - x_{k,m+1}\|.\end{aligned}$$

From equation (3.3.1), the hypothesis statement, and the nonexpansivity property of the projection, we obtain

$$\begin{aligned}\|x_{k,i} - x_{k+1}\| \leq &\left\|\mathcal{P}_X(x_{k,i}) - \mathcal{P}_X\left(x_{k,i} - \gamma_k\left(F_i\left(x_{k,i}\right) + \eta_k\tilde{\nabla} f_i\left(x_{k,i}\right)\right)\right)\right\| \\ &+ \frac{(m-i)\gamma_k\left(C_F + \eta_k C_f\right)}{m} \leq \frac{(m-i+1)\gamma_k\left(C_F + \eta_k C_f\right)}{m}.\end{aligned}$$

This completes the proof of part (b). ■

We note that the generated agent-wise iterates $\bar{x}_{k,i}$ in Algorithm 4, as the scheme proceeds, may not be solutions to $\mathrm{VI}(X, F)$ and so, they may not necessarily be feasible to problem (P_2). To quantify the infeasibility of these iterates, we employ a dual gap function (cf. Chapter 1 in [30]) defined in Definition 2.3.2.

We conclude this section by presenting the following result that will be utilized in the rate analysis.

Lemma 3.3.3 *Let $\beta \in [0, 1)$ and $\Gamma \geq 1$ be given scalars and K be an integer. Then for all $K \geq \left(2^{\frac{1}{1-\beta}} - 1\right)\Gamma$, we have*

$$\frac{(K+\Gamma)^{1-\beta}}{2(1-\beta)} \leq \sum_{k=0}^{K}(k+\Gamma)^{-\beta} \leq \frac{(K+\Gamma)^{1-\beta}}{1-\beta}.$$

Proof. See Appendix A.8. ■

3.4 Rate and Complexity Analysis

In this section we present the convergence and rate analysis of the proposed method under Assumption 3.3.1. After obtaining a preliminary inequality in Lemma 3.4.1 in terms of the sequence generated by the last agent, in Lemma 3.4.2 we derive inequalities that relate the global objective and the dual gap function at the iterate of other agents with those of the last agent. Utilizing these results, in Proposition 3.4.1 we obtain agent-specific bounds on the objective function value and the dual gap function. Consequently, in Theorem 3.4.1 we derive convergence rate statements under suitably chosen sequences for the stepsize and the regularization parameter.

Lemma 3.4.1 *Consider Algorithm 4. Let Assumption 3.3.1 hold. Let $\{\gamma_k\}$ and $\{\eta_k\}$ be nonincreasing and strictly positive sequences. For any arbitrary $y \in X$, for all $k \geq 0$ we have*

$$\begin{aligned} 2\gamma_k^r \left(\eta_k \left(f(x_k) - f(y)\right) + F(y)^T \left(x_k - y\right)\right) \leq & \gamma_k^{r-1} \left\|x_k - y\right\|^2 - \gamma_k^{r-1} \left\|x_{k+1} - y\right\|^2 \\ & + \gamma_k^{r+1} \left(C_F + \eta_k C_f\right)^2 . \end{aligned} \tag{3.4.1}$$

Proof. Let $y \in X$ be an arbitrary vector and $k \geq 0$ be fixed. From the update rule (3.3.1), for $i \in [m]$ we have

$$\left\|x_{k,i+1} - y\right\|^2 = \left\|\mathcal{P}_X \left(x_{k,i} - \gamma_k \left(F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right)\right) - \mathcal{P}_X(y)\right\|^2 .$$

Employing the nonexpansivity of the projection we have

$$\begin{aligned} \left\|x_{k,i+1} - y\right\|^2 &\leq \left\|x_{k,i} - \gamma_k \left(F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right) - y\right\|^2 \\ &= \left\|x_{k,i} - y\right\|^2 + \gamma_k^2 \left\|F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right\|^2 \\ &\quad - 2\gamma_k \left(F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right)^T \left(x_{k,i} - y\right) . \end{aligned}$$

From the triangle inequality and recalling the bounds on $\tilde{\nabla} f_i(x)$ and $F_i(x)$, we obtain

$$\left\|x_{k,i+1} - y\right\|^2 \leq \left\|x_{k,i} - y\right\|^2 + \gamma_k^2 \left(\frac{C_F + \eta_k C_f}{m}\right)^2 + 2\gamma_k \left(F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right)^T \left(y - x_{k,i}\right) . \tag{3.4.2}$$

The last term in the preceding relation is bounded as follows

$$\begin{aligned} 2\gamma_k \left(F_i \left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)\right)^T & \left(y - x_{k,i}\right) \\ &= 2\gamma_k F_i \left(x_{k,i}\right)^T \left(y - x_{k,i}\right) + 2\gamma_k \eta_k \tilde{\nabla} f_i \left(x_{k,i}\right)^T \left(y - x_{k,i}\right) \\ &\leq 2\gamma_k F_i \left(y\right)^T \left(y - x_{k,i}\right) + 2\gamma_k \eta_k \left(f_i \left(y\right) - f_i \left(x_{k,i}\right)\right) , \end{aligned}$$

where the last inequality is implied from the monotonicity of F_i and convexity of f_i. Combining with equation (3.4.2) we have

$$\begin{aligned}\|x_{k,i+1}-y\|^2 \leq \|x_{k,i}-y\|^2 &+ \gamma_k^2\left(\frac{C_F+\eta_k C_f}{m}\right)^2 \\ &+ 2\gamma_k F_i(y)^T(y-x_{k,i}) + 2\gamma_k\eta_k\left(f_i(y)-f_i(x_{k,i})\right).\end{aligned}$$

Adding and subtracting $2\gamma_k F_i(y)^T x_k + 2\gamma_k\eta_k f_i(x_k)$ we get

$$\begin{aligned}\|x_{k,i+1}-y\|^2 \leq \|x_{k,i}-y\|^2 &+ \gamma_k^2\left(\frac{C_F+\eta_k C_f}{m}\right)^2 \\ &+ 2\gamma_k F_i(y)^T(y-x_k) + 2\gamma_k\eta_k\left(f_i(y)-f_i(x_k)\right) \\ &+ 2\gamma_k\left(\left|F_i(y)^T(x_k-x_{k,i})\right| + \eta_k\left|f_i(x_k)-f_i(x_{k,i})\right|\right).\end{aligned}$$

Using the Cauchy-Schwarz inequality and Remark 3.3.2 we obtain

$$\begin{aligned}\|x_{k,i+1}-y\|^2 \leq \|x_{k,i}-y\|^2 &+ \gamma_k^2\left(\frac{C_F+\eta_k C_f}{m}\right)^2 \\ &+ 2\gamma_k F_i(y)^T(y-x_k) + 2\gamma_k\eta_k\left(f_i(y)-f_i(x_k)\right) \\ &+ 2\gamma_k\left(\frac{C_F}{m}\|x_k-x_{k,i}\| + \frac{\eta_k C_f}{m}\|x_k-x_{k,i}\|\right).\end{aligned}$$

Summing over $i\in[m]$ and considering Definition 3.3.1 we have

$$\begin{aligned}\|x_{k+1}-y\|^2 \leq \|x_k-y\|^2 &+ \frac{\gamma_k^2\left(C_F+\eta_k C_f\right)^2}{m} \\ &+ 2\gamma_k F(y)^T(y-x_k) + 2\gamma_k\eta_k\left(f(y)-f(x_k)\right) \\ &+ \frac{2\gamma_k\left(C_F+\eta_k C_f\right)}{m}\sum_{i=1}^m\|x_k-x_{k,i}\|.\end{aligned}$$

From Lemma 3.3.2 we obtain

$$\begin{aligned}\|x_{k+1}-y\|^2 &\leq \|x_k-y\|^2 + \frac{\gamma_k^2\left(C_F+\eta_k C_f\right)^2}{m} \\ &\quad + 2\gamma_k F(y)^T(y-x_k) + 2\gamma_k\eta_k\left(f(y)-f(x_k)\right) \\ &\quad + \frac{2\gamma_k\left(C_F+\eta_k C_f\right)}{m}\sum_{i=1}^m\frac{(i-1)\gamma_k\left(C_F+\eta_k C_f\right)}{m} \\ &= \|x_k-y\|^2 + \gamma_k^2\left(C_F+\eta_k C_f\right)^2 + 2\gamma_k F(y)^T(y-x_k) \\ &\quad + 2\gamma_k\eta_k\left(f(y)-f(x_k)\right).\end{aligned}$$

Multiplying the both sides by γ_k^{r-1} we can obtain the result. ■

In the next result we provide inequalities that relate the objective function and the dual gap function at the generated averaged iterate of the last agent with that of any other agent, respectively. This result will be utilized in Proposition 3.4.1.

Lemma 3.4.2 *Consider problem* (P_2) *and the sequences* $\{\bar{x}_{N,i}\}$ *generated in Algorithm 4 for* $i \in [m]$ *for some* $N \geq 1$. *Let Assumption 3.3.1 hold and let* $\{\gamma_k\}$ *and* $\{\eta_k\}$ *be strictly positive and nonincreasing sequences. Then for any* $i \in [m]$ *we have*

$$f(\bar{x}_{N,i}) - f(\bar{x}_{N,m}) \leq C_f \lambda_{0,N} \|\bar{x}_{0,i} - \bar{x}_{0,m}\| + \frac{(m-i)C_f\,(C_F + \eta_0 C_f)}{m} \sum_{k=0}^{N} \lambda_{k,N}\gamma_k, \tag{3.4.3a}$$

$$\mathrm{GAP}\,(\bar{x}_{N,i}) - \mathrm{GAP}(\bar{x}_{N,m}) \leq C_F \lambda_{0,N} \|\bar{x}_{0,i} - \bar{x}_{0,m}\| + \frac{(m-i)C_F\,(C_F + \eta_0 C_f)}{m} \sum_{k=0}^{N} \lambda_{k,N}\gamma_k, \tag{3.4.3b}$$

where $\lambda_{k,N} \triangleq \frac{\gamma_k^r}{\sum_{j=0}^{N} \gamma_j^r}$ *for* $k \in \{0, \ldots, N\}$.

Proof. Note that the results are trivial when $m = 1$. Throughout, we assume that $m \geq 2$. From the Lipschitz continuity of function f from Remark 3.3.2 and invoking Lemma 3.3.1, we can write the following for all $i \in [m]$.

$$\begin{aligned} f(\bar{x}_{N,i}) - f(\bar{x}_{N,m}) \leq & C_f \lambda_{0,N} \|\bar{x}_{0,i} - \bar{x}_{0,m}\| \\ & + C_f \sum_{k=1}^{N} \lambda_{k,N} \|x_{k-1,i+1} - x_{k-1,m+1}\|. \end{aligned} \tag{3.4.4}$$

Next, using Lemma 3.3.2(b) for any $k \geq 1$ and $i \in [m]$ we have

$$\|x_{k-1,i+1} - x_{k-1,m+1}\| \leq \frac{(m-i)\gamma_{k-1}\,(C_F + \eta_{k-1} C_f)}{m}. \tag{3.4.5}$$

From (3.4.4), (3.4.5), and the nonincreasing sequence $\{\eta_k\}$, we have

$$\begin{aligned} f(\bar{x}_{N,i}) - f(\bar{x}_{N,m}) \leq & \frac{(m-i)C_f\,(C_F + \eta_0 C_f)}{m} \sum_{k=1}^{N} \lambda_{k,N}\gamma_{k-1} \\ & + C_f \lambda_{0,N} \|\bar{x}_{0,i} - \bar{x}_{0,m}\|. \end{aligned}$$

Since $\{\gamma_k\}$ is nonincreasing and $0 \leq r < 1$, we obtain

$$\begin{aligned} \sum_{k=1}^{N} \lambda_{k,N}\gamma_{k-1} &\leq \frac{1}{\sum_{j=0}^{N} \gamma_j^r} \sum_{k=1}^{N} \gamma_{k-1}^{r+1} \leq \frac{1}{\sum_{j=0}^{N} \gamma_j^r} \sum_{k=0}^{N-1} \gamma_k^{r+1} \\ &\leq \frac{1}{\sum_{j=0}^{N} \gamma_j^r} \sum_{k=0}^{N} \gamma_k^{r+1} = \sum_{k=0}^{N} \lambda_{k,N}\gamma_k. \end{aligned}$$

From the last two relations we obtain equation (3.4.3a). Next we show (3.4.3b). From Definition 2.3.2 we have

$$\begin{aligned} \mathrm{GAP}(\bar{x}_{N,i}) &= \sup_{y \in X} F(y)^T (\bar{x}_{N,i} - y) \\ &= \sup_{y \in X} F(y)^T (\bar{x}_{N,i} + \bar{x}_{N,m} - \bar{x}_{N,m} - y) \\ &\leq \sup_{y \in X} F(y)^T (\bar{x}_{N,i} - \bar{x}_{N,m}) + \sup_{y \in X} F(y)^T (\bar{x}_{N,m} - y) \\ &\leq C_F \|\bar{x}_{N,i} - \bar{x}_{N,m}\| + \mathrm{GAP}\,(\bar{x}_{N,m}), \end{aligned}$$

Rearranging the terms we obtain $\mathrm{GAP}(\bar{x}_{N,i}) - \mathrm{GAP}\,(\bar{x}_{N,m}) \leq C_F\,\|\bar{x}_{N,i} - \bar{x}_{N,m}\|$. The rest of the proof can be done in a similar fashion to the proof of (3.4.3a). ■

Next we construct agent-wise error bounds in terms of the objective function value and the dual gap function at the averaged iterates generated in Algorithm 4.

Proposition 3.4.1 (Agent-wise error bounds) *Consider problem* (P_2) *and the averaged sequence* $\{\bar{x}_{k,i}\}$ *generated by agent* i *in Algorithm 4 for* $i \in [m]$. *Let Assumption 3.3.1 hold and* $\{\gamma_k\}$ *and* $\{\eta_k\}$ *be nonincreasing and strictly positive sequences. Then we have for* $i \in [m]$, $N \geq 1$, *and* $r \in [0, 1)$*:*

$$
\begin{aligned}
(a)\ f(\bar{x}_{N,i}) - f^* \leq & \left(\sum_{k=0}^{N} \gamma_k^r\right)^{-1} \left(\frac{2M_X^2 \gamma_N^{r-1}}{\eta_N} + \frac{(C_F + \eta_0 C_f)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k} + \gamma_0^r f(\bar{x}_{0,m})\right. \\
& \left. -\gamma_0^r f(x_{0,1}) + \frac{(m-i) C_f\,(C_F + \eta_0 C_f)}{m} \sum_{k=0}^{N} \gamma_k^{r+1} + C_f \gamma_0^r\,\|\bar{x}_{0,i} - \bar{x}_{0,m}\|\right). \\
(b)\ \mathrm{GAP}(\bar{x}_{N,i}) \leq & \left(\sum_{k=0}^{N} \gamma_k^r\right)^{-1} \left(2M_X^2 \gamma_N^{r-1} + 2M_f \sum_{k=0}^{N} \gamma_k^r \eta_k + \frac{(C_F + C_f \eta_0)^2}{2} \sum_{k=0}^{N} \gamma_k^{r+1}\right. \\
& \left. + \frac{(m-i) C_F\,(C_F + \eta_0 C_f)}{m} \sum_{k=0}^{N} \gamma_k^{r+1} + \gamma_0^r C_F \|\bar{x}_{0,m} - x_{0,1}\| + C_F \gamma_0^r\,\|\bar{x}_{0,i} - \bar{x}_{0,m}\|\right).
\end{aligned}
$$

Proof. (a) Let $x^* \in X$ denote an arbitrary optimal solution to problem (P_2). From feasibility of x^* we have $F(x^*)\,(x_k - x^*) \geq 0$. Substituting y by x^* in relation (3.4.1) and using the preceding relation we have

$$
2\gamma_k^r \eta_k\,(f(x_k) - f^*) \leq \gamma_k^{r-1}\,\|x_k - x^*\|^2 - \gamma_k^{r-1}\,\|x_{k+1} - x^*\|^2 + \gamma_k^{r+1}\,(C_F + \eta_k C_f)^2.
$$

Dividing both sides by $2\eta_k$ we have

$$
\gamma_k^r\,(f(x_k) - f^*) \leq \frac{\gamma_k^{r-1}}{2\eta_k}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) + \frac{\gamma_k^{r+1}}{2\eta_k}\,(C_F + \eta_k C_f)^2. \tag{3.4.6}
$$

Adding and subtracting the term $\frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\|x_k - x^*\|^2$ we have

$$
\begin{aligned}
\gamma_k^r\,(f(x_k) - f^*) \leq & \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\|x_k - x^*\|^2 - \frac{\gamma_k^{r-1}}{2\eta_k}\,\|x_{k+1} - x^*\|^2 \\
& + \underbrace{\left(\frac{\gamma_k^{r-1}}{2\eta_k} - \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\right)}_{\text{term 1}}\|x_k - x^*\|^2 + \frac{\gamma_k^{r+1}}{2\eta_k}\,(C_F + \eta_k C_f)^2.
\end{aligned} \tag{3.4.7}
$$

Recalling the definition of scalar M_X we have

$$
\|x_k - x^*\|^2 \leq 2\|x_k\|^2 + 2\|x^*\|^2 \leq 4M_X^2. \tag{3.4.8}
$$

Taking into account $r < 1$, the nonincreasing property of the sequences $\{\gamma_k\}$ and $\{\eta_k\}$, we have: term $1 \geq 0$. Using (3.4.8) and taking summation from (3.4.7) over $k \in [N]$, we obtain

$$\sum_{k=1}^{N} \gamma_k^r \left(f(x_k) - f^*\right) \leq \frac{\gamma_0^{r-1}}{2\eta_0}\|x_1 - x^*\|^2 - \frac{\gamma_N^{r-1}}{2\eta_N}\|x_{N+1} - x^*\|^2 + \left(\frac{\gamma_N^{r-1}}{2\eta_N} - \frac{\gamma_0^{r-1}}{2\eta_0}\right) 4M_X^2 + \frac{\left(C_F + \eta_0 C_f\right)^2}{2} \sum_{k=1}^{N} \frac{\gamma_k^{r+1}}{\eta_k}. \tag{3.4.9}$$

Rewriting equation (3.4.6) for $k = 0$ and then, adding and subtracting $f(\bar{x}_{0,m})$, we have

$$\gamma_0^r \left(f(\bar{x}_{0,m}) - f^* + f(x_0) - f(\bar{x}_{0,m})\right) \leq \frac{\gamma_0^{r-1}\left\|x_0 - x^*\right\|^2}{2\eta_0} - \frac{\gamma_0^{r-1}\left\|x_1 - x^*\right\|^2}{2\eta_0} + \left(C_F + \eta_0 C_f\right)^2 \frac{\gamma_0^{r+1}}{2\eta_0}.$$

Adding the preceding equation with (3.4.9) we obtain

$$\gamma_0^r \left(f(\bar{x}_{0,m}) - f^*\right) + \sum_{k=1}^{N} \gamma_k^r \left(f(x_k) - f^*\right) \leq \frac{\gamma_0^{r-1}\left\|x_0 - x^*\right\|^2}{2\eta_0} + 2M_X^2 \left(\frac{\gamma_N^{r-1}}{\eta_N} - \frac{\gamma_0^{r-1}}{\eta_0}\right) - \frac{\gamma_N^{r-1}}{2\eta_N}\left\|x_{N+1} - x^*\right\|^2 + \frac{\left(C_F + \eta_0 C_f\right)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k} + \gamma_0^r \left(f(\bar{x}_{0,m}) - f(x_0)\right).$$

From (3.4.8) and neglecting the nonpositive term we obtain

$$\gamma_0^r \left(f(\bar{x}_{0,m}) - f^*\right) + \sum_{k=1}^{N} \gamma_k^r \left(f(x_k) - f^*\right) \leq \frac{2M_X^2 \gamma_N^{r-1}}{\eta_N} + \frac{\left(C_F + \eta_0 C_f\right)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k} + \gamma_0^r \left(f(\bar{x}_{0,m}) - f(x_0)\right).$$

Next, dividing both sides by $\sum_{k=0}^{N} \gamma_k^r$ we have

$$\frac{\gamma_0^r f(\bar{x}_{0,m}) + \sum_{k=1}^{N} \gamma_k^r f(x_k)}{\sum_{k=0}^{N} \gamma_k^r} - f^* \leq \left(\sum_{k=0}^{N} \gamma_k^r\right)^{-1} \left(\frac{2M_X^2 \gamma_N^{r-1}}{\eta_N} + \frac{\left(C_F + \eta_0 C_f\right)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k} + \gamma_0^r \left(f(\bar{x}_{0,m}) - f(x_0)\right)\right).$$

Taking into account the convexity of f we have

$$f\left(\frac{\gamma_0^r \bar{x}_{0,m} + \sum_{k=1}^{N} \gamma_k^r x_{k-1,m+1}}{\sum_{k=0}^{N} \gamma_k^r}\right) \leq \frac{\gamma_0^r f(\bar{x}_{0,m}) + \sum_{k=1}^{N} \gamma_k^r f(x_{k-1,m+1})}{\sum_{k=0}^{N} \gamma_k^r}.$$

Invoking Lemma 3.3.1, from the preceding two relations we obtain

$$f(\bar{x}_{N,m}) - f^* \leq \left(\sum_{k=0}^{N} \gamma_k^r\right)^{-1} \left(\frac{2M_X^2 \gamma_N^{r-1}}{\eta_N} + \frac{(C_F + \eta_0 C_f)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k} + \gamma_0^r f(\bar{x}_{0,m}) - \gamma_0^r f(x_{0,1})\right).$$

Adding equation (3.4.3a) with the preceding inequality we obtain the desired result.
(b) From equation (3.4.1), for an arbitrary $y \in X$ we have

$$2\gamma_k^r F(y)^T (x_k - y) \leq \gamma_k^{r-1} \left(\|x_k - y\|^2 - \|x_{k+1} - y\|^2\right) + 2\gamma_k^r \eta_k \left(f(y) - f(x_k)\right) + \gamma_k^{r+1} (C_F + \eta_k C_f)^2.$$

From the triangle inequality and definition of M_f we have $|f(y) - f(x_k)| \leq 2M_f$. We obtain:

$$2\gamma_k^r F(y)^T (x_k - y) \leq \gamma_k^{r-1} \left(\|x_k - y\|^2 - \|x_{k+1} - y\|^2\right) + 4\gamma_k^r \eta_k M_f + \gamma_k^{r+1} (C_F + \eta_k C_f)^2. \tag{3.4.10}$$

Adding and subtracting $\gamma_{k-1}^{r-1} \|x_k - y\|^2$, we have:

$$2\gamma_k^r F(y)^T (x_k - y) \leq \gamma_{k-1}^{r-1} \|x_k - y\|^2 - \gamma_k^{r-1} \|x_{k+1} - y\|^2 + 4\gamma_k^r \eta_k M_f + \underbrace{\left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right) \|x_k - y\|^2}_{\text{term 2}} + \gamma_k^{r+1} (C_F + \eta_k C_f)^2. \tag{3.4.11}$$

Using the nonincreasing property of $\{\gamma_k\}$ and recalling $0 \leq r < 1$, we have $\gamma_k^{r-1} - \gamma_{k-1}^{r-1} \geq 0$. Thus, we can write: term $2 \leq \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right) 4M_X^2$. Taking summation over $k \in [N]$ in equation (3.4.11) and dropping a nonpositive term we obtain

$$2\sum_{k=1}^{N} \gamma_k^r F(y)^T (x_k - y) \leq \gamma_0^{r-1} \|x_1 - y\|^2 + 4M_f \sum_{k=1}^{N} \gamma_k^r \eta_k + 4M_X^2 \left(\gamma_N^{r-1} - \gamma_0^{r-1}\right) + (C_F + \eta_0 C_f)^2 \sum_{k=1}^{N} \gamma_k^{r+1}. \tag{3.4.12}$$

Writing equation (3.4.10) for $k = 0$ and adding and subtracting $2\gamma_0^r F(y)^T \bar{x}_{0,m}$, we have

$$2\gamma_0^r F(y)^T (\bar{x}_{0,m} - y + x_0 - \bar{x}_{0,m}) \leq 4\gamma_0^r \eta_0 M_f + \gamma_0^{r-1} \left(\|x_0 - y\|^2 - \|x_1 - y\|^2\right) + \gamma_0^{r+1} (C_F + \eta_0 C_f)^2.$$

Adding the preceding relation with equation (3.4.12) we have

$$2\gamma_0^r F(y)^T (\bar{x}_{0,m} - y) + 2\sum_{k=1}^{N} \gamma_k^r F(y)^T (x_k - y) \leq 4M_X^2 \left(\gamma_N^{r-1} - \gamma_0^{r-1}\right) + (C_F + \eta_0 C_f)^2 \sum_{k=0}^{N} \gamma_k^{r+1} + \gamma_0^{r-1} \|x_0 - y\|^2 + 4M_f \sum_{k=0}^{N} \gamma_k^r \eta_k + \underbrace{2\gamma_0^r F(y)^T (\bar{x}_{0,m} - x_0)}_{\text{term 3}}.$$

Using the Cauchy-Schwarz inequality we have: term $3 \leq 2\gamma_0^r C_F \|x_{0,m} - x_{0,1}\|$. We also have $\|x_0 - y\|^2 \leq 4M_X^2$. Dividing the both sides of the preceding inequality by $2\sum_{k=0}^N \gamma_k^r$ and invoking Lemma 3.3.1, we have

$$F(y)^T(\bar{x}_{N,m} - y) \leq \left(\sum_{k=0}^N \gamma_k^r\right)^{-1}\left(2M_X^2\gamma_N^{r-1} + 2M_f\sum_{k=0}^N \gamma_k^r\eta_k \right.$$
$$\left. + \frac{(C_F + \eta_0 C_f)^2}{2}\sum_{k=0}^N \gamma_k^{r+1} + \gamma_0^r C_F\|\bar{x}_{0,m} - x_{0,1}\|\right).$$

Taking the supremum on both sides with respect to y over the set X and recalling Definition 2.3.2 we have

$$\text{GAP}(\bar{x}_{N,m}) \leq \left(\sum_{k=0}^N \gamma_k^r\right)^{-1}\left(\frac{(C_F + \eta_0 C_f)^2}{2}\sum_{k=0}^N \gamma_k^{r+1}\right.$$
$$\left. + 2M_f\sum_{k=0}^N \gamma_k^r\eta_k + 2M_X^2\gamma_N^{r-1} + \gamma_0^r C_F\|\bar{x}_{0,m} - x_{0,1}\|\right).$$

Adding equation (3.4.3b) with the preceding inequality we obtain the desired inequality. ■

In the following we present the main result of this section. We provide non-asymptotic rate statements for each agent $i \in [m]$ in terms of suboptimality measured by the global objective function, and infeasibility characterized by the dual gap function. We note that unlike the analysis of the standard incremental gradient schemes in the literature, here we provide these rate results for individual agents $i \in [m]$.

Theorem 3.4.1 (Agent-wise rate statements for Algorithm 4) *Consider problem* (P_2). *Let the averaged sequence* $\{\bar{x}_{k,i}\}$ *be generated by agent* $i \in [m]$ *using Algorithm 4. Let Assumption 3.3.1 hold. Let the stepsize sequence* $\{\gamma_k\}$ *and the regularization sequence* $\{\eta_k\}$ *be updated using* $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ *and* $\eta_k := \frac{\eta_0}{(k+1)^b}$, *respectively, where* $\gamma_0, \eta_0 > 0$ *and* $0 < b < 0.5$. *Then the following inequalities hold for all* $i \in [m]$, *all* $N \geq 2^{\frac{2}{1-r}} - 1$, *and all* $r \in [0, 1)$:

$$(a)\ f(\bar{x}_{N,i}) - f^* \leq \frac{2-r}{(N+1)^{0.5-b}}\left(\frac{2M_X^2}{\eta_0\gamma_0} + \frac{\gamma_0(C_F + \eta_0 C_f)^2}{\eta_0(1-r+2b)} + f(\bar{x}_{0,m}) - f(x_{0,1})\right.$$
$$\left. + C_f\|\bar{x}_{0,i} - \bar{x}_{0,m}\| + \frac{2(m-i)\gamma_0 C_f(C_F + \eta_0 C_f)}{m(1-r)}\right). \tag{3.4.13}$$

$$(b)\ \text{GAP}(\bar{x}_{N,i}) \leq \frac{2-r}{(N+1)^b}\left(\frac{2M_X^2}{\gamma_0} + \frac{2M_f\eta_0}{1-0.5r-b} + C_F\|\bar{x}_{0,m} - x_{0,1}\| + C_F\|\bar{x}_{0,i} - \bar{x}_{0,m}\|\right.$$
$$\left. + \frac{(C_F + \eta_0 C_f)^2\gamma_0}{1-r} + \frac{2(m-i)C_F(C_F + \eta_0 C_f)\gamma_0}{m(1-r)}\right). \tag{3.4.14}$$

Proof. (a) Consider the inequality in Proposition 3.4.1(a). Substituting γ_k and η_k by their

update rules we obtain

$$f(\bar{x}_{N,i}) - f^* \leq \left(\sum_{k=0}^{N} \frac{\gamma_0^r}{(k+1)^{0.5r}}\right)^{-1} \left(\frac{2M_X^2(N+1)^{0.5(1-r)+b}}{\eta_0\gamma_0^{1-r}}\right.$$
$$+\frac{(C_F+\eta_0 C_f)^2}{2}\sum_{k=0}^{N}\frac{\gamma_0^{1+r}}{\eta_0(k+1)^{0.5(1+r)-b}} + \gamma_0^r f(\bar{x}_{0,m}) - \gamma_0^r f(x_{0,1})$$
$$\left.+\frac{(m-i)C_f\,(C_F+\eta_0 C_f)}{m}\sum_{k=0}^{N}\frac{\gamma_0^{r+1}}{(k+1)^{0.5(1+r)}} + C_f\gamma_0^r\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|\right).$$

In the next step, to apply Lemma 3.3.3 we need to ensure that the conditions in that result are met. From $0 \leq r < 1$ and $0 < b < 0.5$, we have $0 \leq 0.5r < 1$, $0 \leq 0.5(1+r) - b < 1$, $0 \leq 0.5r + b < 1$, and $0 \leq 0.5(1+r) < 1$. Further, from $N \geq 2^{\frac{2}{1-r}} - 1$, $0 < b < 0.5$, and $0 \leq r < 1$ we have that $N \geq \max\left\{2^{1/(1-0.5r)}, 2^{1/(1-0.5(1+r)+b)}, 2^{1/(1-0.5(1+r))}\right\} - 1$.
Therefore, all the necessary conditions of Lemma 3.3.3 are met.

$$f(\bar{x}_{N,i}) - f^* \leq \left(\frac{\gamma_0^r(N+1)^{1-0.5r}}{2(1-0.5r)}\right)^{-1}\left(\frac{2M_X^2(N+1)^{0.5(1-r)+b}}{\eta_0\gamma_0^{1-r}}\right.$$
$$+\frac{\gamma_0^{1+r}\,(C_F+\eta_0 C_f)^2\,(N+1)^{1-0.5(1+r)+b}}{2\eta_0(1-0.5(1+r)+b)}$$
$$+\frac{(m-i)C_f\,(C_F+\eta_0 C_f)\,\gamma_0^{r+1}(N+1)^{1-0.5(1+r)}}{m(1-0.5(1+r))}$$
$$\left.+\gamma_0^r f(\bar{x}_{0,m}) - \gamma_0^r f(x_{0,1}) + C_f\gamma_0^r\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|\right).$$

From the preceding relation we obtain

$$f(\bar{x}_{N,i}) - f^* \leq (2-r)\left(\frac{2M_X^2}{\eta_0\gamma_0(N+1)^{0.5-b}} + \frac{\gamma_0\,(C_F+\eta_0 C_f)^2}{2\eta_0(1-0.5(1+r)+b)(N+1)^{0.5-b}}\right.$$
$$\left.+\frac{(m-i)C_f\,(C_F+\eta_0 C_f)\,\gamma_0}{m(1-0.5(1+r))(N+1)^{0.5}} + \frac{f(\bar{x}_{0,m}) - f(x_{0,1}) + C_f\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|}{(N+1)^{1-0.5r}}\right).$$

Factoring out $1/(N+1)^{0.5-b}$ we obtain

$$f(\bar{x}_{N,i}) - f^* \leq \frac{2-r}{(N+1)^{0.5-b}}\left(\frac{2M_X^2}{\eta_0\gamma_0} + \frac{\gamma_0\,(C_F+\eta_0 C_f)^2}{2\eta_0(1-0.5(1+r)+b)}\right.$$
$$\left.+\frac{(m-i)C_f\,(C_F+\eta_0 C_f)\,\gamma_0}{m(1-0.5(1+r))(N+1)^{b}} + \frac{f(\bar{x}_{0,m}) - f(x_{0,1}) + C_f\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|}{(N+1)^{0.5-0.5r+b}}\right).$$

Note that from $b > 0$ and $r < 1$ we have $0.5 - 0.5r + b > 0$. Hence equation (3.4.13) holds.

(b) Consider the inequality in Proposition 3.4.1(b). We have

$$\begin{aligned}\mathrm{GAP}(\bar{x}_{N,i}) \leq \left(\sum_{k=0}^{N}\gamma_k^r\right)^{-1}\Bigg(&2M_X^2\gamma_N^{r-1}+2M_f\sum_{k=0}^{N}\gamma_k^r\eta_k\\&+\gamma_0^r C_F\|\bar{x}_{0,m}-x_{0,1}\|+\gamma_0^r C_F\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|\\&+\left(\frac{\left(C_F+\eta_0 C_f\right)^2}{2}+\frac{(m-i)C_F\left(C_F+\eta_0 C_f\right)}{m}\right)\sum_{k=0}^{N}\gamma_k^{r+1}\Bigg).\end{aligned}$$

Substituting $\{\gamma_k\}$ and $\{\eta_k\}$ by their update rules we obtain

$$\begin{aligned}\mathrm{GAP}(\bar{x}_{N,i}) \leq \left(\sum_{k=0}^{N}\frac{\gamma_0^r}{(k+1)^{0.5r}}\right)^{-1}\Bigg(&\frac{2M_X^2(N+1)^{0.5(1-r)}}{\gamma_0^{1-r}}\\&+\sum_{k=0}^{N}\frac{2M_f\eta_0\gamma_0^r}{(k+1)^{0.5r+b}}+\gamma_0^r C_F\left(\|\bar{x}_{0,m}-x_{0,1}\|+\|\bar{x}_{0,i}-\bar{x}_{0,m}\|\right)\\&+\left(\frac{\left(C_F+\eta_0 C_f\right)^2}{2}+\frac{(m-i)C_F\left(C_F+\eta_0 C_f\right)}{m}\right)\sum_{k=0}^{N}\frac{\gamma_0^{r+1}}{(k+1)^{0.5(1+r)}}\Bigg).\end{aligned}$$

Utilizing the bounds in Lemma 3.3.3 we obtain

$$\begin{aligned}\mathrm{GAP}(\bar{x}_{N,i}) \leq \left(\frac{\gamma_0^r(N+1)^{1-0.5r}}{2(1-0.5r)}\right)^{-1}\Bigg(&\frac{2M_f\eta_0\gamma_0^r(N+1)^{1-0.5r-b}}{1-0.5r-b}\\&+\frac{2M_X^2(N+1)^{0.5(1-r)}}{\gamma_0^{1-r}}+\gamma_0^r C_F\left(\|\bar{x}_{0,m}-x_{0,1}\|+\|\bar{x}_{0,i}-\bar{x}_{0,m}\|\right)\\&+\left(\frac{\left(C_F+\eta_0 C_f\right)^2}{2}+\frac{(m-i)C_F\left(C_F+\eta_0 C_f\right)}{m}\right)\frac{\gamma_0^{r+1}(N+1)^{1-0.5(1+r)}}{(1-0.5(1+r))}\Bigg).\end{aligned}$$

Rearranging the terms we obtain

$$\begin{aligned}\mathrm{GAP}(\bar{x}_{N,i}) \leq (2-r)\Bigg(&\frac{2M_X^2}{\gamma_0(N+1)^{0.5}}+\frac{2M_f\eta_0}{(1-0.5r-b)(N+1)^b}\\&+\frac{C_F\|\bar{x}_{0,m}-x_{0,1}\|+C_F\,\|\bar{x}_{0,i}-\bar{x}_{0,m}\|}{(N+1)^{1-0.5r}}\\&+\left(\frac{\left(C_F+\eta_0 C_f\right)^2}{2}+\frac{(m-i)C_F\left(C_F+\eta_0 C_f\right)}{m}\right)\frac{\gamma_0}{(1-0.5(1+r))(N+1)^{0.5}}\Bigg).\end{aligned}$$

Note that from $b<0.5$ and $0\leq r<1$ we have $1-0.5r\geq b$. Hence equation (3.4.14) holds. ■

Remark 3.4.1 (Iteration complexity of Algorithm 4) Consider the rate results presented by relations (3.4.13) and (3.4.14). Let us choose $r:=0$ and suppose $\gamma_k:=\frac{(C_F+C_f)^{-1}}{\sqrt{k+1}}$ and $\eta_k:=\frac{1}{\sqrt[4]{k+1}}$ for $k\geq 0$. Let $\epsilon>0$ be an arbitrary small scalar such that $f\left(\bar{x}_{N_\epsilon,i}\right)-$

$f^* + \text{GAP}\left(\bar{x}_{N_\epsilon,i}\right) < \epsilon$ for all $i \in [m]$. Then, we obtain the iteration complexity of $N_\epsilon = \mathcal{O}((C_F + C_f)^4 \epsilon^{-4})$ for each agent. Interestingly, this iteration complexity matches the complexity of the proposed method in our earlier work [52] for addressing formulation (P_2) in a centralized regime where the information of the objective function f is globally known. Importantly, this indicates that there is no sacrifice in the iteration complexity in addressing the distributed formulation (P_2). Another important observation to make is that the iteration complexity of the proposed distributed method is independent of the number of agents m.

3.5 Rate Analysis in the Solution Space

In this section we study the convergence rate properties of the proposed method in the solution space. To this end, we compare the sequences generated from Algorithm 4 with the Tikhonov trajectory (formally introduced in Definition 3.5.1). Throughout this section, we make the following additional assumption.

Assumption 3.5.1 *Consider problem* (P_2). *For all* $i \in [m]$ *let the component function* $f_i : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable and* $\mu_{f_i}-$*strongly convex over the set* X.

Note that under this assumption, the equation (3.3.1) in Algorithm 4 can be written as

$$x_{k,i+1} := \mathcal{P}_X\left(x_{k,i} - \gamma_k\left(F_i\left(x_{k,i}\right) + \eta_k \nabla f_i\left(x_{k,i}\right)\right)\right). \tag{3.5.1}$$

Next, we comment on the strong convexity parameter of the global objective function.

Remark 3.5.1 Under Assumption 3.5.1, we note that any function f_i is strongly convex with a parameter $\mu_{\min} \triangleq \min_{i\in[m]} \mu_{f_i}$. This also implies that the global function $f(x) \triangleq \sum_{i=1}^m f_i(x)$ is $m\mu_{\min}-$ strongly convex. Another implication is that under Assumption 3.3.1(b), (c), and Assumption 3.5.1, problem (P_2) has a unique optimal solution. Throughout this section, we denote the unique optimal solution of (P_2) by x^*.

3.5.1 Preliminaries

Here we provide some preliminary results that will be used later. We first introduce the notion of Tikhonov trajectory.

Definition 3.5.1 (Tikhonov trajectory) Let $\{\eta_k\} \subseteq \mathbb{R}^n_{++}$ be a given sequence. Consider a class of regularized VI problems for $k \geq 0$ given by

$$\text{VI}\left(X, \sum_{i=1}^m \left(F_i + \eta_k \nabla f_i\right)\right). \tag{3.5.2}$$

Let $x^*_{\eta_k}$ denote the unique solution to (3.5.2). The sequence $\{x^*_{\eta_k}\}$ is called the Tikhonov trajectory associated with problem (P_2).

In the following, we establish the convergence of the Tikhonov trajectory to the unique optimal solution of problem (P_2). This result will be used later in Theorem 3.5.1.

Lemma 3.5.1 (Properties of the Tikhonov trajectory) *Consider problem* (P_2) *and Definition 3.5.1. Let Assumption 3.3.1(b), (c), and Assumption 3.5.1 hold. Then:*
(a) Let $\{\eta_k\}$ *be a strictly positive sequence such that* $lim_{k\to\infty}\eta_k = 0$. *Then,* $\lim_{k\to\infty} x^*_{\eta_k}$ *exists and is equal to* x^*.
(b) For any two nonnegative integers k_1 *and* k_2, *we have* $\left\|x^*_{\eta_{k_2}} - x^*_{\eta_{k_1}}\right\| \leq \frac{C_f}{m\mu_{min}}\left|1 - \frac{\eta_{k_2}}{\eta_{k_1}}\right|$.

Proof. The proof can be done in a similar fashion to the proof of Lemma 4.5 in [52]. ■

Next we obtain a recursive bound on an error metric that is characterized by the sequence $\{x_k\}$ in Definition 3.3.1 and the Tikhonov trajectory. This result will be utilized in Theorem 3.5.1.

Lemma 3.5.2 *Let* $\{x_k\}$ *be given by Definition 3.3.1. Let Assumption 3.3.1(b), (c), and Assumption 3.5.1 hold. Suppose* $\{\gamma_k\}$ *and* $\{\eta_k\}$ *are strictly positive and nonincreasing such that* $\gamma_0\eta_0\mu_{min} \leq 0.5$. *Then for any* $k \geq 1$ *we have*

$$\begin{aligned}\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \leq (1 - \gamma_k\eta_k\mu_{min})\left\|x_k - x^*_{\eta_{k-1}}\right\|^2 \\ + \frac{1.5C_f^2}{m^2\gamma_k\eta_k\mu_{min}^3}\left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2 + \gamma_k^2\left(C_F + \eta_k C_f\right)^2. \qquad (3.5.3)\end{aligned}$$

Proof. From Algorithm 4, the nonexpansivity of the projection, and $x^*_{\eta_k} \in X$, for any $i \in [m]$ and $k \geq 1$ we have

$$\begin{aligned}\left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq \left\|x_{k,i} - x^*_{\eta_k}\right\|^2 + \gamma_k^2\left\|F_i(x_{k,i}) + \eta_k\nabla f_i\left(x_{k,i}\right)\right\|^2 \\ - 2\gamma_k\left(F_i(x_{k,i}) + \eta_k\nabla f_i\left(x_{k,i}\right)\right)^T\left(x_{k,i} - x^*_{\eta_k}\right).\end{aligned}$$

From the definition of C_F and C_f we have

$$\begin{aligned}\left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq \left\|x_{k,i} - x^*_{\eta_k}\right\|^2 + \gamma_k^2\left(\frac{C_F + \eta_k C_f}{m}\right)^2 + \\ 2\gamma_k\left(F_i(x_{k,i}) + \eta_k\nabla f_i\left(x_{k,i}\right)\right)^T\left(x^*_{\eta_k} - x_{k,i}\right).\end{aligned}$$

From the strong monotonicity of ∇f_i and the monotonicity of F_i we can write

$$\begin{aligned}2\gamma_k\left(F_i(x_{k,i}) + \eta_k\nabla f_i\left(x_{k,i}\right)\right)^T\left(x^*_{\eta_k} - x_{k,i}\right) \leq 2\gamma_k F_i\left(x^*_{\eta_k}\right)^T\left(x^*_{\eta_k} - x_{k,i}\right) \\ + 2\gamma_k\eta_k\left(\nabla f_i\left(x^*_{\eta_k}\right)^T\left(x^*_{\eta_k} - x_{k,i}\right) - \mu_{\min}\left\|x^*_{\eta_k} - x_{k,i}\right\|^2\right).\end{aligned}$$

From the preceding two relations we obtain

$$\begin{aligned}\left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq (1 - 2\gamma_k\eta_k\mu_{\min})\left\|x_{k,i} - x^*_{\eta_k}\right\|^2 + \gamma_k^2\left(\frac{C_F + \eta_k C_f}{m}\right)^2 \\ + 2\gamma_k F_i\left(x^*_{\eta_k}\right)^T\left(x^*_{\eta_k} - x_{k,i}\right) + 2\gamma_k\eta_k\nabla f_i\left(x^*_{\eta_k}\right)^T\left(x^*_{\eta_k} - x_{k,i}\right).\end{aligned}$$

Adding and subtracting $2\gamma_k F_i(x^*_{\eta_k})^T x_k + 2\gamma_k\eta_k \nabla f_i(x^*_{\eta_k})^T x_k$ in the previous relation we get

$$\begin{aligned}\left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq{}& (1 - 2\gamma_k\eta_k\mu_{\min}) \left\|x_{k,i} - x^*_{\eta_k}\right\|^2 + \gamma_k^2 \left(\frac{C_F + \eta_k C_f}{m}\right)^2 \\ &+ 2\gamma_k F_i \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) + 2\gamma_k \Big(\eta_k \nabla f_i \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) \\ &+ \left|F_i \left(x^*_{\eta_k}\right)^T \left(x_k - x_{k,i}\right)\right|\Big) + 2\gamma_k\eta_k \left|\nabla f_i \left(x^*_{\eta_k}\right)^T \left(x_k - x_{k,i}\right)\right|.\end{aligned}$$

Employing the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}\left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq{}& (1 - 2\gamma_k\eta_k\mu_{\min}) \left\|x_{k,i} - x^*_{\eta_k}\right\|^2 \\ &+ \gamma_k^2 \left(\frac{C_F + \eta_k C_f}{m}\right)^2 + 2\gamma_k F_i \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) \\ &+ 2\gamma_k\eta_k \nabla f_i \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) + \frac{2\gamma_k \left(C_F + \eta_k C_f\right)}{m} \left\|x_k - x_{k,i}\right\|.\end{aligned}$$

Next we take summations over $i \in [m]$ from both sides. Recall $f(x) \triangleq \sum_{i=1}^m f_i(x)$ and $F(x) \triangleq \sum_{i=1}^m F_i(x)$. Using Definition 3.3.1 for $x_{k,1}$ and recalling $1 - 2\gamma_k\eta_k\mu_{\min} < 1$ we have

$$\begin{aligned}\sum_{i=1}^m \left\|x_{k,i+1} - x^*_{\eta_k}\right\|^2 \leq{}& (1 - 2\gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_k}\right\|^2 + 2\gamma_k\eta_k \nabla f \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) \\ &+ \sum_{i=2}^m \left\|x_{k,i} - x^*_{\eta_k}\right\|^2 + \gamma_k^2 \frac{\left(C_F + \eta_k C_f\right)^2}{m} + 2\gamma_k F \left(x^*_{\eta_k}\right)^T \left(x^*_{\eta_k} - x_k\right) \\ &+ \frac{2\gamma_k \left(C_F + \eta_k C_f\right)}{m} \sum_{i=1}^m \left\|x_k - x_{k,i}\right\|. \qquad (3.5.4)\end{aligned}$$

From Lemma 3.3.2, $\|x_k - x_{k,i}\| \leq (i-1)\gamma_k \left(C_F + \eta_k C_f\right)/m$ for all $i \in [m]$. Invoking this relation and Definition 3.3.1 we obtain

$$\begin{aligned}\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \leq{}& (1 - 2\gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_k}\right\|^2 + \gamma_k^2 \left(C_F \right. \\ &\left. + \eta_k C_f\right)^2 + 2\gamma_k \underbrace{\left(F \left(x^*_{\eta_k}\right) + \eta_k \nabla f \left(x^*_{\eta_k}\right)\right)^T \left(x^*_{\eta_k} - x_k\right)}_{\text{term 4}}.\end{aligned}$$

From Definition 3.5.1, $x^*_{\eta_k}$ is the solution to problem (3.5.2). Recalling $x_k \in X$, we have term $4 \leq 0$. We obtain

$$\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \leq (1 - 2\gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_k}\right\|^2 + \gamma_k^2 \left(C_F + \eta_k C_f\right)^2. \qquad (3.5.5)$$

Next, consider the term $\|x_k - x^*_{\eta_k}\|^2$ as follows.

$$\left\|x_k - x^*_{\eta_k}\right\|^2 = \left\|x_k - x^*_{\eta_{k-1}}\right\|^2 + \left\|x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\|^2 + \underbrace{2 \left(x_k - x^*_{\eta_{k-1}}\right)^T \left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right)}_{\text{term 5}}.$$

Next, we bound term 5 by recalling $2a^Tb \leq \|a\|^2/\alpha + \alpha\|b\|^2$ where $a, b \in \mathbb{R}^n$ and $\alpha > 0$. For $\alpha := 1/\gamma_k\eta_k\mu_{\min}$, bounding term 5 in the preceding inequality we obtain

$$\left\|x_k - x^*_{\eta_k}\right\|^2 = (1 + \gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_{k-1}}\right\|^2 + \left(1 + \frac{1}{\gamma_k\eta_k\mu_{\min}}\right) \left\|x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\|^2.$$

From Lemma 3.5.1(b) we obtain

$$\begin{aligned}\left\|x_k - x^*_{\eta_k}\right\|^2 \leq{}& (1 + \gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_{k-1}}\right\|^2 \\ &+ \left(1 + \frac{1}{\gamma_k\eta_k\mu_{\min}}\right) \frac{C_f^2}{m^2\mu_{\min}^2} \left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2. \end{aligned} \tag{3.5.6}$$

From equations (3.5.5) and (3.5.6) we have

$$\begin{aligned}\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \leq{}& (1 - 2\gamma_k\eta_k\mu_{\min}) (1 + \gamma_k\eta_k\mu_{\min}) \left\|x_k - x^*_{\eta_{k-1}}\right\|^2 \\ &+ \left(1 + \frac{1}{\gamma_k\eta_k\mu_{\min}}\right) \frac{C_f^2}{m^2\mu_{\min}^2} \left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2 + \gamma_k^2 (C_F + \eta_k C_f)^2.\end{aligned}$$

Using $0 < \gamma_k\eta_k\mu_{\min} \leq 0.5$ we have the desired result. ■

3.5.2 Convergence analysis

In this section, our goal is to derive a non-asymptotic convergence rate statement that relates the generated sequences by Algorithm 4 to the Tikhonov trajectory. We begin with providing a class of sequences for the stepsize and the regularization parameter and prove some properties for them that will be used in the analysis.

Definition 3.5.2 (Stepsize and regularization parameter) *Let $\gamma_k := \frac{\gamma}{(k+\Gamma)^a}$ and $\eta_k := \frac{\eta}{(k+\Gamma)^b}$ for all $k \geq 0$ where γ, η, Γ, a and b are strictly positive scalars. Let $a > b$, $a + b < 1$, and $3a + b < 2$. Assume that $\Gamma \geq 1$ and it is sufficiently large such that $\Gamma^{a+b} \geq 2\gamma\eta\mu_{min}$ and $\Gamma^{1-a-b} \geq \frac{4}{\gamma\eta\mu_{min}}$.*

Lemma 3.5.3 *Consider Definition 3.5.2. The following results hold.*
(i) $\{\gamma_k\}$ and $\{\eta_k\}$ are strictly positive and nonincreasing such that $\gamma_0\eta_0\mu_{min} \leq 0.5$.
(ii) For all integers k_1 and k_2 such that $k_2 \geq k_1 \geq 0$ we have $1 - \frac{\eta_{k_2}}{\eta_{k_1}} \leq \frac{k_2 - k_1}{k_2 + \Gamma}$.
(iii) For all $k \geq 1$ we have $\frac{1}{\gamma_k^3\eta_k}\left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2 \leq \frac{1}{\gamma^3\eta\Gamma^{2-3a-b}}$.
(iv) For all $k \geq 1$ we have $\frac{\gamma_{k-1}}{\eta_{k-1}} \leq \frac{\gamma_k}{\eta_k}(1 + 0.5\gamma_k\eta_k\mu_{min})$.

Proof. See Appendix A.9. ■

The main contribution in this section is presented by Theorem 3.5.1 where we derive agent-specific rates relating the sequences generated by Algorithm 4 with the Tikhonov trajectory.

Theorem 3.5.1 (Comparison with the Tikhonov trajectory) *Consider problem* (P_2). *Let Assumption 3.3.1(b), (c), and Assumption 3.5.1 hold. Consider* $\{x_k\}$ *and* $\{x^*_{\eta_k}\}$ *given in Definitions 3.3.1 and 3.5.1, respectively. Let the stepsize sequence* $\{\gamma_k\}$ *and the regularization sequence* $\{\eta_k\}$ *be given by Definition 3.5.2. Then for all* $k \geq 0$ *and all* $i \in [m]$ *we have*

$$\left\|x_{k+1,i} - x^*_{\eta_k}\right\|^2 \leq \frac{2(i-1)^2 \left(C_F + \eta_0 C_f\right)^2 \gamma^2}{m^2(k+\Gamma+1)^{2a}} + \frac{2\tau B_0 \gamma}{\mu_{min}\eta(k+\Gamma)^{a-b}},$$

where $\tau \triangleq max\left\{\mu_{min}\eta\gamma^{-1}B_0^{-1}\Gamma^{a-b}\left\|x_1 - x^*_{\eta_0}\right\|^2, 2\right\}$ *and* $B_0 \triangleq \frac{1.5C_f^2}{m^2\mu_{min}^3\gamma^3\eta\Gamma^{2-3a-b}} + \left(C_F + \eta_0 C_f\right)^2$.

Proof. Consider (3.5.3). From Lemma 3.5.3, for $k \geq 1$ we have

$$\begin{aligned}\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \leq &\left(1 - \gamma_k\eta_k\mu_{\min}\right)\left\|x_k - x^*_{\eta_{k-1}}\right\|^2 \\ &+ \frac{1.5C_f^2\gamma_k^2}{m^2\mu_{\min}^3\gamma^3\eta\Gamma^{2-3a-b}} + \gamma_k^2\left(C_F + \eta_0 C_f\right)^2.\end{aligned}$$

Let us define the terms $v_k \triangleq \left\|x_k - x^*_{\eta_{k-1}}\right\|^2$, $\alpha_k \triangleq \gamma_k\eta_k\mu_{\min}$, and $\beta_k \triangleq B_0\gamma_k^2$ for $k \geq 1$. Therefore, for all $k \geq 1$ we have

$$v_{k+1} \leq (1 - \alpha_k)v_k + \beta_k. \tag{3.5.7}$$

From Lemma 3.5.3(iii), for all $k \geq 1$ we have

$$\frac{\beta_{k-1}}{\alpha_{k-1}} \leq \frac{B_0\gamma_k}{\mu_{\min}\eta_k}(1 + 0.5\gamma_k\eta_k\mu_{\min}) = \frac{\beta_k}{\alpha_k}(1 + 0.5\alpha_k). \tag{3.5.8}$$

Next, we show that $v_{k+1} \leq \tau\frac{\beta_k}{\alpha_k}$ for all $k \geq 0$. We apply induction on $k \geq 0$. Note that this relation holds for $k := 0$ as an implication of the definition of τ. Suppose $v_k \leq \tau\frac{\beta_{k-1}}{\alpha_{k-1}}$ holds for some $k \geq 1$. From (3.5.7) we obtain $v_{k+1} \leq (1 - \alpha_k)\frac{\beta_{k-1}}{\alpha_{k-1}}\tau + \beta_k$. Using the upper bound for the right-hand side given by (3.5.8) we have

$$\begin{aligned}v_{k+1} \leq &\tau(1 - \alpha_k)(1 + 0.5\alpha_k)\frac{\beta_k}{\alpha_k} + \beta_k \\ &= \tau(1 - \alpha_k + 0.5\alpha_k - 0.5\alpha_k^2)\frac{\beta_k}{\alpha_k} + \beta_k = \tau\frac{\beta_k}{\alpha_k} \\ &\quad - \tau(1 - 0.5)\beta_k - 0.5\tau\alpha_k\beta_k + \beta_k \leq \tau\frac{\beta_k}{\alpha_k} + (1 - 0.5\tau)\beta_k.\end{aligned}$$

From the definition of τ, $\tau \geq 2$ implying that $1 - 0.5\tau \leq 0$. This completes the proof of induction. Recall from Lemma 3.3.2 that we have $\|x_k - x_{k,i}\| \leq (i-1)\gamma_k\left(C_F + \eta_k C_f\right)/m$ for any $i \in [m]$. For all $k \geq 0$ and $i \in [m]$ we have

$$\begin{aligned}\left\|x_{k+1,i} - x^*_{\eta_k}\right\|^2 &\leq 2\left\|x_{k+1,i} - x_{k+1}\right\|^2 + 2\left\|x_{k+1} - x^*_{\eta_k}\right\|^2 \\ &\leq \frac{2(i-1)^2\left(C_F + \eta_0 C_f\right)^2\gamma_{k+1}^2}{m^2} + \frac{2\tau B_0\gamma_k}{\mu_{\min}\eta_k} \\ &= \frac{2(i-1)^2\left(C_F + \eta_0 C_f\right)^2\gamma^2}{m^2(k+\Gamma+1)^{2a}} + \frac{2\tau B_0\gamma}{\mu_{\min}\eta(k+\Gamma)^{a-b}}.\end{aligned}$$

Hence the proof is completed. ■

3.6 Numerical Results

In this section we present the implementation results of Algorithm 4 in addressing two of the motivating examples discussed in Chapter 1. These include a traffic equilibrium problem and a soft-margin support vector classification problem.

(i) Traffic Equilibrium Problem. For an illustrative example, we consider the transportation network in [23]. We first describe the network and present the NCP formulation. Then, we implement Algorithm 4 to solve model (1.2.2) and compute the best equilibrium.

Consider a transportation network with the set of nodes $\{n_1, n_2\}$ and the set of directed arcs $\{a_1, a_2, a_3, a_4, a_5\}$. As shown in Figure 4, arcs a_1, a_2, and a_3 are directed from node n_1 to node n_2, and arcs a_4 and a_5 are directed in the reverse way. Note that a_1 and a_4 construct a two-way road. The same holds for a_2 and a_5. We let $d \triangleq [d_1, d_2]^T$ denote the expected travel demand vector where d_1 and d_2 correspond to the demand from n_1 to n_2, and from n_2 to n_1, respectively. Let the vector $h \triangleq [h_1, \ldots, h_5]^T$ denote the traffic flow on the arcs. The travel cost on each arc is assumed to be a linear function in terms of h. More precisely, the travel cost on arc i is equal to $[Ch + q]_i$ where we let the cost matrix $C \in \mathbb{R}^{5\times 5}$ and vector $q \in \mathbb{R}^5$ be given by

$$C := \begin{bmatrix} 0.92 & 0 & 0 & 5 & 0 \\ 0 & 5.92 & 0 & 0 & 5 \\ 0 & 0 & 10.92 & 0 & 0 \\ 2 & 0 & 0 & 10.92 & 0 \\ 0 & 1 & 0 & 0 & 15.92 \end{bmatrix}, \qquad q := \begin{bmatrix} 1000 \\ 950 \\ 3000 \\ 1000 \\ 1300 \end{bmatrix}. \tag{3.6.1}$$

We note that the matrix C is positive semidefinite. The diagonal values of C are rounded by two decimal places for the ease of presentation. Intuitively speaking, the structure of C implies that the cost of each arc in a two-way road depends on the flows on the both directions. Let $u \triangleq [u_1, u_2]^T$ denote the (unknown) vector of minimum travel costs between the origin-destination (OD) pairs, i.e., u_1 denotes the minimum travel cost from n_1 to n_2, and u_2 denotes the minimum travel cost from n_2 to n_1. The Wardrop user equilibrium principle represents the path choice behavior of the users based on the following rationale: (i) For any OD pair among all possible arcs, users tend to choose the arc(s) with the minimum cost. (ii) For any OD pair, the arc(s) that have the minimum cost will have positive flows and will have equal costs. (iii) For any OD pair, arcs with higher costs than the minimum value will

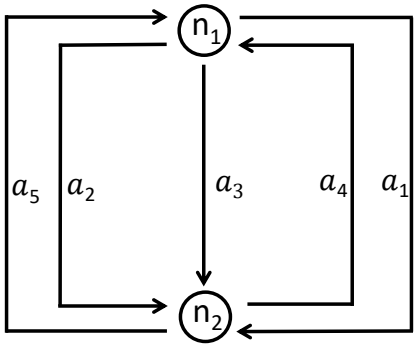

Figure 4: A transportation network with 2 nodes and 5 arcs

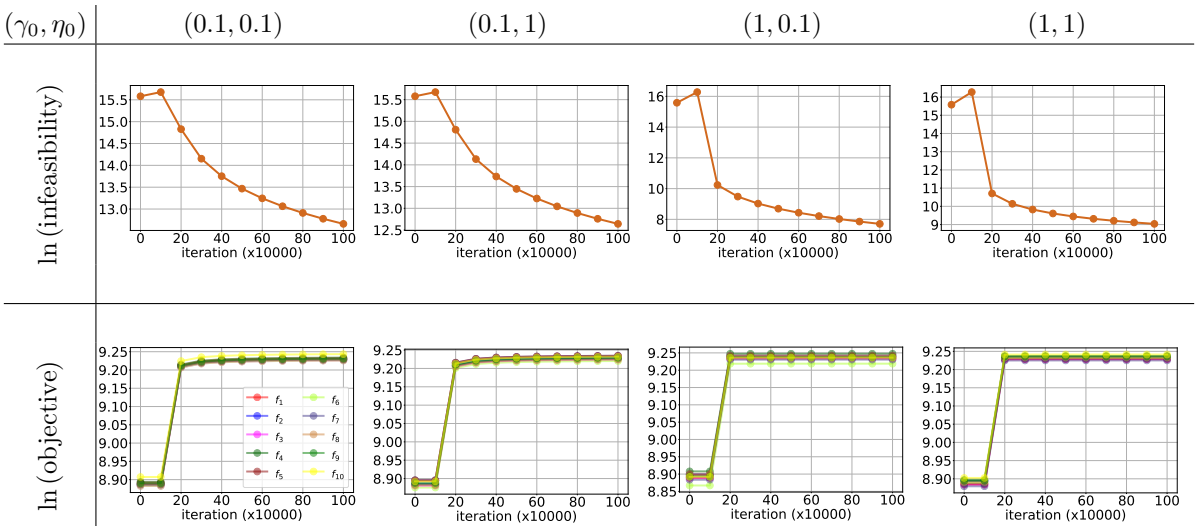

Figure 5: Performance of Algorithm 4 in terms of infeasibility and the objective function value for finding the best equilibrium in the transportation network problem

have no flows. Mathematically the Wardrop's principle can be characterized as

$$0 \leq Ch + q - B^T u \perp h \geq 0, \quad 0 \leq Bh - d \perp u \geq 0, \tag{3.6.2}$$

where $B \in \mathbb{R}^{2\times 5}$ denotes the (OD pair, arc)-incidence matrix given as $B := \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$. Throughout, we assume that the demand vector d and the cost vector q are subject to uncertainties. Let us define decision vector $x \in \mathbb{R}^7$, random variable $\xi \in \mathbb{R}^{10}$, and stochastic mapping $F(\bullet, \xi) : \mathbb{R}^7 \to \mathbb{R}^7$ as

$$x \triangleq \begin{bmatrix} h \\ u \end{bmatrix}, \qquad \xi \triangleq \begin{bmatrix} \tilde{d} \\ \tilde{q} \end{bmatrix}, \qquad F(x, \xi) \triangleq \begin{bmatrix} C & -B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} h \\ u \end{bmatrix} + \begin{bmatrix} \tilde{q} \\ -\tilde{d} \end{bmatrix}.$$

Then from Example 1.2.2 in Chapter 1, the Wardrop equation (3.6.2) can be characterized as VI $\big(\mathbb{R}^7_+, \mathbb{E}[F(\bullet, \xi)]\big)$. Notably due to positive semidefinite property of C, the mapping $\mathbb{E}[F(\bullet, \xi)]$ is merely monotone. Consequently, the aforementioned VI may have multiple equilibria. Among them, we seek to find the best equilibrium with respect to a welfare function f defined as the expected total travel time over the network by all users, i.e., $f(x) \triangleq \mathbb{E}\left[(Ch + \tilde{q})^T \mathbf{1}_5\right]$ where $\mathbf{1}_5 \in \mathbb{R}^n$ denotes a vector with all unit elements.

Set-up. For this experiment, we assume that $\tilde{d}_1 \sim \mathcal{N}(210, 10)$, $\tilde{d}_2 \sim \mathcal{N}(120, 10)$. Also for $i = 1, \ldots, 5$ we let $\tilde{q}_i$ be normally distributed with the mean equal to q_i and the standard deviation of 300, where the vector q is given by (3.6.1). Following the formulation (1.2.2) we generate 1000 samples for each parameter and distribute the data equally among 10 agents. We let $\gamma_k := \frac{\gamma_0}{\sqrt{k+1}}$ and $\eta_k := \frac{\eta_0}{(k+1)^{0.25}}$ and consider different values for the initial stepsize γ_0 and the initial regularization parameter η_0. The results are as shown in Figure 5. We use standard averaging by assuming that $r = 0$. Notably for quantifying the infeasibility, we consider the metric $\phi(x) \triangleq \|\max\{0, -x\}\|^2 + \|\max\{0, -F(x)\}\|^2 + \left|x^T F(x)\right|$, where $F(x) \triangleq \sum_{i=1}^m F_i(x)$ and $F_i(x) \triangleq \sum_{\ell \in \mathcal{S}_i} F(x, \xi_\ell)$. Note that $\phi(x) = 0$ if and only if $0 \leq x \perp F(x) \geq 0$. We choose this metric over the dual gap function employed earlier in the analysis because in

this particular example, the dual gap function becomes infinity at some of the evaluations of the generated iterates. This is due to the unboundedness of the set $X := \mathbb{R}^n_+$. Unlike the dual gap function, $\phi(x)$ stays bounded and is more suitable to plot.

Insights. In Figure 5 we observe that in all four different settings the infeasibility metric decreases as the algorithm proceeds. This indeed implies that the generated iterates by the agents tend to satisfy the NCP constraints with an increasing accuracy. In terms of the suboptimality metric we observe that the each agent's objective value becomes more and more stable over time. Intuitively this implies that the agents asymptotically reach to an equilibrium. We should note that although the function f is minimized, it is minimized only over the set of equilibria. The fact that the objective values in Figure 5 are not necessarily decreasing is mainly because of the impact of feasibility violation of the iterates with respect to the NCP constraints throughout the implementations. As evidenced, pair-IG performs with much robustness to the choice of the initial values of γ_0 and η_0.

3.7 Conclusion

We introduce a new unifying formulation for distributed constrained optimization where the constraint set is characterized as the solution set of a merely monotone variational inequality problem. We develop an iteratively regularized incremental gradient method where at each iteration agents communicate over a cycle graph to update their iterates using their local information about the objective function and the mapping. We derive new iteration complexity bounds in terms of the global objective function and a suitably defined infeasibility metric. To analyze the convergence in the solution space, we also provide non-asymptotic agent-wise convergence rate statements that relate the iterate of each agent with that of the Tikhonov trajectory. We validate the theoretical results on an illustrative transportation network problem.

CHAPTER IV

ILL-POSED HIGH-DIMENSIONAL OPTIMIZATION PROBLEMS

Motivated by high-dimensional nonlinear optimization problems as well as ill-posed optimization problems arising in image processing, in this chapter we consider a bilevel optimization model where we seek among the optimal solutions of the inner level problem, a solution that minimizes a secondary metric, discussed further in Section 4.1. Our goal is to address the high-dimensionality of the bilevel problem, and the nondifferentiability of the objective function. Minimal norm gradient, sequential averaging, and iterative regularization are some of the recent schemes developed for addressing the bilevel problem. But none of them address the high-dimensional structure and nondifferentiability. Section 4.2 includes the summary of literature for addressing problem (1.1.5) and the research gap. We address problem (1.1.5) by proposing a randomized block iterative regularized gradient scheme. The outline of algorithm and the required preliminaries are provided in Section 4.3. We establish the convergence of the sequence generated by Algorithm 5 to the unique solution of the bilevel problem of interest. Furthermore, we derive a rate of convergence with respect to the inner level objective function. Section 4.4 includes the convergence analysis and rate results. In Section 4.5, we demonstrate the performance of Algorithm 5 in solving the ill-posed problems arising in image processing. Section 4.6 summarizes the chapter with conclusions.

4.1 Problem Formulation

We consider a special case of Problem (P_1), large-scale bilevel optimization as follows

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \operatorname{argmin}\{g(x) : x \in X\}, \end{aligned} \tag{1.1.5}$$

where functions f and g are defined as $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. In particular, here we consider the case where the set X has a block structure, i.e., $X = \prod_{i=1}^{d} X_i$, where $X_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^{d} n_i = n$. The size of the solution space (n) can be of the order $10^8 - 10^{12}$. Under the convexity of f, g, and set X_i, problem (1.1.5) can be easily captured by (P_1) such that F is a gradient map, given as $F(x) \triangleq (\nabla_{x^{(1)}} g(x); \ldots; \nabla_{x^{(d)}} g(x))$ with set $X = \prod_{i=1}^{d} X_i$.

The content of this chapter has been published in the Proceedings of 2019 American Control Conference [49].

4.2 Existing Methods and Research Gap

Minimal norm gradient, sequential averaging, and iterative regularization are some of the known schemes developed for addressing problem (1.1.5) are summarized in Table 2 and are described as follows. Given $\eta > 0$, consider the following regularized problem

Table 2: Comparison of schemes for solving bilevel optimization problem

Ref.	Problem formulation	Assumption	Rate
[82]	$\begin{array}{c}\text{minimize } g(x)\\ \text{s.t.} \quad x \in \operatorname{argmin}\{f(x) : x \in X\}\end{array}$	f and g both smooth and convex	–
[10]	$\begin{array}{c}\text{minimize } g(x)\\ \text{s.t.} \quad x \in \operatorname{argmin}\{f(x) : x \in X\}\end{array}$	f convex, Lipschitz cont. g strongly conv.	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$
[75]	$\begin{array}{c}\text{minimize } g(x)\\ x \in \operatorname{argmin}\{f_1(x) + f_2(x) : x \in \mathbb{R}^n\}\end{array}$	f_1, f_2 convex. g is strongly convex. f_1, g Lipschitz cont.	$\mathcal{O}\left(\frac{1}{k}\right)$
[96]	$\begin{array}{c}\text{minimize } \|x\|\\ \text{s. t. } x \in \operatorname{SOL}(X, F), \text{ where } F(x) \triangleq f(x, \xi)\end{array}$	F is monotone and continuous.	$\mathcal{O}\left(\frac{1}{k^{1/6-\delta}}\right)$
[90]	$\min_{x_i \in X_i, z \in Z} \sum_{i=1}^N f_i(x_i) \quad \text{s.t. } Dx + Hz = 0$	f_i is convex and possibly non-smooth.	$\mathcal{O}\left(\frac{1}{k}\right)$
[8]	$\begin{array}{c}\text{for } \mathcal{G}(\mathcal{N}, \mathcal{E}), \ \min \sum_{i \in \mathcal{N}} \xi_i(x) + f_i(x)\\ \text{s.t. } x \in \mathbb{R}^n, \ x_i = x_j \ \text{ for all } (i,j) \in \mathcal{E}\end{array}$	ξ_i, f_i are convex, f_i Lipschitz continuous.	$\mathcal{O}(1/k)$
[92]	$\begin{array}{c}\min\limits_x \ g_1(x) + g_2(x)\\ \text{s.t. } Ax = b\end{array}$	g_1 is convex and Lipschitz continuous. g_2 is convex.	$\mathcal{O}(1/k^2)$
This work	$\begin{array}{c}\text{minimize } g(x)\\ x \in \operatorname{argmin}\{f(x) : x \in X\}; \ X = \prod_{i=1}^d X_i\end{array}$	f is convex and g is strongly convex.	$\mathcal{O}\left(\frac{1}{k^{0.5-\delta}}\right)$

$$\begin{array}{ll}\text{minimize} & g(x) + \eta f(x)\\ \text{s.t.} & x \in X.\end{array} \tag{P_η}$$

Particular case of the above problem (P_η) was discussed in (1.1.6). Tikhonov in [85] showed that under some assumptions, the solution of regularized problem (P_η) converges to the solution of the inner level problem of (1.1.5) as the regularization parameter η goes to zero. Later, the threshold value of η, under which the solution of (P_η) is the same as the solution of the inner level problem of (1.1.5), was studied under the area of *exact regularization* [32, 38]. There have been numerous theoretical studies in the '80s, '90s [11, 12, 18, 38] and more recently [16, 24] on finding the suitable η, but in practice there is not much guidance on tuning this parameter. Finding a suitable η necessitates solving a sequence of problems (P_η) for η_k, where $\eta_k \to 0$. This *two-loop* scheme is significantly inefficient, especially in high-dimensional spaces.

In the past decade, interest has been shifted to solving the bilevel problem (1.1.5) using *single-loop schemes*. Solodov in [82] showed that for both functions g and f in (1.1.5) with Lipschitz gradient, and f to be a composite function with the indicator function, solutions to (1.1.5) can be found by an iterative regularized gradient descent with sequence $\eta_k \to 0$ and $\sum_{k=1}^\infty \eta_k = \infty$. In (P_η), when g is ℓ_2 norm in variational inequality regimes, Yousefian et

al. [96] showed that solution to (1.1.5) can be found by employing an iterative regularized smoothing gradient scheme.

In 2014, the minimal norm gradient (MNG) scheme was proposed [10]. MNG is a *two-loop* scheme where an optimization problem needs to be solved at each iteration k, making MNG to be computationally expensive for the large-scale problems. Later, in [75] a sequential averaging scheme (BiG-SAM) was developed with a rate of convergence $\mathcal{O}(1/k)$. Recently in [33], a general iterative regularized algorithm based on a primal-dual diagonal descent method was proposed to solve (1.1.5).

In all the aforementioned papers, the missing part is addressing the high-dimensional structure, which is common in the high-resolution image processing applications such as hyper-spectral imaging. Our goal is to bridge this gap by developing a *single-loop* randomized block-coordinate iterative regularized subgradient scheme.

High-dimensional nonlinear constrained optimization is another motivating example to our work. One of the popular primal-dual methods is Alternating Direction Method of Multipliers (ADMM) [8,90,92]. One of the underlying assumptions for ADMM is the linearity of the constraints.

4.3 Algorithm Outline

Here we propose Algorithm 5 and include the preliminaries for the convergence and rate analysis.

Assumption 4.3.1 *Consider the optimization problem* (1.1.5)*. Let the following hold:*

(a) Any block i of set X ($X_i \subseteq \mathbb{R}^{n_i}$) is assumed to be nonempty, closed, and convex for all $i = 1, \ldots, d$.

(b) $g : \mathbb{R}^n \to (-\infty, \infty]$ is a nondifferentiable, proper, and convex function.

(c) $f : \mathbb{R}^n \to (-\infty, \infty]$ is a nondifferentiable, proper, and μ-strongly convex function ($\mu > 0$).

(d) $X \subseteq \operatorname{int}(\operatorname{dom}(f) \cap \operatorname{dom}(g))$.

Next, a randomized block-coordinate iterative regularized subgradient Algorithm 5 is proposed for solving (1.1.5). In Algorithm 5, both the sequences of regularization parameter η_k and stepsize parameter γ_k are in terms of iteration k. The update rules of γ_k and γ_k are finalized later (in Theorem 4.4.2). To address the high-dimensionality, at each iteration we update a random block of the iterate x_k. Selection of block i_k at iteration k is governed by Assumption 4.3.2. Finally, averaging is employed which will be helpful in deriving the rate statement.

Assumption 4.3.2 (Block-coordinate selection rule) *At each iteration $k \geq 0$, the random variable i_k is generated from an independent and identically distributed discrete probability distribution such that* $\operatorname{Prob}(i_k = i) = \mathsf{p}_i$ *where $\mathsf{p}_i > 0$ for $i \in \{1, \ldots, d\}$ and $\sum_{i=1}^{d} \mathsf{p}_i = 1$.*

Algorithm 5 Randomized block iterative regularized gradient (RB-IRG)

1: **Initialize:** Set k = 0, select a point $x_0 \in X$, parameters $\gamma_0 > 0$, and $\eta_0 > 0, S_0 = \gamma_0^r$, and $\bar{x}_0 = x_0$.

2: **for** k = 0, 1, ..., N-1 **do**

3: i_k is generated by Assumption 4.3.2.

4: Compute $\tilde{\nabla}_i g(x_k) \in \partial g\left(x_k^{(i)}\right)$ and $\tilde{\nabla}_i f(x_k) \in \partial f\left(x_k^{(i)}\right)$ for $x_k^{(i)} \in X_i$.

5: Update $x_{k+1}^{(i_k)}$:=

$$\begin{cases} \mathcal{P}_{X_i}\left(x_k^{(i)} - \gamma_k\left(\tilde{\nabla}_i g\left(x_k\right) + \eta_k \tilde{\nabla}_i f\left(x_k\right)\right)\right) & \text{if } i = i_k. \\ x_k^{(i)} \quad \text{if } i \neq i_k. \end{cases} \tag{RB-IRG}$$

6: Update $\overline{x}_k$ as following,

$$S_{k+1} = S_k + \gamma_{k+1}^r, \quad \overline{x}_{k+1} = \frac{S_k \overline{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}}. \tag{4.3.1}$$

7: **end for**

4.3.1 Preliminaries

In this subsection, we list all the required preliminaries for the convergence and rate analysis. Throughout, we use x_f^* and $x_{\eta_k}^*$ to denote the unique minimizers of (1.1.5) and (P_η), respectively.

Remark 4.3.1 From Assumptions 4.3.1 (b, c), the objective function of (P_η), is a strongly convex. The feasible region of (P_η) is closed and convex (from Assumption 4.3.1(a)). Therefore (P_η) has a unique minimizer. (cf. Ch. 2 of [30]). Similarly, we can claim that (1.1.5) has a unique minimizer.

Remark 4.3.2 In problem (1.1.5), for any $x_1, x_2 \in X$, for a convex function g and μ-strongly convex function f,
$\left(\tilde{\nabla} g(x_1) - \tilde{\nabla} g(x_2)\right)^T (x_1 - x_2) \geq 0,$
$\left(\tilde{\nabla} f(x_1) - \tilde{\nabla} f(x_2)\right)^T (x_1 - x_2) \geq \mu \|x_1 - x_2\|^2.$

The following lemma is used in proving the convergence.

Lemma 4.3.1 (Lemma 10, pg. 49 of [69]) *Let $\{v_k\}$ be a sequence of nonnegative random variables, where $E[v_0] < \infty$, and let $\{\alpha_k\}$ and $\{\beta_k\}$ be deterministic scalar sequences such that: $\mathsf{E}[v_{k+1}|v_0, \ ..., v_k] \leq (1 - \alpha_k)v_k + \beta_k$ for all $k \geq 0$, $0 \leq \alpha_k \leq 1$, $\beta_k \geq 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, $\lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0$. Then, $v_k \to 0$, a.s., and $\lim_{k\to\infty} \mathsf{E}[v_k] = 0$.*

The next result will be used in our analysis.

Lemma 4.3.2 (Theorem 6, pg. 75 of [54]) *Let $\{u_t\}$ ($\subset \mathbb{R}^n$) be a convergent sequence such that it has a limit point $\hat{u} \in \mathbb{R}^n$ and consider another sequence $\{\alpha_k\}$ of positive numbers*

such that $\sum_{k=0}^{\infty}\alpha_k = \infty$. *Suppose* v_k *is given by* $v_k = \frac{\sum_{t=0}^{k-1}(\alpha_t u_t)}{\sum_{t=0}^{k-1}\alpha_t}$, *for all* $k \geq 1$. *Then* $\lim_{k\to\infty} v_k = \hat{u}$.

Remark 4.3.3 From Assumption 4.3.1 (b, c, d), for all $x \in X$, the set $\partial f(x)$ is nonempty and bounded (cf. Ch. 3 of [9]). Similarly $\partial g(x)$ is nonempty and bounded for all $x \in X$.

Remark 4.3.4 From Remark 4.3.3, let us say that for any $x^{(i)} \in X_i$, there exists a scalar $C_{f,i}$ such that $\left\|\tilde{\nabla}_i f(x)\right\| \leq C_{f,i}$. Let $C_f \triangleq \sqrt{\sum_{i=1}^{d} C_{f,i}^2}$. Now we have, $\left\|\tilde{\nabla} f(x)\right\| \leq C_f$ for all $x \in X$. Similarly, $\left\|\tilde{\nabla} g(x)\right\| \leq C_g$ for all $x \in X$.

In the following lemma, we present the properties of $\{x^*_{\eta_k}\}$, which denotes the of solution of (P_η) for $\eta \in \{\eta_k\}$.

Lemma 4.3.3 *Consider problem* (1.1.5) *and* (P_η). *Let Assumption 4.3.1 hold. Then, for the sequence* $\{x^*_{\eta_k}\}$, *and* x^*_f *for any* $k \geq 1$, *we have*
(a) $\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\| \leq \frac{C_f}{\mu}\left|\frac{\eta_{k-1}}{\eta_k} - 1\right|$.
(b) When $\{\eta_k\}$ *goes to zero,* $\{x^*_{\eta_k}\}$ *converges to* x^*_f.

Proof. Please see the proof of Lemma 2.5.1. ■

Our objective is to show $\|x_{k+1} - x^*_f\| \to 0$. Now from the triangle inequality, $\|x_{k+1} - x^*_{\eta_k}\| \to 0$ and $\|x^*_{\eta_k} - x^*_f\| \to 0$. We know $\|x^*_{\eta_k} - x^*_f\| \to 0$ as $\eta_k \to 0$. Our main objective is to show $\|x_{k+1} - x^*_{\eta_k}\| \to 0$. Next we define an error function which will be used in the convergence analysis.

Definition 4.3.1 *Let Assumption 4.3.2 hold. Then for any* $x, y \in \mathbb{R}^n$, *function* $\mathcal{D}(x,y) = \sum_{i=1}^{d} \mathsf{p}_i^{-1}\left\|x^{(i)} - y^{(i)}\right\|^2$.

The following corollary holds from Definition 4.3.1.

Corollary 4.3.1 *Consider Definition 4.3.1,* p_{max} *and* p_{min} *as defined in the notation, and let Assumption 4.3.2 hold. Then for any* $x, y \in \mathbb{R}^n$, $\mathsf{p}_{max}\mathcal{D}(x,y) \leq \|x-y\|^2 \leq \mathsf{p}_{min}\mathcal{D}(x,y)$.

4.4 Convergence and Rate Analysis

Here we begin with deriving a recursive error bound, that will be used later to show the almost sure convergence.

Lemma 4.4.1 (Recursive relation for $\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)$**)** *Consider problem* (1.1.5) *and* (P_η). *Let Assumptions 4.3.1 and 4.3.2 hold. Let* $\{x_k\}$ *be the sequence generated from Algorithm 1. Let positive sequences* $\{\gamma_k\}$, *and* $\{\eta_k\}$ *be non-increasing and* $\gamma_0\,\eta_0 < 1/\mu\mathsf{p}_{min}$. *Then the following relation holds,*

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq (1 - \mu\gamma_k\eta_k\mathsf{p}_{min})\,\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) + \frac{2C_f^2}{\mathsf{p}_{min}^2\mu^3\gamma_k\eta_k}\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 + 2\gamma_k^2(C_g^2 + \eta_0^2 C_f^2).$$

Proof. Consider $\mathcal{D}(x_{k+1}, x^*_{\eta_k})$. From the Definition 4.3.1,

$$\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) = \sum_{i=1}^{d} \mathsf{p}_i^{-1} \left\| x^{(i)}_{k+1} - x^{*(i)}_{\eta_k} \right\|^2 = \sum_{i=1, i\neq i_k}^{d} \mathsf{p}_i^{-1} \left\| x^{(i)}_{k} - x^{*(i)}_{\eta_k} \right\|^2 + \mathsf{p}_{i_k}^{-1} \underbrace{\left\| x^{(i_k)}_{k+1} - x^{*(i_k)}_{\eta_k} \right\|^2}_{\text{term-1}} \tag{4.4.1}$$

Since $x^*_{\eta_k} \in X$, we have $x^{*(i_k)}_{\eta_k} \in X_{i_k}$. Now from the non-expansive property of projection operator, term-1 becomes,

$$\left\| x^{(i_k)}_{k+1} - x^{*(i_k)}_{\eta_k} \right\|^2 \leq \left\| x^{(i_k)}_{k} - \gamma_k \left(\tilde{\nabla}_{i_k} f\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} g\left(x_k\right) \right) - x^{*(i_k)}_{\eta_k} \right\|^2.$$

From the two preceding relations, we have,

$$\begin{aligned}\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) = &\sum_{i=1, i\neq i_k}^{d} \mathsf{p}_i^{-1} \left\| x^{(i)}_{k} - x^{*(i)}_{\eta_k} \right\|^2 + \mathsf{p}_{i_k}^{-1} \left\| x^{(i_k)}_{k} - x^{*(i_k)}_{\eta_k} \right\|^2 \\ &- 2\,\mathsf{p}_{i_k}^{-1} \gamma_k \left(x^{(i_k)}_{k} - x^{*(i_k)}_{\eta_k} \right)^T \left(\tilde{\nabla}_{i_k} g\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} f\left(x_k\right) \right) + \mathsf{p}_{i_k}^{-1} \underbrace{\gamma_k^2 \left\| \tilde{\nabla}_{i_k} g\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} f\left(x_k\right) \right\|^2}_{\text{term-2}}.\end{aligned} \tag{4.4.2}$$

From Assumptions 4.3.1 (d) and Remark 4.3.4,

$$\text{term-2} = \gamma_k^2 \left\| \tilde{\nabla}_{i_k} g(x_k) + \eta_k \tilde{\nabla}_{i_k} f(x_k) \right\|^2 \leq 2\gamma_k^2 C^2_{g,i_k} + 2\gamma_k^2 \eta_k^2 C^2_{f,i_k}.$$

Thus from (4.4.2), and Definition 4.3.1, we obtain,

$$\begin{aligned}\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) \leq &\mathcal{D}\left(x_k, x^*_{\eta_k}\right) + \mathsf{p}_{i_k}^{-1} 2\gamma_k^2 \left(C^2_{g,i_k} + \eta_k^2 C^2_{f,i_k} \right) \\ &- 2\mathsf{p}_{i_k}^{-1} \gamma_k \left(x^{(i_k)}_{k} - x^{*(i_k)}_{\eta_k} \right)^T \left(\tilde{\nabla}_{i_k} g\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} f\left(x_k\right) \right).\end{aligned}$$

Now taking conditional expectation on both sides, and taking into account $\mathcal{D}\left(x_k, x^*_{\eta_k}\right)$ is $\mathcal{F}_k$ measurable,

$$\begin{aligned}\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right) | \mathcal{F}_k\right] \leq \mathcal{D}\left(x_k, x^*_{\eta_k}\right) &+ 2\gamma_k^2 \underbrace{\mathsf{E}\left[\mathsf{p}_{i_k}^{-1} C^2_{g,i_k} | \mathcal{F}_k\right]}_{\text{term-3}} + 2\gamma_k^2 \eta_k^2 \underbrace{\mathsf{E}\left[\mathsf{p}_{i_k}^{-1} C^2_{f,i_k} | \mathcal{F}_k\right]}_{\text{term-4}} - \\ &\underbrace{2\gamma_k \mathsf{E}\left[\mathsf{p}_{i_k}^{-1} \left(x^{(i_k)}_{k} - x^{*(i_k)}_{\eta_k} \right)^T \left(\tilde{\nabla}_{i_k} g\left(x_k\right) + \eta_k \tilde{\nabla}_{i_k} f\left(x_k\right) \right) | \mathcal{F}_k \right]}_{\text{term-5}}.\end{aligned}$$

$\text{term-3} = C_g^2, \quad \text{term-4} = C_f^2, \quad \text{term-5} = \sum_{i=1}^{d} \mathsf{p}_i \left(\mathsf{p}_i^{-1} \left(x^{(i)}_{k} - x^{*(i)}_{\eta_k} \right)^T \left(\tilde{\nabla}_i g\left(x_k\right) + \eta_k \tilde{\nabla}_i f\left(x_k\right) \right) \right)$
$= \left(x_k - x^*_{\eta_k} \right)^T \left(\tilde{\nabla} g\left(x_k\right) + \eta_k \tilde{\nabla} f\left(x_k\right) \right).$

Substituting the values of term-3, term-4 and term-5, we obtain,

$$\begin{aligned}\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] =& \mathcal{D}\left(x_k, x^*_{\eta_k}\right) + 2\gamma_k^2 C_g^2 + 2\gamma_k^2\eta_k^2 C_f^2 \\ & - 2\,\gamma_k \left(x_k - x^*_{\eta_k}\right)^T \left(\tilde{\nabla} g\left(x_k\right) + \eta_k \tilde{\nabla} f\left(x_k\right)\right). \end{aligned} \tag{4.4.3}$$

Now from Remark 4.3.2, for $x_1 = x_k$ and $x_2 = x^*_{\eta_k}$, we have,

$$\begin{aligned}&\left(\tilde{\nabla} g\left(x_k\right) + \eta_k \tilde{\nabla} f\left(x_k\right)\right)^T \left(x_k - x^*_{\eta_k}\right) - \left(\tilde{\nabla} g\left(x^*_{\eta_k}\right) + \eta_k \tilde{\nabla} f\left(x^*_{\eta_k}\right)\right)^T \left(x_k - x^*_{\eta_k}\right) \\ &\geq \eta_k \mu \left\| x_k - x^*_{\eta_k}\right\|^2. \end{aligned} \tag{4.4.4}$$

From the optimality conditions on (P_η), we have,

$$\left(\tilde{\nabla} g\left(x^*_{\eta_k}\right) + \eta_k \tilde{\nabla} f\left(x^*_{\eta_k}\right)\right)^T \left(x_k - x^*_{\eta_k}\right) \geq 0.$$

Thus,

$$\left(\tilde{\nabla} g\left(x_k\right) + \eta_k \tilde{\nabla} f\left(x_k\right)\right)^T \left(x_k - x^*_{\eta_k}\right) \geq \eta_k \mu \left\| x_k - x^*_{\eta_k}\right\|^2$$

Now, from (4.4.3) and the preceding inequality, we can write,

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq \mathcal{D}\left(x_k, x^*_{\eta_k}\right) - 2\gamma_k\eta_k\mu \underbrace{\left\| x_k - x^*_{\eta_k}\right\|^2}_{\text{term-6}} + 2\gamma_k^2 C_g^2 + 2\gamma_k^2\eta_k^2 C_f^2.$$

From Corollary 4.3.1, bounding term-6, we have,

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq (1 - 2\gamma_k\eta_k\mu\mathsf{p}_{min})\,\mathcal{D}\left(x_k, x^*_{\eta_k}\right) + 2\gamma_k^2 C_g^2 + 2\gamma_k^2\eta_k^2 C_f^2. \tag{4.4.5}$$

Now consider $\left\| x_k - x^*_{\eta_k}\right\|^2$. It can be written as,

$$\left\| x_k - x^*_{\eta_k}\right\|^2 = \left\| x_k - x^*_{\eta_{k-1}}\right\|^2 + \left\| x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\|^2 + \underbrace{2(x_k - x^*_{\eta_{k-1}})^T(x^*_{\eta_{k-1}} - x^*_{\eta_k})}_{\text{term-7}}. \tag{4.4.6}$$

$$\text{For } \; c \in \mathbb{R}^n, \; \text{term-7} \leq \left(c\left\| x_k - x^*_{\eta_{k-1}}\right\|\right)^2 + \left(\frac{\left\| x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\|}{c}\right)^2.$$

Substituting above in equation (4.4.6), with $c = \sqrt{\mathsf{p}_{min}\mu\gamma_k\eta_k}$,

$$\left\| x_k - x^*_{\eta_k}\right\|^2 \leq (1 + \mathsf{p}_{min}\mu\gamma_k\eta_k)\left\| x_k - x^*_{\eta_{k-1}}\right\|^2 + \left(1 + \frac{1}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\left\| x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\|^2.$$

From Lemma 4.3.3, and Corollary 4.3.1, we obtain,

$$\mathsf{p}_{min}\,\mathcal{D}\left(x_k, x^*_{\eta_k}\right) \leq (1 + \mathsf{p}_{min}\mu\gamma_k\eta_k)\,\mathsf{p}_{max}\,\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) + \left(1 + \frac{1}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\frac{C_g^2}{\mu^2}\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2.$$

Dividing both sides of previous inequality by p_{min}, and substituting this in (4.4.5), we have,

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq 12\gamma_k^2 C_g^2 + 2\gamma_k^2\eta_k^2 C_f^2 \frac{\mathsf{p}_{max}}{\mathsf{p}_{min}} \underbrace{\left(1 - 2\gamma_k\eta_k\mu\mathsf{p}_{min}\right)\left(1 + \mathsf{p}_{min}\mu\gamma_k\eta_k\right)}_{\text{term-9}} \mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)$$
$$- 2\gamma_k\eta_k\mu\mathsf{p}_{min}\frac{C_f^2}{\mu^2\,\mathsf{p}_{min}}\left(1 + \frac{1}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 + \frac{C_f^2}{\mu^2\,\mathsf{p}_{min}}\left(1 + \frac{1}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2.$$

We have, term-9 $\leq 1 - \mu\gamma_k\eta_k\mathsf{p}_{min}$, now we can write,

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq 2\gamma_k^2 C_g^2 + \frac{\mathsf{p}_{max}}{\mathsf{p}_{min}}(1 - \mu\gamma_k\eta_k\mathsf{p}_{min})\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) + 2\gamma_k^2\eta_k^2 C_f^2$$
$$+ \underbrace{\frac{\left(1 - 2\gamma_k\eta_k\mu\mathsf{p}_{min}\right)C_f^2}{\mu^2\,\mathsf{p}_{min}}\left(1 + \frac{1}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2}_{\text{term-10}}.$$

We have $\gamma_0\eta_0 < \frac{d}{\mathsf{p}_{min}\mu}$, Bounding term-10, we have,

$$\mathsf{E}\left[\mathcal{D}\left(x_{k+1}, x^*_{\eta_k}\right)|\mathcal{F}_k\right] \leq \left(1 - \mu\gamma_k\eta_k\mathsf{p}_{min}\right)\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right) + \frac{C_f^2}{\mathsf{p}_{min}\mu^2}\left(\frac{2}{\mathsf{p}_{min}\mu\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2$$
$$+ 2\gamma_k^2 C_g^2 + 2\gamma_k^2\eta_k^2 C_f^2.$$

Bounding non-increasing sequence, η_k we get the result. ■

Remark 4.4.1 Throughout the analysis, we assume that blocks are randomly selected using a uniform distribution.

Assumption 4.4.1 *Let the following hold:*

(a) $\{\gamma_k\}$ *and* $\{\eta_k\}$ *are positive sequences for* $k \geq 0$ *converging to zero such that* $\gamma_0\eta_0 < \frac{d}{\mu}$*;*

(b) $\sum_{k=0}^{\infty}\gamma_k\eta_k = \infty$; *(c)* $\sum_{k=0}^{\infty}\left(\frac{1}{\gamma_k\eta_k}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 < \infty$;

(d) $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$; *(e)* $\lim_{k\to\infty}\left(\frac{1}{\gamma_k^2\eta_k^2}\right)\left(\frac{\eta_{k-1}}{\eta_k} - 1\right)^2 = 0$;

(f) $\lim_{k\to\infty}\frac{\gamma_k}{\eta_k} = 0$.

Next, we show the a.s. convergence of the sequence $\{x_k\}$.

Theorem 4.4.1 (a.s. convergence of $\{x_k\}$) *Consider* (1.1.5) *and* (P_η)*. Let Assumption 4.4.1 hold. Consider the sequence* $\{x_k\}$ *is obtained by Algorithm 1, and the sequence* $\{x^*_{\eta_k}\}$ *suppose obtained by solving* (P_η)*. Then,* $\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)$ *goes to zero a.s. and* $\lim_{k\to\infty}\mathsf{E}\left[\mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)\right] = 0.$

Proof. We apply Lemma 4.3.1 to the result of Lemma 4.4.1. $v_k \triangleq \mathcal{D}\left(x_k, x^*_{\eta_{k-1}}\right)$, $\alpha_k \triangleq \frac{\mu\gamma_k\eta_k}{d}$, $\beta_k \triangleq \left(\frac{2d^2}{\mu\gamma_k\eta_k}\right)\frac{C_f^2}{\mu^2}\left(\frac{\eta_{k-1}}{\eta_k}-1\right)^2 + 2\gamma_k^2(C_g^2+\eta_0^2C_f^2)$. Now, in order to claim the convergence of v_k, we show that all conditions of Lemma 4.3.1 hold. Note that $\mathsf{p}_i = 1/d$. From Assumption 4.4.1 (a), definition of $\{\gamma_k\}$, $\{\eta_k\}$, and from $\gamma_0\eta_0 < \frac{d}{\mu}$, the first condition of Lemma 4.3.1 is satisfied. Now consider sequence β_k. From Assumption 4.4.1 (a), sequences $\{\gamma_k\}$, $\{\eta_k\}$ and the constant μ are positive, so the second condition of Lemma 4.3.1 is satisfied. Now in $\sum_{k=0}^{\infty}\alpha_k$, i.e. $\sum_{k=0}^{\infty}\frac{\mu\gamma_k\eta_k}{d}$. From Assumption 4.4.1(b), the third condition of Lemma 4.3.1 holds. Now from the definition of β_k and from Assumption 4.4.1(c) and (d), the fourth condition of Lemma 4.3.1 holds. Finally consider $\lim_{k\to\infty}\left(\frac{\beta_k}{\alpha_k}\right)=0$. Using the definition of β_k and Assumption 4.4.1(e, f), condition 5 of Lemma 4.3.1 holds. Thus we get the required result. ■

Next in Lemma 4.4.2 we give the choice of sequences γ_k and η_k that satisfy Assumption 4.4.1.

Lemma 4.4.2 *Let Assumption 4.3.2 hold. Then sequences $\{\gamma_k\}$ and $\{\eta_k\}$ given by $\gamma_k = \gamma_0(k+1)^{-a}$ and $\eta_k = \eta_0(k+1)^{-b}$ where a, and b satisfy, $a>0, \quad b>0, \quad a+b<1, \quad b<a, \quad a>0.5$, where $\gamma_0>0$ and $\eta_0>0$. Then $\{\gamma_k\}$ and $\{\eta_k\}$ satisfy Assumption 4.4.1.*

Proof. Similar to the proof of Lemma 5 in [96]. Omitted because of the space requirements. ■

Next, we show the a.s. convergence of the sequence $\{\bar{x}_k\}$.

Theorem 4.4.2 (a.s. convergence of $\{\bar{x}_k\}$) *Consider problem* (1.1.5). *Let γ_k and η_k be the sequences defined by Lemma 4.4.2 where $\gamma_0>0$, $\eta_0>0$, and $ar<1$. Then $\{\bar{x}_k\}$ converges to the unique solution of* (1.1.5), *x_f^* a.s.*

Proof. From $\lambda_{t,k} = \gamma_t^r / \sum_{j=0}^{k}\gamma_j^r$,

$$\left\|\bar{x}_k - x_f^*\right\| = \left\|\sum_{t=0}^{k}\lambda_{t,k}x_t - \sum_{t=0}^{k}\lambda_{t,k}x_f^*\right\| = \left\|\sum_{t=0}^{k}\lambda_{t,k}\left(x_t - x_f^*\right)\right\|.$$

Using the triangle inequality,

$$\left\|\bar{x}_k - x_f^*\right\| \le \sum_{t=0}^{k}\lambda_{t,k}\left\|x_t - x_f^*\right\|.$$

From definition of $\lambda_{t,k}$,

$$\left\|\bar{x}_k - x_f^*\right\| \le \frac{\sum_{t=0}^{k}\gamma_t^r\left\|x_t - x_f^*\right\|}{\sum_{j=0}^{k}\gamma_j^r}.$$

Comparing with Lemma 4.3.2,

$$\alpha_k \triangleq \gamma_k^r,\ u_k \triangleq \left\|x_k - x_f^*\right\|,\ v_{k+1} \triangleq \sum_{t=0}^{k}\gamma_t^r\left\|x_t - x_f^*\right\|.$$

Consider $\sum_{k=0}^{\infty}\alpha_t$, i.e. $\sum_{k=0}^{\infty}(1+k)^{-at}$. Now, we have $at<1$, so $\sum_{k=0}^{\infty}(1+k)^{-at}=\infty$. From Theorem 4.4.1, $\left\|x_k-x_f^*\right\|\to 0$ a.s. Therefore, Using Lemma 4.3.2, we get the required result. ■

Now we derive the rate of convergence for Algorithm 5.

Lemma 4.4.3 (Feasibility error bound for Algorithm 1) *Consider problem* (1.1.5) *and* $\{\bar{x}_k\}$*, the sequence generated by Algorithm 1. Let Assumption 4.3.1 hold,* $r(<1)$ *be an arbitrary scalar, and* γ_k *be a non-increasing sequence. Let* η_k *be a non-increasing sequence and* X *to be bounded, i.e.* $\|x\|\leq M$ *for all* $x\in X$ *for some* $M>0$*. Then for any* $z\in X$*, the following holds,*

$$\mathsf{E}[f(\bar{x}_N)]-f(z)\leq\left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2M_g\sum_{k=0}^{N-1}\gamma_k^r\eta_k+\right.$$
$$\left.2\mathsf{p}_{max}^{-1}M^2\left(\gamma_0^{r-1}+\gamma_{N-1}^{r-1}\right)+\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right)\left(C_f^2+C_g^2\eta_0^2\right)\right),$$

where $M_f(>0)$ *is a scalar such that* $f(x)\leq M_f$ *for all* $x\in X$*.*

Proof. Consider equation (4.3.1) in step 6 of Algorithm 5. Note that using induction, it can be shown that $\bar{x}_k=\sum_{i=0}^{k}\lambda_{t,k}x_t$, where $\lambda_{t,k}\triangleq\gamma_t^r/\sum_{j=0}^{k}\gamma_j^r$.

Next, consider $\{x_k\}$ be the sequence generated from Algorithm 1 and $z\in X$. Then from Definition 4.3.1, we have,

$$\mathcal{D}(x_{k+1},z)=\sum_{i=1,i\neq i_k}^{d}\mathsf{p}_i^{-1}\left\|x_k^{(i)}-z^{(i)}\right\|^2+\mathsf{p}_{i_k}^{-1}\underbrace{\left\|x_k^{(i_k)}-z^{(i_k)}\right\|^2}_{\text{term-11}}.$$

Consider term-11. From Algorithm 5, substituting $x_{k+1}^{(i_k)}$ and using the non-expansiveness property of the projection operator,

$$\left\|x_{k+1}^{(i_k)}-z^{(i_k)}\right\|^2\leq\left\|x_k^{(i_k)}-z^{(i_k)}\right\|^2+\gamma_k^2\left\|\tilde{\nabla}_{i_k}g(x_k)+\eta_k\tilde{\nabla}_{i_k}f(x_k)\right\|^2$$
$$-2\gamma_k\left(x_k^{(i_k)}-z^{(i_k)}\right)^T\left(\tilde{\nabla}_{i_k}g(x_k)+\eta_k\tilde{\nabla}_{i_k}f(x_k)\right).$$

Substituting the bound on term-1, we obtain,

$$\mathcal{D}(x_{k+1},z)=\mathcal{D}(x_k,z)+\mathsf{p}_{i_k}^{-1}\gamma_k^2\underbrace{\left\|\tilde{\nabla}_{i_k}g(x_k)+\eta_k\tilde{\nabla}_{i_k}f(x_k)\right\|^2}_{\text{term-12}}$$
$$-\mathsf{p}_{i_k}^{-1}2\gamma_k\left(x_k^{(i_k)}-z^{(i_k)}\right)^T\left(\tilde{\nabla}_{i_k}g(x_k)+\eta_k\tilde{\nabla}_{i_k}f(x_k)\right),$$

here we used Definition 4.3.1. From Remark 4.3.4, bounding term-12,

$$\left\|\tilde{\nabla}_{i_k}g(x_k)+\eta_k\tilde{\nabla}_{i_k}f(x_k)\right\|^2\leq 2C_{g,i_k}^2+2\eta_k^2C_{f,i_k}^2.$$

Substituting the bound of term-12, we get,

$$\mathcal{D}(x_{k+1},z) \leq \mathcal{D}(x_k,z) + 2\mathsf{p}_{i_k}^{-1}\gamma_k^2 C_{g,i_k}^2 + 2\mathsf{p}_{i_k}^{-1}\eta_k^2\gamma_k^2 C_{f,i_k}^2 - 2\mathsf{p}_{i_k}^{-1}\gamma_k\left(x_k^{(i_k)} - z^{(i_k)}\right)^T\left(\tilde{\nabla}_{i_k}g(x_k) + \eta_k\tilde{\nabla}_{i_k}f(x_k)\right).$$

By taking conditional expectation on the both sides of equation above, and taking into account $\mathcal{D}(x_k,z)$ is $\mathcal{F}_k$ measurable, we have:

$$\mathsf{E}[\mathcal{D}(x_{k+1},z)\,|\mathcal{F}_k] \leq \mathcal{D}(x_k,z) + 2\gamma_k^2\underbrace{\mathsf{E}\left[\mathsf{p}_{i_k}^{-1}C_{g,i_k}^2|\mathcal{F}_k\right]}_{\text{term-13}} + 2\eta_k^2\gamma_k^2\underbrace{\mathsf{E}\left[\mathsf{p}_{i_k}^{-1}C_{f,i_k}^2|\mathcal{F}_k\right]}_{\text{term-14}} - 2\gamma_k\underbrace{\mathsf{E}\left[\mathsf{p}_{i_k}^{-1}\left(x_k^{(i_k)} - z^{(i_k)}\right)^T\left(\tilde{\nabla}_{i_k}g(x_k) + \eta_k\tilde{\nabla}_{i_k}f(x_k)\right)|\mathcal{F}_k\right]}_{\text{term-15}}. \quad (4.4.7)$$

Using definition of expectation, term-13 $= C_g^2$, term-14 $= C_f^2$,
term-15 $= (x_k - z)^T\left(\tilde{\nabla}g(x_k) + \eta_k\tilde{\nabla}f(x_k)\right)$. From (4.4.7),

$$\mathsf{E}[\mathcal{D}(x_{k+1},z)\,|\mathcal{F}_k] \leq \mathcal{D}(x_k,z) + 2\gamma_k^2C_g^2 + 2\eta_k^2\gamma_k^2C_f^2 + 2\gamma_k\underbrace{(z-x_k)^T\left(\tilde{\nabla}g(x_k) + \eta_k\tilde{\nabla}f(x_k)\right)}_{\text{term-16}}. \quad (4.4.8)$$

Using the definition of subgradient at point x_k,

$$\begin{aligned}\text{term-16} &= (z-x_k)^T\tilde{\nabla}g(x_k) + \eta_k(z-x_k)^T\tilde{\nabla}f(x_k)\\ &\leq g(z) - g(x_k) + \eta_k f(z) - \eta_k f(x_k).\end{aligned}$$

Bounding term-16, using conditional and total expectation,

$$\mathsf{E}[\mathcal{D}(x_{k+1},z)] \leq \mathsf{E}[\mathcal{D}(x_k,z)] + 2\gamma_k^2C_g^2 + 2\eta_k^2\gamma_k^2C_f^2 + 2\gamma_k\left(g(z) + \eta_k f(z) - \mathsf{E}[g(x_k) + \eta_k f(x_k)]\right). \quad (4.4.9)$$

Multiplying the both sides of equation (4.4.9) by γ_k^{r-1}, and adding, subtracting $\gamma_{k-1}^{r-1}\mathsf{E}[\mathcal{D}(x_k,z)]$ on the left-hand side,

$$\gamma_k^{r-1}\mathsf{E}[\mathcal{D}(x_{k+1},z)] - \left(\gamma_k^{r-1} - \gamma_{k-1}^{r-1}\right)\underbrace{\mathsf{E}[\mathcal{D}(x_k,z)]}_{\text{term-17}} - \gamma_{k-1}^{r-1}\mathsf{E}[\mathcal{D}(x_k,z)] \leq 2\gamma_k^{r+1}C_g^2 + 2\gamma_k^{r+1}\eta_k^2C_f^2 + 2\gamma_k^r\left(g(z) + \eta_k f(z) - \mathsf{E}[g(x_k) + \eta_k f(x_k)]\right). \quad (4.4.10)$$

Since $r < 1$ and γ_k is a non-increasing, $\gamma_{k-1}^{r-1} - \gamma_k^{r-1}$ is a non-negative sequence. From Lemma 4.3.1, $\mathcal{D}(x_k,z) \leq \mathsf{p}_{max}\|x_k - z\|^2 \leq 2\mathsf{p}_{max}(\|x_k\|^2 + \|z\|^2)$. From the boundedness of set X, $\mathsf{E}[\mathcal{D}(x_k,k)] \leq 4\mathsf{p}_{max}M^2$. Substituting bound on term-17 in (4.4.10) and summing up over $k = 1,\ldots,N-1$,

$$-\gamma_0^{r-1}\mathsf{E}[\mathcal{D}(x_1,z)] - 4\gamma_{N-1}^{r-1}\mathsf{p}_{max}M^2 \leq 2C_g^2\sum_{k=1}^{N-1}\gamma_k^{r+1} + 2C_f^2\sum_{k=1}^{N-1}\gamma_k^{r+1}\eta_k^2 + 2\sum_{k=1}^{N-1}\gamma_k^r\left(g(z) + \eta_k f(z)\right) - 2\sum_{k=1}^{N-1}\gamma_k^r\mathsf{E}[g(x_k) + \eta_k f(x_k)]. \quad (4.4.11)$$

putting $k = 0$ in (4.4.9),

$$\mathsf{E}[\mathcal{D}(x_1, z)] \leq \underbrace{\mathsf{E}[\mathcal{D}(x_0, z)]}_{\text{term-18}} + 2\gamma_0^2 C_g^2 + 2\eta_0^2\gamma_0^2 C_f^2 + 2\gamma_0 \left(g(z) + \eta_0 f(z) - \mathsf{E}[g(x_0) + \eta_0 f(x_0)]\right).$$

Now, term-18 $\leq 4\mathsf{p}_{max}M^2 + 2\gamma_0^2 C_g^2 + 2\eta_0^2\gamma_0^2 C_f^2 + 2\gamma_0 \left(g(z) + \eta_0 f(z) - \mathsf{E}[g(x_0) + \eta_0 f(x_0)]\right)$. Multiplying the both sides of equation with γ_0^{r-1}, we get,

$$\begin{aligned}\gamma_0^{r-1}\mathsf{E}[\mathcal{D}(x_1, z)] - 4\gamma_0^{r-1}\mathsf{p}_{max}M^2 \leq & 2\gamma_0^{r+1}C_g^2 + 2\eta_0^2\gamma_0^{r+1} \\ & C_f^2 + 2\gamma_0^r \left(g(z) + \eta_0 f(z) - \mathsf{E}[g(x_0) + \eta_0 f(x_0)]\right). \quad (4.4.12)\end{aligned}$$

Adding (4.4.11) and (4.4.12) together, and combining the terms,

$$\begin{aligned}-4\mathsf{p}_{max}M^2\left(\gamma_0^{r-1} + \gamma_{N-1}^{r-1}\right) \leq & 2C_g^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right) + 2C_f^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\eta_k^2\right) \\ & + 2\left(\sum_{k=0}^{N-1}\gamma_k^r\left(g(z) + \eta_k f(z)\right)\right) - 2\left(\sum_{k=0}^{N-1}\gamma_k^r\mathsf{E}[g(x_k) + \eta_k f(x_k)]\right).\end{aligned}$$

Dividing the both sides by $\sum_{i=0}^{N-1}\gamma_i^r$, and denoting $\frac{\gamma_k^r}{\sum_{i=0}^{N-1}\gamma_i^r} = \lambda_{k,N-1}$, we get,

$$\underbrace{\sum_{k=0}^{N-1}\lambda_{k,N-1}\mathsf{E}[g(x_k) + \eta_k f(x_k)]}_{\text{term-19}} - \sum_{k=0}^{N-1}\lambda_{k,N-1}\left(g(z) + \eta_k f(z)\right) \leq$$

$$\left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2\mathsf{p}_{max}M^2\left(\gamma_0^{r-1} + \gamma_{N-1}^{r-1}\right) + C_g^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right) + C_f^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\eta_k^2\right)\right).$$

By updating term-19 and rearranging the original terms,

$$\begin{aligned}\underbrace{\mathsf{E}\left[\sum_{k=0}^{N-1}\lambda_{k,N-1}g(x_k)\right]}_{\text{term-20}} - \underbrace{\sum_{k=0}^{N-1}\lambda_{k,N-1}g(z)}_{\text{term-21}} \leq & \sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f(z) - \mathsf{E}\left[\sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f(x_k)\right] \\ & + \left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2\mathsf{p}_{max}M^2\left(\gamma_0^{r-1} + \gamma_{N-1}^{r-1}\right) + C_g^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right) + C_f^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\eta_k^2\right)\right).\end{aligned}$$

Using the convexity of g and the definition of $\lambda_{k,N-1}$, we have term-20 $\leq \sum_{k=0}^{N-1}\lambda_{k,N-1}f(x_k)$, and term-21 $= g(z)$.

$$\begin{aligned}\mathsf{E}[g(\bar{x}_N)] - g(z) \leq & \underbrace{\sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f(z) - \mathsf{E}\left[\sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f(x_k)\right]}_{\text{term-22}} \\ & + \left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2\mathsf{p}_{max}M^2\left(\gamma_0^{r-1} + \gamma_{N-1}^{r-1}\right) + C_g^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\right) + C_f^2\left(\sum_{k=0}^{N-1}\gamma_k^{r+1}\eta_k^2\right)\right).\end{aligned}$$

Using definition of M_f, we obtain,

$$\text{term-22} = \mathsf{E}\left[\sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f(z) - \sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k f\left(x_k\right)\right] \leq \mathsf{E}\left[\sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k |f(z) - f\left(x_k\right)|\right]$$
$$\leq 2M_f \sum_{k=0}^{N-1}\lambda_{k,N-1}\eta_k.$$

Bounding term-22 and using the definition of $\lambda_{k,N-1}$,

$$\mathsf{E}[g\left(\bar{x}_N\right)] - g(z) \leq$$
$$\left(\sum_{i=0}^{N-1}\gamma_i^r\right)^{-1}\left(2M_f\sum_{k=0}^{N-1}\gamma_k^r\eta_k + C_g^2\sum_{k=0}^{N-1}\gamma_k^{r+1} + 2\mathsf{p}_{max}M^2\left(\gamma_0^{r-1}+\gamma_{N-1}^{r-1}\right) + C_f^2\sum_{k=0}^{N-1}\gamma_k^{r+1}\eta_k^2\right)$$

Here, since η_k is a non-increasing sequence, bounding it by η_0, we get the required result. ■

Next, we state Lemma 4.4.4 (see Lemma 9, pg. 418 of [96]) and use it in Theorem 4.4.3 to derive the rate of convergence.

Lemma 4.4.4 *For a scalar $\alpha \neq -1$ and integers l, N, where $0 \leq l \leq N-1$, we have*

$$\frac{N^{\alpha+1}-(l+1)^{\alpha+1}}{\alpha+1} \leq \sum_{k=l}^{N-1}(k+1)^{\alpha} \leq (l+1)^{\alpha} + \frac{(N+1)^{\alpha+1}-(l+1)^{\alpha+1}}{\alpha+1}.$$

In Theorem 4.4.3, we show the rate of convergence for the sequence generated from Algorithm 5.

Theorem 4.4.3 *Consider problem* (1.1.5) *and the sequence generated from Algorithm 1 $\{\bar{x}_N\}$. Let Assumptions 4.3.1, and 4.3.2 with a uniform distribution. Let the sequence $\{\gamma_k\}$ and $\{\eta_k\}$ are given by the following, $\gamma_k = \gamma_0/(k+1)^{0.5+0.1\delta}$ and $\eta_k = \eta_0/(k+1)^{0.5-\delta}$, such that $\gamma_0 \triangleq \gamma\sqrt{d}$, for some $\gamma > 0$, $\eta_0 > 0$, $\gamma\eta_0 < \frac{\sqrt{d}}{\mu}$, $0 < \delta < 0.5$, and $r < 1$. Then, (i) Sequence $\{\bar{x}_N\}$ converges to x_f^* almost surely. (ii) $\mathsf{E}[f\left(\bar{x}_N\right)] \to f^*$ with the rate $\mathcal{O}\left(\sqrt{d}/N^{0.5-\delta}\right)$.*

Proof. (i) Consider the sequences given for γ_k and η_k. By denoting $a = 0.5+0.1\delta$ and $b = 0.5-\delta$, we have, $\gamma_k = \gamma_0/(k+1)^a$ and $\eta_k = \eta_0/(k+1)^b$. Also we know that $0 < \delta < 0.5$ and $r < 1$. Therefore, we have: $a, b > 0$, $b < a$, $0.5 < a < 0.55$, $0 < b < 0.5$, $a+b < 1$ and $ar < 1$. So, γ_k and η_k satisfy all the conditions of Lemma 4.4.2.
(ii) Substituting γ_k, η_k, and $z := x_f^*$ in Lemma 4.4.3, we obtain,

$$\mathsf{E}[g\left(\bar{x}_N\right)] - g^* \leq \left(\gamma_0^r\sum_{i=0}^{N-1}\frac{1}{(k+1)^{ar}}\right)^{-1}\left(2\mathsf{p}_{max}^{-1}M^2\gamma_0^{r-1}\left(N^{a(1-r)}+1\right)\right.$$
$$\left.+\gamma_0^r\left(2M_f\eta_0\sum_{k=0}^{N-1}\frac{1}{(k+1)^{ar+b}} + \left(C_g^2+C_f^2\eta_0^2\right)\gamma_0\underbrace{\sum_{k=0}^{N-1}\frac{1}{(k+1)^{ar+a}}}_{\text{term-23}}\right)\right).$$

modifying term-23 in equation above, and expanding terms,

$$\mathsf{E}[g(\bar{x}_N)] - g^* \leq 2\mathsf{p}_{max}^{-1} M^2 \gamma_0^{-1} \left(\underbrace{\left(\sum_{i=0}^{N-1} \frac{1}{(k+1)^{ar}} \right)^{-1} N^{a(1-r)}}_{\text{term-24}} + \underbrace{\left(\sum_{i=0}^{N-1} \frac{1}{(k+1)^{ar}} \right)^{-1}}_{\text{term-25}} \right)$$
$$+ \left(2M_f \eta_0 + \gamma_0 \left(C_g^2 + C_f^2 \eta_0^2 \right) \right).$$

The above equation can also be written as

$$\mathsf{E}[g(\bar{x}_N)] - g^* \leq 2dM^2 \gamma^{-1} d^{-0.5} \left(\text{term-24} + \text{term-25}\right) + \text{term-26} \left(2M_f \eta_0 + \gamma \sqrt{d} \left(C_g^2 + C_f^2 \eta_0^2 \right) \right).$$

From Lemma 4.4.4, we have,

$$\text{term-25} \leq \frac{1-ar}{N^{-ar+1}-1} = \mathcal{O}\left(N^{-(1-ar)}\right), \quad \text{term-24} \leq \frac{(1-ar)N^{a(1-r)}}{N^{-ar+1}-1} = \mathcal{O}\left(N^{-(1-a)}\right),$$
$$\text{term-26} \leq \left(\frac{1-ar}{N^{-ar+1}-1} \right) \left(1 + \frac{(N+1)^{1-(ar+b)} - 1}{1-(ar+b)} \right) = \mathcal{O}\left(N^{-(1-ar)}\right) + \mathcal{O}\left(N^{-b}\right).$$

Now, substituting bounds of terms-24, 25, and 26, we have,

$$\mathsf{E}[g(\bar{x}_N)] - g^* \leq \mathcal{O}\left(\sqrt{d} \mathsf{max} \left\{ N^{-(1-ar)}, N^{-(1-a)}, N^{-b} \right\} \right) = \mathcal{O}\left(\sqrt{d} N^{-\mathsf{min}\{1-ar,\ 1-a,\ b\}} \right).$$

From definitions of $a, r,$ and δ, we obtain the result. ■

4.5 Numerical Results

In the literature, one of the ways to address the ill-posedness in image deblurring is employing the regularization technique. The ill-posed problem (1.1.3) is converted into the regularized problem (P_η) by, for example, substituting functions $g(x) := \|Ax - b\|_2^2$, and $f(x) := \|x\|_2^2 + \|x\|_1$.

As the value of regularization parameter η changes, a different optimization problem (P_η) is solved. The basic idea is, $\eta \in (0, +\infty)$ governs the way by which solutions of linear inverse problem (1.1.3) are approximated by (P_η). This is a *two-loop* scheme, as explained earlier in the introduction. It is computationally inefficient to find a suitable regularization parameter η. In this section, we address this challenge in an image deblurring using problem formulation (1.1.4) (specific case of problem (1.1.5)), by avoiding the conventional *two-loop* regularization technique. The values of regularization parameter η_k and stepsize parameter γ_k are updated iteratively, explained below in the inference.

We are provided with the blurred noisy image in Figure 1(a), which is further converted into the column vector b. Our objective is to get the original image, Figure 1 (b) using image deblurring. Here we compare two ways of deblurring: standard regularization, and using Algorithm 5.

(a) η=0 (b) η=0.001 (c) η=0.01 (d) η=0.1 (e) η=1

(f) $k = 10^2$ (g) $k = 10^3$ (h) $k = 10^4$ (i) $k = 10^5$ (j) $k = 10^6$

Figure 6: First set of images (a)-(e) are obtained using the sequential regularization at different values of η, running for 10^5 iterations. Second set of images (f)-(j) are obtained using Algorithm 5 by stopping at different iterations k.

Inference. Figure 6(a)–(e) show the deblurred images obtained by conventional regularization at different η for 10^5 iterations. Figure 6(f)–(j) show the deblurred images using Algorithm 5 scheme with stopping at different iteration. Our Algorithm 5 is computationally efficient, because unlike the case of *two-loop* regularization, in Algorithm 5 we implement the scheme once. A question then is: what iteration k we should stop the scheme at? Stopping at a suitable iteration k is desired because that governs the deblurred image quality. In particular, this *single-loop* scheme seems to be promising because one can generate images after every fixed number of iterations and stop the implementation once the generated deblurred image is good enough. Note that, image deblurring is used for a toy example here, to demonstrate the performance of Algorithm 5. This application can be extended for deblurring of images with a higher resolution. In this work, our intention is to demonstrate the performance of Algorithm 5 on the well-known example of cameraman.

4.6 Concluding Remarks

We address ill-posed optimization problem with a high-dimensional solution space and nondifferentiable objective function. A randomized block-coordinate iterative regularized subgradient Algorithm 5 is developed to address problem (1.1.5). We establish the convergence of the sequence generated from Algorithm 5 to the unique solution of (1.1.5) in an almost sure sense. Furthermore, we derive a rate of convergence $\mathcal{O}\left(\frac{\sqrt{d}}{k^{0.5-\delta}}\right)$, with respect to the inner level objective of the bilevel problem (1.1.5). Our ground assumptions in the convergence proof and rate analysis are mild, such that f and g can be nondifferentiable functions. Demonstration of Algorithm 5 on an image deblurring example shows that the proposed *single-loop* scheme computationally performs well compared to the conventional *two-loop* regularization schemes.

CHAPTER V

LARGE-SCALE DISTRIBUTED NONLINEARLY CONSTRAINED OPTIMIZATION

In this chapter, motivated by applications arising from sensor networks and machine learning, we consider the problem of minimizing a finite sum of nondifferentiable convex functions where each component function is associated with an agent and a hard-to-project constraint set. We consider a special case of Problem (P_2). Section 5.1 provides the problem formulation. Among well-known avenues to address finite sum problems is the class of incremental gradient (IG) methods where a single component function is selected at each iteration in a cyclic or randomized manner. When the problem is constrained, the existing IG schemes (including projected IG, proximal IAG, and SAGA) require a projection step onto the feasible set at each iteration. Consequently, the performance of these schemes is afflicted with costly projections when the problem includes: (1) nonlinear constraints, or (2) a large number of linear constraints. Our focus in this chapter lies in addressing both of these challenges. Section 5.2 provides the available methods to address problem (1.2.3) and the research gap. We develop an algorithm called averaged iteratively regularized incremental gradient (aIR-IG) that does not involve any hard-to-project computation. Section 5.3 includes the algorithm outline. Under mild assumptions, we derive non-asymptotic rates of convergence for both suboptimality and infeasibility metrics. Section 5.4 includes the convergence and rate analysis. Numerically, we show that the proposed scheme outperforms the standard projected IG methods on distributed soft-margin support vector machine problems in Section 5.5. Section 5.6 includes the concluding remarks.

5.1 Problem Formulation

We consider a finite sum minimization subject to nonlinear inequality and linear equality functional constraints as follows.

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & f(x) \triangleq \sum_{i=1}^{m} f_i(x) && && (1.2.3)\\ \text{subject to} \quad & h_i(x) \leq 0 && \text{for all } i \in \{1,\dots,m\},\\ & A_i x = b_i && \text{for all } i \in \{1,\dots,m\},\\ & x^{(j)} \geq 0 && \text{for all } j \in J,\\ & x \in X, \end{aligned}$$

The content of this chapter has been published in the proceedings of 2021 American Control Conference [51].

where the component functions $f_i : \mathbb{R}^n \to \mathbb{R}$ and $h_i : \mathbb{R}^n \to \mathbb{R}$ are nonsmooth convex, $A_i \in \mathbb{R}^{d_i \times n}$, and $b_i \in \mathbb{R}^{d_i}$, for all $i \in \{1, \ldots, m\}$. Also, $X \subseteq \mathbb{R}^n$ is an easy-to-project convex set and $J \subseteq \{1, \ldots, n\}$. The information about f_i, h_i, A_i, and b_i is locally known by agent i, while the sets X and J are globally known. Parameters n, m, and $p \triangleq \sum_{i=1}^m d_i$ are possibly large. Note that this is a special case of problem (P_2).

5.2 Existing Methods and Research Gap

Table 3: Comparison and memory requirements of incremental gradient schemes for addressing constrained finite sum problems.

Ref.	Scheme	Problem class	Formulation	Convergence rate	Memory
[62]	Projected IG	C_0^0	$\min_{x\in X} \sum_{i=1}^m f_i(x)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}(n)$
[17,34]	IAG	$C_{\mu,L}^{1,1}$	$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m f_i(x)$	linear	$\mathcal{O}(mn)$
[29]	SAGA	$C_{0,L}^{1,1}$, $C_{\mu,L}^{1,1}$	$\min_{x\in X} \sum_{i=1}^m f_i(x)$	$\mathcal{O}\left(\frac{1}{k}\right)$, linear	$\mathcal{O}(mn)$
[87]	Proximal IAG	$C_{\mu,L}^{1,1}$	$\min_{x\in X} \sum_{i=1}^m f_i(x)$	linear	$\mathcal{O}(mn)$
[35]	IG	$C_{0,L}^{2,1}, C_{\mu,L}^{2,1}$	$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m f_i(x)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$, $\mathcal{O}\left(\frac{1}{k}\right)$	$\mathcal{O}(n)$
This work	aIR-IG	C_0^0	$\min_{x\in X} \sum_{i=1}^m f_i(x)$ $h_i(x) \leq 0 \ \ \forall i \in [m]$ $A_i x = b_i \ \ \forall i \in [m]$ $x^{(j)} \geq 0 \ \ \forall j \in J$	suboptimality: $\mathcal{O}\left(k^{-0.5+b}\right)$ infeasibility: $\mathcal{O}\left(k^{-b}\right)$ for $0 < b < 0.5$	$\mathcal{O}(n)$

Problem (1.2.3) arises in a breadth of applications including expected loss minimization in statistical learning [74] where f_i is associated with a data block, as well as distributed optimization in wireless sensor networks where f_i represents the local performance measure of the i^{th} agent [70]. One of the popular methods in addressing finite sum problems, in particular, in the unconstrained regime, is the class of incremental gradient (IG) methods where utilizing the additive structure of the problem, the algorithm cycles through the data blocks and updates the local estimates of the optimal solution in a sequential manner [14]. While the first variants of IG schemes find their roots in addressing neural networks as early as in the '80s [15], the complexity analysis of these schemes has been a trending research topic in the fields of control and machine learning in the past two decades. In addressing constrained problems with easy-to-project constraint sets, the projected incremental gradient (P-IG) method and its subgradient variant were developed [63]. In the smooth case, it is described as follows: given an initial point $x_{0,1} \in X$, where $X \subseteq \mathbb{R}^n$ denotes the constraint set, for each $k \geq 1$, consider the following update rule:

$$\begin{aligned} x_{k,i+1} &:= \mathcal{P}_X\left(x_{k,i} - \gamma_k \nabla f_i\left(x_{k,i}\right)\right) \quad \text{for all } i = 1, \ldots, m, \\ x_{k+1,1} &:= x_{k,m+1} \qquad \text{for all } k \geq 0, \end{aligned}$$

where $\mathcal{P}$ denotes the Euclidean projection operator and is defined as $\mathcal{P}_X(z) \triangleq \operatorname{argmin}_{x\in X}\|x - z\|_2$ and $\gamma_k > 0$ is the stepsize parameter. Recently, under the assumption of strong convexity

and twice continuous differentiability of the objective function, the standard IG method was proved to converge with the rate $\mathcal{O}(1/k)$ in the unconstrained case [35]. This is an improvement to the previously known rate of $\mathcal{O}(1/\sqrt{k})$ for the merely convex case. Accelerated variants of IG schemes with provable convergence speeds were also developed, including the incremental aggregated gradient method (IAG) [17, 34], SAG [74], and SAGA [29]. While addressing the merely convex case, SAGA using averaging achieves a sublinear convergence rate, assuming strong convexity and smoothness, this is improved for non-averaging variants of SAGA and IAG to a linear rate.

Existing gap. Despite the faster rates of convergence in comparison with the standard IG method, the aforementioned methods require an excessive memory of $\mathcal{O}(mn)$ which limits their applications in the large-scale settings. Another existing challenge in the implementation of these schemes lies in addressing the hard-to-project constraints. Contending with the presence of constraints, projected (and more generally proximal) variants of the aforementioned IG schemes have been developed. However, the performance of these schemes is afflicted with costly projections when the problem includes: (1) nonlinear constraints, or (2) a large number of linear constraints. In the area of distributed optimization over networks, addressing constraints has been done to a limited extent through employing duality theory, projection, or penalty methods (see [6, 21, 36, 65, 79]). We also note that a celebrated variant of the dual based schemes is the alternating direction method of multipliers (ADMM) (e.g., see [7, 53, 60, 83, 84]). Despite the recent advancements in this area, most ADMM methods cannot address inequality constraints with a separable structure as in (1.2.3). Also, ADMM schemes often work under the premise that the communication graph is undirected. Indeed, despite the wide-spread application of the theory of duality and Lagrangian relaxation in addressing constrained problems in centralized regimes, there have been a limited work in the area of distributed optimization that can cope with hard-to-project constraints (see [6, 13, 36] and the references therein). Nevertheless, the problem formulation (1.2.3) is not addressed in the aforementioned articles. Recently, primal-dual algorithms are proposed for finite sum convex optimization problems with conic constraints [6, 37]. A recent work [42] introduced primal-dual incremental gradient method for nonsmooth convex optimization problems. Moreover, iterative regularization (IR) has been employed as a new constraint-relaxation strategy in regimes where addressing the constraints are challenging (e.g., see [3, 52, 96, 98]). Our work in this chapter has been motivated by the recent success of the IR approach. To this end, our goal lies in employing the IR approach to develop an IG algorithm that can address formulation (1.2.3) without requiring any hard-to-project computation.

5.3 Algorithm Outline

In this section, we first provide the main assumptions on problem (1.2.3) and present the outline of the algorithm. Then, we present a few preliminary results that will be used in the analysis.

Assumption 5.3.1 (Properties of problem (1.2.3)**)** *Let the following hold:*
(a) Component function $f_i : \mathbb{R}^n \to \mathbb{R}$ is merely convex and subdifferentiable with bounded subgradients for all $i \in [m]$.

(b) Function $h_i : \mathbb{R}^n \to \mathbb{R}$ is convex and subdifferentiable with bounded subgradients for all $i \in [m]$.
(c) The set X is compact and convex.
(d) The feasible set of problem (1.2.3) *is nonempty.*

An underlying idea in development of Algorithm 6 is to define a regularized error metric.

Definition 5.3.1 Consider the following term for measuring infeasibility for an agent i

$$\phi_i(x) \triangleq \frac{1}{2}\|A_i x - b_i\|^2 + h_i^+(x) + \sum_{j\in J} \frac{\max\left\{-x^{(j)}, 0\right\}}{m},$$

where $h_i^+(x) \triangleq \max\{0, h_i(x)\}$ for $i \in [m]$ and all $x \in \mathbb{R}^n$. Further, we define $\phi(x) = \sum_{i=1}^m \phi_i(x)$.

Then, for each agent i, we consider a regularized metric defined as $\phi_i(x) + \eta_k f_i(x)$ at iteration k. This metric captures both infeasibility and objective component function of the agent. Next, we derive a subgradient to this metric.

Let $\partial h_i^+(x)$ denote the subdifferential set of the function h_i^+ at x. Consider the vector $\tilde{\nabla} h_i^+(x)$ defined as $\tilde{\nabla} h_i^+(x) \triangleq h_i^+(x)\tilde{\nabla} h_i(x)$ where $\tilde{\nabla} h_i(x)$ denotes a subgradient of function h_i at x. Then, from the definition of subgradient mapping and the definition of $h_i^+(x)$, we have that $\tilde{\nabla} h_i^+(x) \in \partial h_i^+(x)$. Next, consider the function $\frac{1}{m}\sum_{j\in J} \max\left\{0, -x^{(j)}\right\}$. A subgradient to this function is the vector $\frac{\mathbb{1}^-(x)}{m}$ where $\mathbb{1}^-(x)$ is defined a column vector $\in \mathbb{R}^n$ and the value of any component $i \in \{1, \ldots, n\}$ is -1 when $x^{(i)} < 0$ and $i \in J$, otherwise that component is 0. Let $x_{k,i}$ in $\mathbb{R}^n$ denote the iterate of agent i at iteration k. From the above discussion, we can conclude that the subgradient of the regularized error metric for agent i, is given as follows

$$A_i^T\left(A_i x_{k,i} - b_i\right) + \tilde{\nabla} h_i^+(x_{k,i}) + \frac{\mathbb{1}^-\left(x_{k,i}\right)}{m} + \eta_k \tilde{\nabla} f_i\left(x_{k,i}\right).$$

We are now ready to present the outline of aIR-IG scheme presented by Algorithm 6. At each iteration, agents update their iterates in a cyclic manner by employing the aforementioned subgradient. Each agent uses its local information including subgradients of functions f_i, h_i, as well as matrix A_i and vector b_i. Here γ_k and η_k are the stepsize and regularization parameters, respectively. These parameters are updated at each iteration. This, indeed, is important because the convergence and rate analysis mainly depend on the choice of γ_k and η_k. The key research question lies in finding suitable update rules for the two sequences so that we can achieve convergence and rate results. For the rate analysis, we employ averaging which is characterized by stepsize γ_k and a scalar $0 \leq r < 1$.

In the following, we claim the boundedness of the subgradients $\tilde{\nabla}\phi_i(x)$ and $\tilde{\nabla} f_i(x)$ which will be used in the rate analysis in the next section.

Remark 5.3.1 Under Assumption 5.3.1, from compactness of the set X, the term $A_i^T\left(A_i x - b_i\right)$ is bounded. Also, from the boundedness of subgradients of function h_i and continuity of the function h_i that is implied from convexity of h_i, we can claim that the subgradient $\tilde{\nabla} h_i^+(x) \triangleq h_i^+(x)\tilde{\nabla} h_i(x)$ is bounded on the set X. Consequently, we have that

Algorithm 6 Averaged Iteratively Regularized Incremental Gradient (aIR-IG)

1: **Input**: $x_0 \in \mathbb{R}^n$, $\bar{x}_0 := x_0$, $S_0 := \gamma_0^r$, and $0 \leq r < 1$.
2: **for** $k = 0, 1, \ldots, N-1$ **do**
3: Let $x_{k,1} := x_k$ and select $\gamma_k > 0$, $\eta_k > 0$
4: **for** $i = 1, \ldots, m$ **do**
5:

$$x_{k,i+1} := \mathcal{P}_X \left(x_{k,i} - \gamma_k \left(A_i^T \left(A_i x_{k,i} - b_i \right) + \tilde{\nabla} h_i^+(x_{k,i}) + \frac{\mathbb{1}^- \left(x_{k,i} \right)}{m} + \eta_k \tilde{\nabla} f_i \left(x_{k,i} \right) \right) \right)$$

6: **end for**
7: Set $x_{k+1} \triangleq x_{k,m+1}$.
8: Update the weighted average iterate as

$$\bar{x}_{k+1} := \frac{S_k \bar{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}}, \text{ where } S_{k+1} := S_k + \gamma_{k+1}^r.$$

9: **end for**
10: **return**: $\bar{x}_N$.

$\tilde{\nabla}\phi_i(x) \triangleq A_i^T \left(A_i x - b_i \right) + \tilde{\nabla} h_i^+(x) + \frac{\mathbb{1}^-(x)}{m}$ is a bounded subgradient of ϕ_i for all $x \in X$. This implies that there exists a scalar $C > 0$ such that for all $x \in X$, we have

$$\sum_{i=1}^m \tilde{\nabla}\phi_i(x) \leq C \quad \text{and} \quad \tilde{\nabla}\phi_i(x) \leq \frac{C}{m} \text{ for all } i \in [m].$$

Remark 5.3.2 From Assumption 5.3.1, taking into account the subdifferentiability and boundedness of subgradient of function f_i, there exists a scalar $C_f > 0$ such that for all $x \in X$,

$$\sum_{i=1}^m \left\| \tilde{\nabla} f_i \left(x \right) \right\| \leq C_f \text{ and } \left\| \tilde{\nabla} f_i \left(x \right) \right\| \leq \frac{C_f}{m} \text{ for all } i \in [m].$$

Remark 5.3.3 Taking into account Assumption 5.3.1, from Theorem 3.61 in [9], functions f_i and ϕ_i are Lipschitz continuous over set X. Therefore for $x, y \in X$, and $i \in [m]$, $|f_i(x) - f_i(y)| \leq \frac{C_f}{m} \|x - y\|$ and $|\phi_i(x) - \phi_i(y)| \leq \frac{C}{m} \|x - y\|$.

Next, we show that the sequence $\bar{x}_k$, employed in Algorithm 6, is a well-defined weighted average.

Remark 5.3.4 From Algorithm 6, the average of the iterate can be written as $\bar{x}_{k+1} = \sum_{t=0}^k \lambda_{t,k} x_t$, where $\lambda_{t,k} \triangleq \frac{\gamma_t^r}{\sum_{j=0}^k \gamma_j^r}$ for $t \in \{0, \ldots, k\}$ denote the weights. This can be shown using induction on $k \geq 0$. For $k = 0$, the relation holds directly due to the initialization $\bar{x}_0 := x_0$. To show the relation for $k+1$, assuming that it holds for k, using the step 7 in

Algorithm 6, and that $S_k := \sum_{j=0}^{k} \gamma_j^r$, we have

$$\bar{x}_{k+1} = \frac{S_k \bar{x}_k + \gamma_{k+1}^r x_{k+1}}{S_{k+1}} = \frac{\sum_{t=0}^{k} \gamma_t^r x_t + \gamma_{k+1}^r x_{k+1}}{S_{k+1}} = \frac{\sum_{t=0}^{k+1} \gamma_t^r x_t}{\sum_{t=0}^{k+1} \gamma_t^r} = \sum_{t=0}^{k+1} \lambda_{t,k} x_t.$$

In this work, the average of the m^{th} agent's iterate is considered in the analysis.

The next result will be employed in the rate analysis.

Lemma 5.3.1 (Lemma 2.14 in [52]**)** *For any scalar $\alpha \in [0, 1)$ and integer N such that $N \geq 2^{\frac{1}{1-\alpha}} - 1$, we have*

$$\frac{(N+1)^{1-\alpha}}{2(1-\alpha)} \leq \sum_{k=0}^{N} (k+1)^{-\alpha} \leq \frac{(N+1)^{1-\alpha}}{1-\alpha}.$$

5.4 Convergence Analysis

We begin with obtaining an error bound that will be employed later in the construction of bounds on the objective value and infeasibility metrics for Algorithm 6.

Lemma 5.4.1 *Let the sequence $\{x_k\}$ be generated by Algorithm 6 and $\{\gamma_k\}$ and $\{\eta_k\}$ be nonincreasing positive sequences. Let Assumption 5.3.1 hold, $0 \leq r < 1$, and scalars $C, C_f > 0$ be defined as in Remarks 5.3.1 and 5.3.2, respectively. Then, for any $y \in X$ and $k \geq 0$, we have*

$$2\gamma_k^r \eta_k \left(f(x_k) - f(y)\right) + 2\gamma_k^r \left(\phi\left(x_k\right) - \phi\left(y\right)\right) \leq \gamma_k^{r-1} \left\|x_k - y\right\|^2 - \gamma_k^{r-1} \left\|x_{k+1} - y\right\|^2 + \left(1 + \frac{1}{m}\right) \gamma_k^{r+1} \left(C + \eta_k C_f\right)^2. \tag{5.4.1}$$

Proof. Consider the update rule in step 4 in Algorithm 6. For iteration $k \geq 0$, agent $i \in \{1, \ldots, m\}$, and $y \in X$, we have

$$\left\|x_{k,i+1} - y\right\|^2 := \Big\|\mathcal{P}_X\Big(x_{k,i} - \gamma_k \Big(A_i^T \left(A_i x_{k,i} - b_i\right) + \tilde{\nabla} h_i^+(x_{k,i}) + \frac{\mathbb{1}^-\left(x_{k,i}\right)}{m} + \eta_k \tilde{\nabla} f_i\left(x_{k,i}\right)\Big)\Big) - \mathcal{P}_X(y)\Big\|^2.$$

Employing the non-expansiveness of the projection operator, and recalling Definition 5.3.1 for $\phi_i(x)$, we have

$$\begin{aligned}
\left\|x_{k,i+1} - y\right\|^2 &\leq \left\|x_{k,i} - \gamma_k \left(\tilde{\nabla}\phi_i\left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i\left(x_{k,i}\right)\right) - y\right\|^2 \\
&= \left\|x_{k,i} - y\right\|^2 + \underbrace{\gamma_k^2 \left\|\tilde{\nabla}\phi_i\left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i\left(x_{k,i}\right)\right\|^2}_{\text{term 1}} \\
&\quad - 2\gamma_k \left(\tilde{\nabla}\phi_i\left(x_{k,i}\right) + \eta_k \tilde{\nabla} f_i\left(x_{k,i}\right)\right)^T \left(x_{k,i} - y\right).
\end{aligned}$$

Consider term 1. Employing the triangle inequality, taking into account the definitions of scalars C, and C_f, we obtain

$$\|x_{k,i+1}-y\|^2 \leq \|x_{k,i}-y\|^2+\gamma_k^2\left(\frac{C+\eta_k C_f}{m}\right)^2 \underbrace{-2\gamma_k\left(\tilde{\nabla}\phi_i\left(x_{k,i}\right)+\eta_k\tilde{\nabla}f_i\left(x_{k,i}\right)\right)^T\left(x_{k,i}-y\right)}_{\text{term 2}}.$$

Bounding term 2 by invoking the definition of subgradient and the convexity of $\phi_i(x)$ and $f_i(x)$, we obtain

$$\begin{aligned}\|x_{k,i+1}-y\|^2 \leq \|x_{k,i}-y\|^2+\gamma_k^2\left(\frac{C+\eta_k C_f}{m}\right)^2 \\ +2\gamma_k\eta_k\left(f_i(y)-f_i(x_{k,i})\right)+2\gamma_k\left(\phi_i(y)-\phi_i\left(x_{k,i}\right)\right).\end{aligned}$$

Taking summation over all the agents $i\in\{1,\ldots,m\}$,

$$\begin{aligned}\|x_{k+1}-y\|^2 \leq \|x_k-y\|^2+2\gamma_k\eta_k\sum_{i=1}^m\left(f_i(y)-f_i(x_{k,i})\right) \\ +\gamma_k^2\sum_{i=1}^m\left(\frac{C+\eta_k C_f}{m}\right)^2+2\gamma_k\sum_{i=1}^m\left(\phi_i(y)-\phi_i\left(x_{k,i}\right)\right).\end{aligned}$$

Adding and subtracting $2\gamma_k\sum_{i=1}^m\phi_i\left(x_k\right)+2\gamma_k\eta_k\sum_{i=1}^m f_i\left(x_k\right)$, and taking into account Definition 5.3.1, we have

$$\begin{aligned}\|x_{k+1}-y\|^2 &\leq \|x_k-y\|^2+\frac{\gamma_k^2\left(C+\eta_k C_f\right)^2}{m}+2\gamma_k\eta_k\left(f(y)-f(x_k)\right)+2\gamma_k\left(\phi(y)-\phi\left(x_k\right)\right) \\ &\quad+2\gamma_k\sum_{i=1}^m\left(\phi_i\left(x_k\right)-\phi_i\left(x_{k,i}\right)+\eta_k\left(f_i\left(x_k\right)-f_i\left(x_{k,i}\right)\right)\right), \\ &\leq \|x_k-y\|^2+\frac{\gamma_k^2\left(C+\eta_k C_f\right)^2}{m}+2\gamma_k\eta_k\left(f(y)-f(x_k)\right)+2\gamma_k\left(\phi(y)-\phi\left(x_k\right)\right) \\ &\quad+2\gamma_k\sum_{i=1}^m\left(\underbrace{\left|\phi_i\left(x_k\right)-\phi_i\left(x_{k,i}\right)\right|}_{\text{term 3}}+\eta_k\underbrace{\left|f_i\left(x_k\right)-f_i\left(x_{k,i}\right)\right|}_{\text{term 4}}\right).\end{aligned}$$

From Remark 5.3.3, bounding terms 3 and 4, we have

$$\begin{aligned}\|x_{k+1}-y\|^2 \leq \|x_k-y\|^2+\frac{\gamma_k^2\left(C+\eta_k C_f\right)^2}{m}+2\gamma_k\eta_k\left(f(y)-f(x_k)\right)+2\gamma_k\left(\phi(y)-\phi\left(x_k\right)\right) \\ +\frac{2\gamma_k\left(C+\eta_k C_f\right)}{m}\sum_{i=2}^m\underbrace{\left\|x_k-x_{k,i}\right\|}_{\text{term 5}}.\end{aligned} \tag{5.4.2}$$

Note that from Algorithm 6, for $i=1$, we have $\|x_k-x_{k,1}\|=0$. Consider term 5 in relation (5.4.2). Applying induction on i, we bound term 5 as $\|x_k-x_{k,i}\|\leq(i)\gamma_k\left(C+\eta_k C_f\right)/m$ for

any $i = 2, \ldots, m$. For $i = 2$, from Algorithm 6, we have

$$\begin{aligned}\|x_k - x_{k,2}\| &= \left\|\mathcal{P}_X\left(x_{k,1}\right) - \mathcal{P}_X\left(x_{k,1} - \gamma_k\left(\tilde{\nabla}\phi_1(x_{k,1}) + \eta_k\tilde{\nabla}f_1\left(x_{k,1}\right)\right)\right)\right\| \\ &\leq \gamma_k\left\|\tilde{\nabla}\phi_1(x_{k,1}) + \eta_k\tilde{\nabla}f_1\left(x_{k,1}\right)\right\| \leq \gamma_k\left(C + \eta_k C_f\right)/m.\end{aligned}$$

Now, suppose the hypothesis statement holds for some $i \geq 2$. Then, we can write

$$\begin{aligned}\|x_k - x_{k,i+1}\| &= \left\|\mathcal{P}_X\left(x_k\right) - \mathcal{P}_X\left(x_{k,i} - \gamma_k\left(\tilde{\nabla}\phi_i(x_{k,i}) + \eta_k\tilde{\nabla}f_i\left(x_{k,i}\right)\right)\right)\right\| \\ &\leq \|x_k - x_{k,i}\| + \gamma_k\left\|\tilde{\nabla}\phi_i(x_{k,i}) + \eta_k\tilde{\nabla}f_i\left(x_{k,i}\right)\right\| \\ &\leq \|x_k - x_{k,i}\| + \frac{\gamma_k\left(C + \eta_k C_f\right)}{m} \leq \frac{(i+1)\gamma_k\left(C + \eta_k C_f\right)}{m}.\end{aligned}$$

Therefore, the hypothesis statement holds for any $i \geq 2$. Substituting the bound for term 5 in equation (5.4.2), we have

$$\begin{aligned}\|x_{k+1} - y\|^2 \leq{}& \|x_k - y\|^2 + \frac{\gamma_k^2\left(C + \eta_k C_f\right)^2}{m} + 2\gamma_k\eta_k\left(f(y) - f(x_k)\right) + 2\gamma_k\left(\phi(y) - \phi\left(x_k\right)\right) \\ &+ \frac{2\gamma_k\left(C + \eta_k C_f\right)}{m}\sum_{i=2}^{m}\frac{(i)\gamma_k\left(C + \eta_k C_f\right)}{m} \\ ={}& \|x_k - y\|^2 + \left(1 + \frac{1}{m}\right)\gamma_k^2\left(C + \eta_k C_f\right)^2 + 2\gamma_k\eta_k\left(f(y) - f(x_k)\right) \\ &+ 2\gamma_k\left(\phi(y) - \phi\left(x_k\right)\right).\end{aligned}$$

Multiplying both sides by the positive term γ_k^{r-1}, we obtain the desired result. ■

Next we construct the error bounds for Algorithm 6 in terms of the sequences $\{\gamma_k\}$ and $\{\eta_k\}$.

Proposition 5.4.1 (Error bounds for Algorithm 6) *Consider problem* (1.2.3). *Let* $\bar{x}_N$ *be generated by Algorithm 6 after* N *iterations and* $\{\gamma_k\}$ *and* $\{\eta_k\}$ *be nonincreasing and strictly positive sequences. Further, let Assumption 5.3.1 hold, scalars* $C_f, C > 0$, *and parameter* $0 \leq r < 1$. *Let scalars* $M, M_f > 0$ *be defined such that we have:* $\|x\| \leq M$ *and* $|f(x)| \leq M_f$ *for all* $x \in X$. *Then for any optimal solution* x^* *to* (1.2.3), *we have the following:*

$$(a)\ f(\bar{x}_N) - f(x^*) \leq \left(\sum_{k=0}^{N}\gamma_k^r\right)^{-1}\left(\frac{2M^2\gamma_N^{r-1}}{\eta_N} + \left(1 + \frac{1}{m}\right)\frac{\left(C + \eta_0 C_f\right)^2}{2}\sum_{k=0}^{N}\frac{\gamma_k^{r+1}}{\eta_k}\right).$$

$$(b)\ \phi\left(\bar{x}_N\right) \leq \left(\sum_{k=0}^{N}\gamma_k^r\right)^{-1}\left(2M^2\gamma_N^{r-1} + 2M_f\sum_{k=0}^{N}\gamma_k^r\eta_k + \left(1 + \frac{1}{m}\right)\frac{\left(C + \eta_0 C_f\right)^2}{2}\sum_{k=0}^{N}\gamma_k^{r+1}\right).$$

Proof. Consider relation (5.4.1) from Lemma 5.4.1, for any $y \in X$. Substituting y by x^* and taking into account the feasibility of the vector x^* to problem (1.2.3), we obtain

$$\begin{aligned}2\gamma_k^r\eta_k\left(f(x_k) - f(x^*)\right) + 2\gamma_k^r\phi\left(x_k\right) \leq{}& \gamma_k^{r-1}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \\ &+ \left(1 + \frac{1}{m}\right)\gamma_k^{r+1}\left(C + \eta_k C_f\right)^2.\end{aligned}$$

Taking into account the nonnegativity of $2\gamma_k^r \phi(x_k)$ and dividing both sides by $2\eta_k$, we have

$$\gamma_k^r \left(f(x_k) - f(x^*)\right) \leq \frac{\gamma_k^{r-1}}{2\eta_k} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) + \left(1 + \frac{1}{m}\right) \frac{\gamma_k^{r+1} \left(C + \eta_k C_f\right)^2}{2\eta_k}. \tag{5.4.3}$$

Adding and subtracting $\frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}} \|x_k - x^*\|^2$ in the above,

$$\begin{aligned}\gamma_k^r \left(f(x_k) - f(x^*)\right) \leq \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}} \|x_k - x^*\|^2 \quad - \frac{\gamma_k^{r-1}}{2\eta_k} \|x_{k+1} - x^*\|^2 + \underbrace{\left(\frac{\gamma_k^{r-1}}{2\eta_k} - \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\right)}_{\text{term 6}} \|x_k - x^*\|^2 \\ + \underbrace{\left(1 + \frac{1}{m}\right) \frac{\gamma_k^{r+1} \left(C + \eta_k C_f\right)^2}{2\eta_k}}_{\text{term 7}}.\end{aligned} \tag{5.4.4}$$

Recalling the definition for scalar M, we have:

$$\|x_k - x^*\|^2 \leq 2\|x_k\|^2 + 2\|x^*\|^2 \leq 4M^2. \tag{5.4.5}$$

Taking into account $r < 1$ and the nonincreasing property of the sequences $\{\gamma_k\}$ and $\{\eta_k\}$, we have term $6 \geq 0$. Bounding term 7 in equation (5.4.4), we have

$$\begin{aligned}\gamma_k^r \left(f(x_k) - f(x^*)\right) \leq & \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}} \|x_k - x^*\|^2 \quad - \frac{\gamma_k^{r-1}}{2\eta_k} \|x_{k+1} - x^*\|^2 + \left(\frac{\gamma_k^{r-1}}{2\eta_k} - \frac{\gamma_{k-1}^{r-1}}{2\eta_{k-1}}\right) 4M^2 \\ & + \left(1 + \frac{1}{m}\right) \frac{\left(C + \eta_0 C_f\right)^2 \gamma_k^{r+1}}{2\eta_k}.\end{aligned}$$

Next, taking summations over $k = 1, \ldots, N$, we obtain

$$\begin{aligned}\sum_{k=1}^{N} \gamma_k^r \left(f(x_k) - f(x^*)\right) \leq & \frac{\gamma_0^{r-1}}{2\eta_0} \|x_1 - x^*\|^2 - \frac{\gamma_N^{r-1}}{2\eta_N} \|x_{N+1} - x^*\|^2 + \left(\frac{\gamma_N^{r-1}}{2\eta_N} - \frac{\gamma_0^{r-1}}{2\eta_0}\right) 4M^2 \\ & + \left(1 + \frac{1}{m}\right) \frac{\left(C + \eta_0 C_f\right)^2}{2} \sum_{k=1}^{N} \frac{\gamma_k^{r+1}}{\eta_k}.\end{aligned} \tag{5.4.6}$$

Rewriting equation (5.4.3) for $k = 0$, we have

$$\gamma_0^r \left(f(x_0) - f(x^*)\right) \leq \frac{\gamma_0^{r-1}}{2\eta_0} \left(\|x_0 - x^*\|^2 - \|x_1 - x^*\|^2\right) + \left(1 + \frac{1}{m}\right) \frac{\gamma_0^{r+1} \left(C^2 + \eta_0 C_f\right)^2}{2\eta_0}.$$

Adding the preceding relation with (5.4.6), we obtain

$$\begin{aligned}\sum_{k=0}^{N} \gamma_k^r \left(f(x_k) - f(x^*)\right) \leq & 2M^2 \left(\frac{\gamma_N^{r-1}}{\eta_N} - \frac{\gamma_0^{r-1}}{\eta_0}\right) - \frac{\gamma_N^{r-1}}{2\eta_N} \|x_{N+1} - x^*\|^2 + \frac{\gamma_0^{r-1} \|x_0 - x^*\|^2}{2\eta_0} \\ & + \left(1 + \frac{1}{m}\right) \frac{\left(C + \eta_0 C_f\right)^2}{2} \sum_{k=0}^{N} \frac{\gamma_k^{r+1}}{\eta_k}.\end{aligned}$$

Further from (5.4.5), and neglecting the nonpositive term,

$$\sum_{k=0}^{N}\gamma_k^r\left(f(x_k)-f(x^*)\right)\leq 2M^2\gamma_N^{r-1}/\eta_N+\left(1+\frac{1}{m}\right)\frac{\left(C+\eta_0C_f\right)^2}{2}\sum_{k=0}^{N}\frac{\gamma_k^{r+1}}{\eta_k}.$$

Next, dividing both sides by $\sum_{k=0}^{N}\gamma_k^r$,

$$\left(\sum_{k=0}^{N}\gamma_k^r\right)^{-1}\sum_{k=0}^{N}\gamma_k^r\left(f(x_k)-f(x^*)\right)\leq\left(\sum_{k=0}^{N}\gamma_k^r\right)^{-1}\left(2M^2\gamma_N^{r-1}/\eta_N+\left(1+\frac{1}{m}\right)\frac{\left(C+\eta_0C_f\right)^2}{2}\sum_{k=0}^{N}\frac{\gamma_k^{r+1}}{\eta_k}\right).$$

Taking into account the convexity of f and recalling Remark 5.3.4, we obtain the result.
(b) Consider equation (5.4.1). Writing it for $y:=x^*\in X$,

$$\begin{aligned}2\gamma_k^r\phi\left(x_k\right)\leq&2\gamma_k^r\eta_k\left(f\left(x^*\right)-f(x_k)\right)+\gamma_k^{r-1}\left(\left\|x_k-x^*\right\|^2-\left\|x_{k+1}-x^*\right\|^2\right)\\&+\left(1+\frac{1}{m}\right)\gamma_k^{r+1}\left(C+\eta_kC_f\right)^2.\end{aligned}$$

Recalling the definition of M_f, we have, $|f(x^*)-f(x_k)|\leq 2M_f$. Bounding the preceding inequality,

$$2\gamma_k^r\phi(x_k)\leq\gamma_k^{r-1}\left(\left\|x_k-x^*\right\|^2-\left\|x_{k+1}-x^*\right\|^2\right)+4\gamma_k^r\eta_kM_f+\left(1+\frac{1}{m}\right)\gamma_k^{r+1}\left(C+\eta_kC_f\right)^2. \tag{5.4.7}$$

Adding and subtracting $\gamma_{k-1}^{r-1}\|x_k-x^*\|^2$ in the above,

$$\begin{aligned}2\gamma_k^r\phi(x_k)\leq&\gamma_{k-1}^{r-1}\left\|x_k-x^*\right\|^2-\gamma_k^{r-1}\left\|x_{k+1}-x^*\right\|^2+4\gamma_k^r\eta_kM_f+\underbrace{\left(\gamma_k^{r-1}-\gamma_{k-1}^{r-1}\right)\left\|x_k-x^*\right\|^2}_{\text{term 8}}\\&+\underbrace{\left(1+\frac{1}{m}\right)\gamma_k^{r+1}\left(C+\eta_kC_f\right)^2}_{\text{term 9}}.\end{aligned}$$

Using the nonincreasing property of $\{\gamma_k\}$ and $\{\eta_k\}$, recalling $0\leq r<1$, we have $\gamma_k^{r-1}-\gamma_{k-1}^{r-1}>0$, and $\left(1+\frac{1}{m}\right)\gamma_k^{r+1}>0$. Further, from the boundedness of set X, we have term $8<\left(\gamma_k^{r-1}-\gamma_{k-1}^{r-1}\right)4M^2$, and term $9<\left(1+\frac{1}{m}\right)\gamma_k^{r+1}\left(C+\eta_0C_f\right)^2$. Next, taking summations over $k=1,\ldots,N$, and dropping the nonpositive terms, we obtain

$$\begin{aligned}2\sum_{k=1}^{N}\gamma_k^r\phi(x_k)\leq&\gamma_0^{r-1}\left\|x_1-x^*\right\|^2+4M^2\left(\gamma_N^{r-1}-\gamma_0^{r-1}\right)\\&+\left(1+\frac{1}{m}\right)\left(C+\eta_0C_f\right)^2\sum_{k=1}^{N}\gamma_k^{r+1}+4M_f\sum_{k=1}^{N}\gamma_k^r\eta_k.\end{aligned} \tag{5.4.8}$$

Writing equation (5.4.7) for $k = 0$, we have

$$2\gamma_0^r\phi(x_0) \leq \gamma_0^{r-1}\left(\|x_0 - x^*\|^2 - \|x_1 - x^*\|^2\right) + 4\gamma_0^r\eta_0 M_f + \left(1 + \frac{1}{m}\right)\gamma_0^{r+1}\left(C + \eta_0 C_f\right)^2.$$

Adding this into equation (5.4.8), we have

$$\begin{aligned} 2\sum_{k=0}^{N}\gamma_k^r\phi(x_k) \leq & \gamma_0^{r-1}\|x_0 - x^*\|^2 + 4M^2\left(\gamma_N^{r-1} - \gamma_0^{r-1}\right) \\ & + \left(1 + \frac{1}{m}\right)\left(C + \eta_0 C_f\right)^2\sum_{k=0}^{N}\gamma_k^{r+1} + 4M_f\sum_{k=0}^{N}\gamma_k^r\eta_k. \end{aligned}$$

Bounding $\|x_0 - x^*\|^2$ from equation (5.4.5), dividing both sides by $\sum_{k=0}^{N}\gamma_k^r$, taking into account the convexity of $\phi(x_k)$, and from Remark 5.3.4, we obtain the required result. ■

Next, we present the suboptimality and infeasibility convergence rate statements for the proposed algorithm.

Theorem 5.4.1 (Suboptimality and infeasibility rate results) *Consider Algorithm 6. Let Assumption 5.3.1 hold. Consider scalars $M, M_f > 0$ such that $\|x\| \leq M$ and $|f(x)| \leq M_f$ for all $x \in X$. Let $\bar{x}_N$ be generated by Algorithm 6 after N iterations. Let $\{\gamma_k\}$ and $\{\eta_k\}$ be the stepsize and regularization parameter sequences generated using $\gamma_k = \frac{\gamma_0}{\sqrt{1+k}}$, $\eta_k = \frac{\eta_0}{(1+k)^b}$, where $\gamma_0, \eta_0 > 0$, and $0 < b < 0.5$. Then, for any optimal solution x^* to problem (1.2.3), we have:*

$$(a)\ f\left(\bar{x}_N\right) - f(x^*) \leq \frac{2-r}{\gamma_0^r(N+1)^{0.5-b}}\left(\frac{2M^2}{\eta_0\gamma_0^{1-r}} + \frac{(m+1)\,\gamma_0^{1+r}\left(C + \eta_0 C_f\right)^2}{2m\eta_0(0.5 - 0.5r + b)}\right). \tag{5.4.9}$$

$$(b)\ \phi\left(\bar{x}_N\right) \leq \frac{(2-r)}{(N+1)^b}\left(\frac{2M^2}{\gamma_0} + \frac{2M_f\eta_0}{(1 - 0.5r - b)} + \frac{(m+1)\left(C + \eta_0 C_f\right)^2\gamma_0}{2m(0.5 - 0.5r)}\right). \tag{5.4.10}$$

Proof. Taking Proposition 5.4.1 (a) and (b) into account, let us define the following terms

$$\Lambda_{N,1} \triangleq \sum_{k=0}^{N}\gamma_k^r, \quad \Lambda_{N,2} \triangleq \frac{2M^2\gamma_N^{r-1}}{\eta_N}, \quad \Lambda_{N,3} \triangleq \left(1 + \frac{1}{m}\right)\frac{\left(C + \eta_0 C_f\right)^2}{2}\sum_{k=0}^{N}\eta_k^{-1}\gamma_k^{r+1},$$

$$\Lambda_{N,4} \triangleq 2M^2\gamma_N^{r-1}, \quad \Lambda_{N,5} \triangleq 2M_f\sum_{k=0}^{N}\eta_k\gamma_k^r, \quad \Lambda_{N,6} \triangleq \left(1 + \frac{1}{m}\right)\frac{\left(C + \eta_0 C_f\right)^2}{2}\sum_{k=0}^{N}\gamma_k^{r+1}.$$

From Proposition 5.4.1 (a) and (b), we have

$$f\left(\bar{x}_N\right) - f(x^*) \leq \left(\Lambda_{N,2} + \Lambda_{N,3}\right)/\Lambda_{N,1}, \quad \phi\left(\bar{x}_N\right) \leq \left(\Lambda_{N,4} + \Lambda_{N,5} + \Lambda_{N,6}\right)/\Lambda_{N,1}. \tag{5.4.11}$$

Next, applying Lemma 5.3.1 and substituting $\{\gamma_k\}$ and $\{\eta_k\}$ by their update rules, we obtain

$$\Lambda_{N,1} = \sum_{k=0}^{N} \frac{\gamma_0^r}{(k+1)^{0.5r}} \geq \frac{\gamma_0^r (N+1)^{1-0.5r}}{2(1-0.5r)}. \quad \Lambda_{N,2} = \frac{2M^2(N+1)^{0.5(1-r)+b}}{\eta_0 \gamma_0^{1-r}}.$$

$$\Lambda_{N,3} = \left(1+\frac{1}{m}\right) \frac{(C+\eta_0 C_f)^2}{2} \sum_{k=0}^{N} \frac{\gamma_0^{1+r}}{\eta_0 (k+1)^{0.5(1+r)-b}}$$

$$\Lambda_{N,3} \leq \frac{(m+1)\,\gamma_0^{1+r}\,(C+\eta_0 C_f)^2\,(N+1)^{1-0.5(1+r)+b}}{2m\eta_0(1-0.5(1+r)+b)}.$$

$$\Lambda_{N,4} = \frac{2M^2(N+1)^{0.5(1-r)}}{\gamma_0^{1-r}}. \quad \Lambda_{N,5} = \sum_{k=0}^{N} \frac{2M_f \eta_0 \gamma_0^r}{(k+1)^{0.5r+b}} \leq \frac{2M_f \eta_0 \gamma_0^r (N+1)^{1-0.5r-b}}{1-0.5r-b}.$$

$$\Lambda_{N,6} = \left(1+\frac{1}{m}\right) \frac{(C+\eta_0 C_f)^2}{2} \sum_{k=0}^{N} \frac{\gamma_0^{r+1}}{(k+1)^{0.5(1+r)}}$$

$$\leq \frac{(m+1)\,(C+\eta_0 C_f)^2\,\gamma_0^{r+1}(N+1)^{1-0.5(1+r)}}{2m(1-0.5(1+r))}.$$

For these inequalities to hold, we need to ensure that conditions of Lemma 5.3.1 are met. Accordingly, we must have $0 \leq 0.5r < 1$, $0 \leq 0.5(1+r) - b < 1$, $0 \leq 0.5r + b < 1$, and $0 \leq 0.5(1+r) < 1$. These relations hold because $0 \leq r < 1$ and $0 < b < 0.5$. Another set of conditions when applying Lemma 5.3.1 includes $N \geq \max\left\{2^{1/(1-0.5r)}, 2^{1/(1-0.5(1+r)+b)}, 2^{1/(1-0.5r-b)}, 2^{1/(1-0.5(1+r))}\right\} - 1$. Note that this condition is satisfied as a consequence of $N \geq 2^{\frac{2}{1-r}} - 1$, $b > 0$, and $0 \leq r < 1$. We conclude that all the necessary conditions for applying Lemma 5.3.1 and obtaining the aforementioned bounds for the terms $\Lambda_{N,i}$ are satisfied. To show that the inequalities (5.4.9) and (5.4.10), it suffices to substitute the preceding bounds of $\Lambda_{N,i}$, in the inequalities (5.4.11).

$$f(\bar{x}_N) - f(x^*) \leq \frac{2-r}{\gamma_0^r (N+1)^{1-0.5r}} \left(\frac{2M^2(N+1)^{0.5-0.5r+b}}{\eta_0 \gamma_0^{1-r}} + \frac{(m+1)\,\gamma_0^{1+r}\,(C+\eta_0 C_f)^2\,(N+1)^{0.5-0.5r+b}}{2m\eta_0(1-0.5(1+r)+b)} \right).$$

Inequality (5.4.9) is obtained by rearranging the terms in the preceding relation. Next, consider the following

$$\phi(\bar{x}_N) \leq \frac{2-r}{\gamma_0^r (N+1)^{1-0.5r}} \left(\frac{2M_f \eta_0 \gamma_0^r (N+1)^{1-0.5r-b}}{1-0.5r-b} + \frac{2M^2(N+1)^{0.5-0.5r}}{\gamma_0^{1-r}} + \frac{(m+1)\,(C+\eta_0 C_f)^2\,\gamma_0^{r+1}(N+1)^{0.5-0.5r}}{2m(1-0.5(1+r))} \right),$$

$$\leq (2-r) \left(\frac{2M^2}{\gamma_0 (N+1)^{0.5}} + \frac{2M_f \eta_0}{(1-0.5r-b)(N+1)^b} + \frac{(m+1)\,(C+\eta_0 C_f)^2\,\gamma_0}{2m(0.5-0.5r)(N+1)^{0.5}} \right).$$

Taking into account $0 < b < 0.5$, equation (5.4.10) is obtained by rearranging the terms in the preceding inequality. ■

Remark 5.4.1 The convergence rates derived in Theorem 5.4.1 can be improved under stronger assumptions such as smoothness and strong convexity of the functions f_i. Indeed, this is a future direction of our study. We note that a preliminary version of this chapter where a better rate has been derived under such assumptions is [52]. We have omitted such discussions due to the space limitation.

5.5 Numerical Results

Consider a distributed soft-margin support vector machine (SVM) problem described as follows. Let a dataset be given as $\mathcal{D} \triangleq \{(u_j, v_j) \mid u_j \in \mathbb{R}^n,\ v_j \in \{-1, +1\},\ \text{for } j \in S\}$ where u_j and v_j denote the attributes and the binary label of the j^{th} data point, respectively, and S denotes the index set. Let the data be distributed among m agents by defining $\mathcal{D}_i$ such that $\cup_{i=1}^m \mathcal{D}_i = \mathcal{D}$. Let S_i denote the index set corresponding to agent i such that $\sum_{i=1}^m |S_i| = |S|$. Then given $\lambda > 0$, the distributed SVM problem is given as

$$\begin{aligned} \underset{w,b,z}{\text{minimize}} \quad & \sum_{i=1}^m \left(\frac{1}{2m}\|w\|^2 + \frac{1}{\lambda} \sum_{j \in S_i} z_j \right) \\ \text{subject to} \quad & v_j(w^T u_j + b) \geq 1 - z_j, \ \text{for } j \in S_i \text{ and } i \in [m], \\ & z_j \geq 0, \qquad\qquad\qquad \text{for } j \in S_i \text{ and } i \in [m]. \end{aligned} \tag{5.5.1}$$

To solve problem (5.5.1) using Algorithm 4, we apply Lemma 1.2.1 by casting (5.5.1) as model (1.1.2). We also implement some of the well-known existing IG schemes, namely projected IG, proximal IAG, and SAGA. Unlike Algorithm 4, these schemes require a projection step onto the constraint set. To compute the projections we use the Gurobi-Python solver.

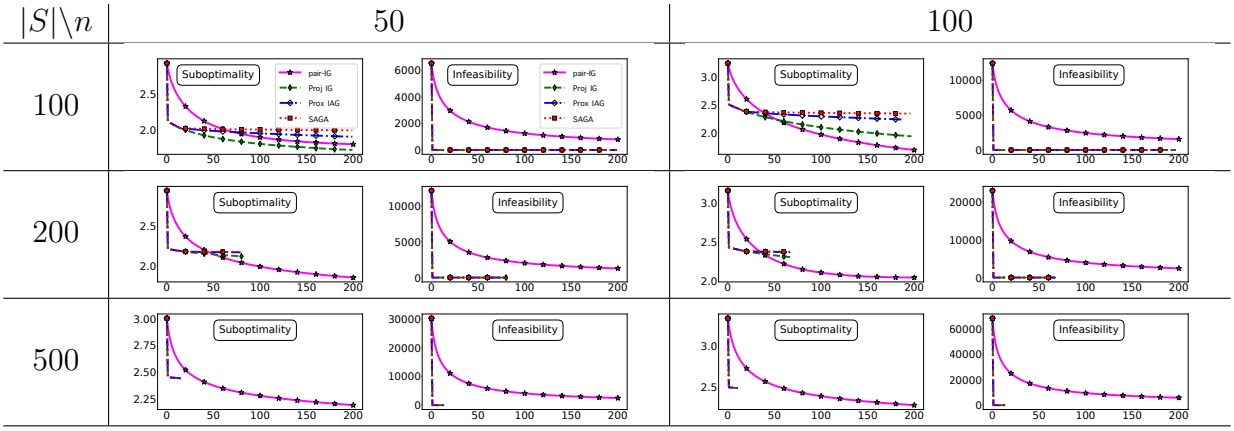

Figure 7: Comparison of Algorithm 4 with standard IG methods in solving an SVM model

Set-up. We consider 20 agents and assume that $\lambda := 10$. We let $\gamma_k := \frac{1}{\sqrt{k+1}}$ and $\eta_k := \frac{1}{(k+1)^{0.25}}$ in Algorithm 4. We use identical initial stepsizes in all the methods. We provide the comparisons with respect to the runtime and report the performance of each scheme over 200 seconds. We use a synthetic dataset with different values for n and $|S|$. The suboptimality is characterized in terms of the global objective and the infeasibility is measured by quantifying the violation of constraints of problem (5.5.1) aligned with ideas in Lemma 1.2.1.

Insights. In Figure 7, we observe that with an increase in the dimension of the solution space, i.e., n, or the size of the training dataset, i.e., $|S|$, the projection evaluations in the standard IG schemes take longer and consequently, the performance of the IG methods is deteriorated in large-scale settings. However, utilizing the reformulation in Lemma 1.2.1, the proposed method does not require any projection operations for addressing problem (5.5.1). As such, the performance of Algorithm 4 does not get affected severely with the increase in n or $|S|$. Note that in Figure 7, the reason that the IG schemes do not show any updates for $|S| = 200$ and $|S| = 500$ beyond a time threshold is because of the interruption in their last update when the method reaches the 200 seconds time limit.

5.6 Concluding remarks

We consider the problem of minimizing the finite sum with separable (agent-wise) nonlinear inequality and linear equality and inequality constraints. Our work is motivated by the computational challenges in the projected incremental gradient schemes under the presence of hard-to-project constraints. We develop an averaged iteratively regularized incremental gradient scheme where we employ a novel regularization-based relaxation technique. The proposed algorithm is designed in a way that it does not require a hard-to-project computation. We establish the rates of convergence for the objective function value and the infeasibility of the generated iterates. We compare the proposed scheme with the state-of-the-art incremental gradient schemes including projected IG, proximal IAG, and SAGA. We observe that the proposed scheme outperforms the projected schemes as the number of samples or the dimension of the solution space increases.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusion

In this dissertation, we consider a unifying class of optimization problems with variational inequality (VI) constraints. Traditionally, constrained optimization models include functional constraints in the form of inequalities and equations. VI constraints allow for capturing a wide range of optimization problems that cannot be formulated by the existing standard constrained models, especially when the constraint set is complicated by equilibrium constraints, complementarity constraints, or an inner-level large-scale optimization problem. The main motivation arises from the notion of efficiency of equilibria in multi-agent networks, e.g. communication networks and power systems.

In the first part of this dissertation, we consider a class of optimization problems with Cartesian variational inequality (CVI) constraints where the objective function is convex and the CVI is associated with monotone mapping and a convex Cartesian product set. In the literature, the iteration complexity of the existing solution methods for optimization problems with the CVI constraints is unknown. To address this shortcoming, we develop a first-order method called averaged randomized block iteratively regularized gradient (aRB-IRG) scheme. The main contributions include: (i) In the case where the associated set of the CVI is bounded and the objective function is nondifferentiable and convex, we derive new non-asymptotic suboptimality and infeasibility convergence rate statements in an ergodic sense. We also obtain deterministic variants of the convergence rates when we suppress the randomized block selection in aRB-IRG scheme. Importantly, this is the first work to provide these rate guarantees for this class of problems. (ii) In the case where the CVI set is unbounded and the objective function is smooth and strongly convex, utilizing the properties of the Tikhonov trajectory, we establish the global convergence of aRB-IRG in an almost sure and a mean sense. Further, we provide numerical experiments for computing the best Nash equilibrium in a networked Cournot competition model. In image deblurring applications where the VI constraints represent the first-order optimality conditions of a convex optimization problem, we devise a randomized block iteratively regularized subgradient scheme (RB-IRG). Under a uniform probability distribution in selecting the blocks and a careful choice of the stepsize and regularization parameter sequences, we establish an almost sure convergence of the generated sequence of the algorithm. Furthermore, we derive a non-asymptotic convergence rate in terms of the expected objective value of the inner level problem.

In the second part of this dissertation, we consider a class of constrained multi-agent optimization problems where the goal is to cooperatively minimize the sum of agent-specific

objective functions. Here, we consider a distributed framework where the objective function and the mappings of the VI are locally known by the agents. We develop an iteratively regularized incremental gradient (pair-IG) method where the agents communicate over a cycle graph at each iteration. We derive new non-asymptotic agent-wise convergence rates for the suboptimality, and infeasibility of the VI constraint. To analyze the convergence rate in the solution space, assuming the objective function is strongly convex and smooth, we derive non-asymptotic agent-wise rates on an error metric that relates the generated iterate with the Tikhonov trajectory. We provide preliminary numerical experiments for computing the best equilibrium in a transportation network problem. We also apply the proposed scheme to address a special case of this distributed formulation, where the VI constraint characterizes a feasible set. We compare the performance of the proposed scheme with that of the existing IG methods in addressing on distributed soft-margin support vector machine problem.

6.2 Future Directions

For addressing optimization problems with variational inequality constraints, in this dissertation, we have designed algorithms considering cyclic and star shaped communication among the agents. One of the ideas for extending the results in this dissertation could be to debate upto what extent the assumptions on the network topology and the type of communications among the agents per iteration could be relaxed. Further, can we consider stochastic regimes where agents would only have access to unbiased estimators of the gradient of their objective functions? Motivated by big data applications, can we account for the high-dimensionality of the solution space where the computation of the local gradient mappings might become expensive? To address these research questions, one avenue lies in employing the randomized block variant of distributed stochastic gradient tracking schemes in both synchronous and asynchronous settings. Specifically, consider function $f_i : \mathbb{R}^n \to \mathbb{R}$ and the following distributed optimization problem

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} f_i(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{6.2.1}$$

where agents would be allowed to have an asynchronous communication over an undirected graph denoted by $\mathcal{G} \equiv (\mathcal{N}, \mathcal{E})$. Note that $\mathcal{N}$ is a set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of ordered pairs of vertices.

While block-coordinate schemes have been studied for distributed optimization problem (6.2.1) before, the convergence rate analysis is not yet done for asynchronous communication settings in the literature. An intriguing question is, can we develop a randomized block asynchronous scheme to address problem (6.2.1) and obtain the convergence rate statements?

REFERENCES

[1] T. Alpcan and T. Başar, *A game-theoretic framework for congestion control in general topology networks*, Proceedings of the 41st IEEE Conference on Decision and Control (CDC), 2002, pp. 1218–1224.

[2] ———, *Distributed algorithms for Nash equilibria of flow control games*, Advances in dynamic games, 2003, pp. 473–498.

[3] M. Amini and F. Yousefian, *An iterative regularized incremental projected subgradient method for a class of bilevel optimization problems*, American Control Conference (ACC), Philadelphia, PA, USA (2019), 4069–4074.

[4] ———, *An iterative regularized mirror descent method for ill-posed nondifferentiable stochastic optimization* (2019). arXiv:1901.09506v2.

[5] E. Anshelevich, A. Dasgupta, J. Kleinberg, É. Tardos, T. Wexler, and T. Roughgarden, *The price of stability for network design with fair cost allocation*, SIAM Journal on Computing **38** (2008), no. 4, 1602—1623.

[6] N. S. Aybat and E. Y. Hamedani, *A primal-dual method for conic constrained distributed optimization problems*, Advances in Neural Information Processing Systems (2016), 5049–5057.

[7] ———, *A distributed ADMM-like method for resource sharing over time-varying networks*, SIAM Journal on Optimization **29** (2019), no. 4, 3036–3068.

[8] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, *Distributed linearized alternating direction method of multipliers for composite convex consensus optimization*, IEEE Transactions on Automatic Control **63** (2018), no. 1, 5–20.

[9] A. Beck, *First-order methods in optimization*, MOS-SIAM Series on Optimization, Philadelphia, PA, 2017.

[10] A. Beck and S. Sabach, *A first order method for finding minimal norm-like solutions of convex optimization problems*, Mathematical Programming **147** (2014), no. 1-2, 25–46.

[11] D. P. Bertsekas, *Necessary and sufficient conditions for a penalty method to be exact*, Mathematical Programming **9** (1975), no. 1, 87–99.

[12] ———, *Constrained optimization and Lagrange multiplier methods*, Academic Press, New York, 1982.

[13] ———, *Incremental aggregated proximal and augmented Lagrangian algorithms* (2015). arXiv:1509.09257.

[14] ———, *Nonlinear programming: 3rd edition*, Athena Scientific, Bellmont, MA, 2016.

[15] ———, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey* (2017).

[16] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*, Athena Scientific, Belmont, MA, 2003.

[17] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM Journal on Optimization **18** (2007), no. 1, 29–51.

[18] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM Journal on Control And Optimization **29** (1991), no. 4, 968–998.

[19] Y. Censor, A. Gibali, and S. Reich, *The subgradient extragradient method for solvingvariational inequalities in hilbert space*, Journal of Optimization Theory and Applications **148** (2011), 318–335.

[20] ———, *Extensions of korpelevich's extragradient method for the variational inequality problem in euclidean space*, Optimization **61** (2012), no. 9, 1119–1132.

[21] T.-H. Chang, A. Nedić, and A. Scaglione, *Distributed constrained optimization by consensus-based primal-dual perturbation method*, IEEE Transactions on Automatic Control **59** (2014), no. 6, 1524–1538.

[22] T. Chen, A. Mokhtari, X. Wang, A. Ribeiro, and G. B. Giannakis, *Stochastic averaging for constrained optimization with application to online resource allocation*, IEEE Transactions on Signal Processing **65** (2017), no. 12, 3078–3093.

[23] X. Chen, C. Zhang, and M. Fukushima, *Robust solution of monotone stochastic linear complementarity problems*, Mathematical Programming **117** (2009), 51–80.

[24] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust region methods*, MPS-SIAM Series on Optimization, Society of Industrial and Applied Mathematics, Philadelphia, 2000.

[25] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses, *Selfish routing in capacitated networks*, Mathematics of Operations Research **29** (2004), no. 4, 961–976.

[26] S. Cui, U. V. Shanbhag, and F. Yousefian, *Complexity guarantees for an implicit smoothing-enabled method for stochastic MPECs* (2021). arXiv:2104.08406.

[27] C. D. Dang and G. Lan, *On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators*, Technical Report, Department of Industrial and Systems Engineering, University of Florida (2013).

[28] ———, *Stochastic block mirror descent methods for nonsmooth and stochastic optimization*, SIAM Journal on Optimization **25** (2015), no. 2, 856–881.

[29] 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.

[30] F. Facchinei and J-S. Pang, *Finite-dimensional variational inequalities and complementarity problems. Vols. I,II*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[31] M. C. Ferris and J-S. Pang, *Engineering and economic applications of complementarity problems*, SIAM Review **39** (1997), no. 4, 669–713.

[32] M. P. Friedlander and P. Tseng, *Exact regularization of convex programs*, SIAM Journal on Optimization **18** (2007), no. 4, 1326–1350.

[33] G. Garrigos, L. Rosasco, and S. Villa, *Iterative regularization via dual diagonal descent*, Journal of Mathematical Imaging and Vision **60** (2018), no. 2, 189–215.

[34] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, *On the convergence rate of incremental aggregated gradient algorithms*, SIAM Journal on Optimization **27** (2017), no. 2, 1035–1048.

[35] ———, *Convergence rate of incremental gradient and incremental Newton methods*, SIAM Journal on Optimization **29** (2019), no. 4, 2542–2565.

[36] E. Y. Hamedani and N. S. Aybat, *A primal-dual algorithm for general convex-concave saddle point problems* (2019). arXiv:1803.01401.

[37] ———, *A decentralized primal-dual method for constrained minimization of a strongly convex function* (2020). arXiv:1908.11835v2.

[38] S. P. Han and O. L. Mangasarian, *Exact penalty function in nonlinear programming*, Mathematical Programming **17** (1979), no. 1, 251–269.

[39] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM Journal on Optimization **27** (2016), no. 2, 686–724.

[40] A. N. Iusem, A. Jofré, and P. Thompson, *Incremental constraint projection methods for monotone stochastic variational inequalities*, Mathematics of Operations Research **44** (2019), no. 1, 236–263.

[41] A. N. Iusem and M. Nasri, *Korpelevich's method for variational inequality problems in banach spaces*, Journal of Global Optimization **50** (2011), 59–76.

[42] A. Jalilzadeh, *Primal-dual incremental gradient method for nonsmooth and convex optimization problems*, Optimization Letters (2021).

[43] H. Jiang and H. Xu, *Stochastic approximation approaches to the stochastic variational inequality problem*, IEEE Transactions on Automatic Control, **53** (2008), no. 6, 1462–1475.

[44] R. Johari, *Efficiency loss in market mechanisms for resource allocation*, Ph.D. Thesis, 2004.

[45] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stochastic Systems **1** (2011), no. 1, 17–58.

[46] A. Kannan and U. V. Shanbhag, *Distributed computation of equilibria in monotone nash games via iterative regularization techniques*, SIAM Journal on Optimization **22** (2012), no. 4, 1177–1205.

[47] A. Kannan, U. V. Shanbhag, and H. M. Kim, *Strategic behavior in power markets under uncertainty*, Energy Systems **2** (2011), 115–141.

[48] ———, *Addressing supply-side risk in uncertain power markets: stochastic Nash models, scalable algorithms and error analysis*, Optimization Methods and Software **28** (2013), 1095–1138.

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

[50] ———, *Distributed optimization for problems with variational inequality constraints* (2021). arXiv:2105.14205.

[51] ———, *An incremental gradient method for large-scale distributed nonlinearly constrained optimization*, Proceedings of the 2021 American Control Conference (ACC), 2021, pp. 953–958.

[52] ———, *A method with convergence rates for optimization problems with variational inequality constraints*, SIAM Journal on Optimization **31** (2021), no. 3, 2171–2198.

[53] V. Khatana and M. V. Salapaka, *DC-DistADMM: ADMM algorithm for constrained distributed optimization over directed graphs* (2020). arXiv:2003.13742.

[54] K. Knopp, *Theory and applications of infinite series*, Blackie & Son Ltd., Glasgow, Great Britain, 1951.

[55] G. M. Korpelevich, *An extragradient method for finding saddle points and for other problems*, Eknomika i Matematicheskie Metody **12** (1976), no. 4, 747—756.

[56] J. Koshal, A. Nedić, and U. V. Shanbhag, *Multiuser optimization: Distributed algorithms and error analysis*, SIAM Journal on Optimization **21** (2011), no. 3, 1046–1081.

[57] ———, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Transactions on Automatic Control **58** (2013), no. 3, 594–609.

[58] J. Lei, U. V. Shanbhag, J. S. Pang, and S. Sen, *On synchronous, asynchronous, and randomized best-response schemes for stochastic Nash games*, Mathematics of Operations Research **45** (2020), no. 1, 157–190.

[59] C. E. Lemke and J. T. Howson Jr., *Equilibrium points of bimatrix games*, Journal of the Society for Industrial and Applied Mathematics **12** (1964), no. 2, 413–423.

[60] A. Makhdoumi and A. Ozdaglar, *Convergence rate of distributed ADMM over networks*, IEEE Transactions on Automatic Control **62** (2017), no. 10, 5082 –5095.

[61] P. Marcotte and D. Zhu, *Weak sharp solutions of variational inequalities*, SIAM Journal on Optimization **9** (1998), no. 1, 179–189.

[62] A. Nedić, *Subgradient methods for convex minimization*, Ph.D. Thesis, 2002.

[63] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM Journal on Optimization **12** (2001), no. 1, 109–138.

[64] A. Nedić and J. Liu, *Distributed optimization for control*, Annual Review of Control, Robotics, and Autonomous Systems **1** (2018), 77–103.

[65] A. Nedić and T. Tatarenko, *Convergence rate of a penalty method for strongly convex problems with linear constraints*, 59th IEEE Conference on Decision and Control (CDC), 2020, pp. 372–377.

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

[67] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*, Cambridge University Press, New York, NY, USA, 2007.

[68] M. J. Osborne and A. Rubinstein, *A course in game theory*, MIT Press, Cambridge, Massachusetts, 1994.

[69] B. T. Polyak, *Introduction to optimization*, Optimization Software, Inc., New York, 1987.

[70] M. Rabbat and R. D. Nowak, *Distributed optimization in sensor networks*, The International Conference on Information Processing in Sensor Networks (2004), 20–27.

[71] P. Ricktárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming **144** (2014), no. 1-2, 1–38.

[72] R. T. Rockafellar and R.J-B Wets, *Variational analysis*, Springer, Berlin, 1998.

[73] T. Roughgarden, *Stackelberg scheduling strategies*, SIAM Journal on Computing **33** (2004), no. 2, 332–350.

[74] N. L. Roux, M. Schmidt, and F.R. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, Advances in Neural Information Processing Systems (2012), 2663–2671.

[75] S. Sabach and S. Shtern, *A first order method for solving convex bilevel optimization problems*, SIAM Journal on Optimization **27** (2017), no. 2, 640–660.

[76] H. Scarf, *The approximation of fixed points of a continuous mapping*, SIAM Journal on Applied Mathematics **15** (1967), no. 5, 1328–1343.

[77] G. Scutari, D. P. Palomar, F. Facchinei, and J-S. Pang, *Convex optimization, game theory, and variational inequality theory*, IEEE Signal Processing Magazine **27** (2010), no. 3, 35–49.

[78] G. Scutari, D. P. Palomar, F. Facchinei, and J. S. Pang, *Monotone games for cognitive radio systems* (2012), 83–112.

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

[80] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research **14** (2013), 567–599.

[81] U. V. Shanbhag, G. Infanger, and P. W. Glynn, *A complementarity framework for forward contracting under uncertainty*, Operations Research **59** (2011), no. 4, 810–834.

[82] M. V. Solodov, *An explicit descent method for bilevel convex optimization*, Journal of Convex Analysis **14** (2007), no. 2, 227–237.

[83] K. Sun and X. A. Sun, *A two-level distributed algorithm for general constrained nonconvex optimization with global convergence* (2021). arXiv:1902.07654.

[84] W. Tang and P. Daoutidis, *Distributed nonlinear model predictive control through accelerated parallel ADMM*, Proceedings of the 2019 American Control Conference (ACC), 2019, pp. 1406–1411.

[85] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill posed problems*, Winston and Sons, Washington DC., 1977.

[86] B. Turan, C. A. Uribe, H. T. Wai, and M. Alizadeh, *Resilient primal–dual optimization algorithms for distributed resource allocation*, IEEE Transactions on Control of Network Systems **8** (2021), no. 1, 282–294.

[87] N. D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar, *Global convergence rate of proximal incremental aggregated gradient methods*, SIAM Journal on Optimization **28** (2018), no. 2, 1282–1300.

[88] J. Wang, G. Scutari, and D. P. Palomar, *Robust mimo cognitive radio via game theory*, IEEE Transactions on Signal Processing **59** (2011), no. 3, 1183–1201.

[89] M. Wang and D. P. Bertsekas, *Incremental constraint projection methods for variational inequalities*, Mathematical Programming (Series A.) **150** (2015), no. 2, 321–363.

[90] E. Wei and A. Ozdaglar, *On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers*, In 2013 IEEE Global Conference on Signal and Information Processing (2013), 551–554.

[91] H.-K. Xu, *Viscosity approximation methods for nonexpansive mappings*, Journal of Mathematical Analysis and Applications **298** (2004), no. 1, 279–291.

[92] Y. Xu, *Accelerated first-order primal dual proximal methods for linearly constrained composite convex programming*, SIAM Journal on Optimization **27** (2017), no. 3, 1459–1484.

[93] H. Yin, U. V. Shanbhag, and P. G. Mehta, *Nash equilibrium problems with scaled congestion costs and shared constraints*, IEEE Transactions on Automatic Control **56** (2011), no. 7, 1702–1708.

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

[95] F. Yousefian, A. Nedić, and U. V. Shanbhag, *On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems*, Mathematical Programming **165** (2017), no. 1, 391–431. DOI: 10.1007/s10107-017-1175-y.

[96] ———, *On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems*, Mathematical Programming **165** (2017), 391–431.

[97] ———, *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. DOI: 10.1007/s11228-018-0472-9.

[98] ———, *On stochastic and deterministic quasi-Newton methods for nonstrongly convex optimization: Asymptotic convergence and rate analysis*, SIAM Journal on Optimization **30** (2020), no. 2, 1144–1172.

[99] C. Zhang and X. Chen, *Stochastic nonlinear complementarity problem and applications to traffic equilibrium under uncertainty*, Journal of Optimization Theory and Applications **137** (2008), no. 277, 51–80.

APPENDICES

A.1 Proof of Lemma 1.1.1

Proof. Let us define the function $\phi : \mathbb{R}^n \to \mathbb{R}$ as $\phi(x) \triangleq \frac{1}{2}\|Ax-b\|^2+\frac{1}{2}\sum_{j=1}^{J}\left(\max\{0, h_j(x)\}\right)^2$. We first note that ϕ is a differentiable function such that $\nabla\phi(x) = F(x)$ where F is given by Lemma 1.1.1 (e.g., see page 380 in [14]). Next, we also note that ϕ is convex. To see this, note that from the convexity of $h_j(x)$, the function $h_j^+(x) \triangleq \max\{0, h_j(x)\}$ is convex. Then, the function $\left(h_j^+(x)\right)^2$ can be viewed as a composition of $s(u) \triangleq u^2$ for $u \in \mathbb{R}$ and the convex function h_j^+. Since h_j^+ is nonnegative on its domain and $s(u)$ is nondecreasing on $[0, +\infty)$, we have that $\left(h_j^+(x)\right)^2$ is a convex function. As such, ϕ is a convex function as well. Consequently, from the first-order optimality conditions for convex programs, we have $\mathrm{SOL}(X, F) = \mathrm{argmin}_{x\in X}\,\phi(x)$. To show the desired equivalence between problems Problem (P_1) and Problem 1.1.2, it suffices to show that $\mathcal{X} = \mathrm{argmin}_{x\in X}\,\phi(x)$ where $\mathcal{X}$ denotes the feasible set of problem Problem 1.1.2. To show this statement, first we let $\bar{x} \in \mathcal{X}$. Then, from the definition of $\phi(x)$, we have $\phi(\bar{x}) = 0$. This implies that $\bar{x} \in \mathrm{argmin}_{x\in X}\,\phi(x)$. Thus, we have $\mathcal{X} \subseteq \mathrm{argmin}_{x\in X}\,\phi(x)$. Second, let $\tilde{x} \in \mathrm{argmin}_{x\in X}\,\phi(x)$. The feasibility assumption of the set $\mathcal{X}$ implies that there exists an $x_0 \in X$ such that $Ax_0 = b$ and $h_j(x_0) \leq 0$ for all j. This implies that $\phi(x_0) = 0$. From the nonnegativity of ϕ and that $\tilde{x} \in \mathrm{argmin}_{x\in X}\,\phi(x)$, we must have $\phi(\tilde{x}) = 0$ and $\tilde{x} \in X$. Therefore, we obtain $A\tilde{x} = b$, $h_j(\tilde{x}) \leq 0$ for all j, and $\tilde{x} \in X$. Thus, we have $\mathrm{argmin}_{x\in X}\,\phi(x) \subseteq \mathcal{X}$. Hence, we conclude that $\mathcal{X} = \mathrm{argmin}_{x\in X}\,\phi(x) = \mathrm{SOL}(X, F)$ and the proof is completed. ■

A.2 Proof of Lemma 1.2.1

Proof. For each $i \in [m]$, let function $\Theta_i : \mathbb{R}^n \to \mathbb{R}$ as

$$\Theta_i(x) \triangleq \frac{1}{2}\|A_ix - b_i\|^2 + \frac{1}{2}\sum_{j=1}^{n_i}\left(\max\{0, g_{i,j}(x)\}\right)^2.$$

Note that $\frac{1}{2}\|A_ix - b_i\|^2$ is a continuously differentiable and convex function for all i. Also, for all i and j, the function $\frac{1}{2}\left(\max\{0, g_{i,j}(x)\}\right)^2$ is continuously differentiable with the gradient map of $\max\{0, g_{i,j}(x)\}\nabla g_{i,j}(x)$ (see page 380 in [14]). Thus, we have that $\nabla\Theta_i(x) = F_i(x)$ where $F_i(x)$ is given in the statement of Lemma 1.2.1. Next, we show that F_i is a monotone mapping. From the convexity of $g_{i,j}(x)$, the function $\max\{0, g_{i,j}(x)\}$ is convex. Now, note that the function $\frac{1}{2}\left(\max\{0, g_{i,j}(x)\}\right)^2$ can be viewed as the composition of the nondecreasing function $h(y) \triangleq \frac{1}{2}y^2$ for $y \in \mathbb{R}_+$ and the convex function $\max\{0, g_{i,j}(x)\}$. Thus, $\frac{1}{2}\left(\max\{0, g_{i,j}(x)\}\right)^2$ is convex. This implies that Θ_i is a convex function and consequently, its gradient mapping that is $F_i(x)$, is monotone. Recalling the first-order optimality conditions for the convex optimization problems and taking into account the definition

of $\mathrm{SOL}(X, \sum_{i=1}^m F_i)$, we have that $\mathrm{SOL}(X, \sum_{i=1}^m F_i) = \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$. Let $\mathcal{Z}$ denote the feasible set of problem (1.1.2). To complete the proof, it suffices to show that $\mathcal{Z} = \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$. First, consider an arbitrary $\bar{x} \in \mathcal{Z}$. Then, from the definition of $\Theta_i(x)$ we must have $\Theta_i(\bar{x}) = 0$ for all i, implying that $\sum_{i=1}^m \Theta_i(\bar{x}) = 0$. Since the feasible set of problem (1.1.2) is nonempty and that $\sum_{i=1}^m \Theta_i(x) \geq 0$ for all $x \in X$, we conclude that $\bar{x} \in \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$. Thus, we showed that $\mathcal{Z} \subseteq \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$. Now, consider an arbitrary $\tilde{x} \in \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$. Thus, $\tilde{x} \in X$. Also, the assumption that the feasible set of problem (1.1.2) is nonempty implies that there exits $x_0 \in X$ such that $A_i x_0 = b_i$, $g_{i,j}(x_0) \leq 0$ for all $i \in [m]$ and $j \in [n_i]$. Thus, we have $\sum_{i=1}^m \Theta_i(x_0) = 0$. From the nonnegativity of the function $\sum_{i=1}^m \Theta_i(x)$ and that $\tilde{x} \in \operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x)$, we must have $\sum_{i=1}^m \Theta_i(\tilde{x}) = 0$. Therefore, we obtain $A_i\tilde{x} = b_i$, $g_{i,j}(\tilde{x}) \leq 0$ for all $i \in [m]$ and $j \in [n_i]$. Thus, we have shown that $\operatorname{argmin}_{x\in X} \sum_{i=1}^m \Theta_i(x) \subseteq \mathcal{Z}$. Hence, we have $\mathcal{Z} = \mathrm{SOL}(X, \sum_{i=1}^m F_i)$ and the proof is completed. ■

A.3 Proof of Lemma 2.3.1

Proof. We use induction to show $\bar{x}_N = \sum_{k=0}^N \lambda_{k,N} x_k$ for any $N \geq 0$. For $N = 0$, the relation holds due to the initialization $\bar{x}_0 := x_0$ in Algorithm 2 and that $\lambda_{0,0} = 1$. Next, let the relation hold for some $N \geq 0$. From the hypothesis, equation (2.3.2), and that $S_N = \sum_{k=0}^N \gamma_k^r$ for all $N \geq 0$, we can write

$$\bar{x}_{N+1} = \frac{S_N \bar{x}_N + \gamma_{N+1}^r x_{N+1}}{S_{N+1}} = \frac{\sum_{k=0}^{N+1} \gamma_k^r x_k}{\sum_{k=0}^{N+1} \gamma_k^r} = \sum_{k=0}^{N+1} \lambda_{k,N+1} x_k,$$

implying that the induction hypothesis holds for $N + 1$. Thus, we conclude that the desired averaging formula holds for all $N \geq 0$. To complete the proof, note that since $\sum_{k=0}^N \lambda_{k,N} = 1$, under the convexity of the set X, we have $\bar{x}_N \in X$. ■

A.4 Proof of Lemma 2.3.2

Proof. (a) From Definition 2.3.5, we can write

$$\mathsf{E}[\Delta_k \mid \mathcal{F}_k] = F(x_k) - \sum_{i=1}^d \mathsf{p}_i \mathsf{p}_i^{-1} \mathbf{U}_i F_i(x_k) = F(x_k) - \sum_{i=1}^d \mathbf{U}_i F_i(x_k) = 0.$$

The relation $\mathsf{E}[\delta_k \mid \mathcal{F}_k] = 0$ can be shown in a similar fashion.
(b) We can write

$$\begin{aligned}
\mathsf{E}\big[\|\Delta_k\|^2 \mid \mathcal{F}_k\big] &= \sum_{i=1}^d \mathsf{p}_i \left\|F(x_k) - \mathsf{p}_i^{-1} \mathbf{U}_i F_i(x_k)\right\|^2 \\
&= \sum_{i=1}^d \mathsf{p}_i \left(\|F(x_k)\|^2 + \mathsf{p}_i^{-2} \|\mathbf{U}_i F_i(x_k)\|^2 - 2\mathsf{p}_i^{-1} F(x_k)^T \mathbf{U}_i F_i(x_k)\right) \\
&= \|F(x_k)\|^2 + \sum_{i=1}^d \mathsf{p}_i^{-1} \|\mathbf{U}_i F_i(x_k)\|^2 - 2\sum_{i=1}^d \|F_i(x_k)\|^2 \leq \left(\mathsf{p}_{min}^{-1} - 1\right) C_F^2.
\end{aligned}$$

The relation $\mathsf{E}[\|\delta_k\|^2 \mid \mathcal{F}_k] \leq \left(\mathsf{p}_{min}^{-1} - 1\right) C_f^2$ can be shown using a similar approach. ■

A.5 Proof of Lemma 2.3.3

Proof. Given $0 \leq \alpha < 1$, let us define the function $\phi : \mathbb{R}_{++} \to \mathbb{R}$ as $\phi(x) \triangleq x^{-\alpha}$ for all $x > 0$. Since $\alpha > 0$, the function ϕ is nonincreasing. We can write

$$\sum_{k=0}^{N} \frac{1}{(k+1)^{\alpha}} = 1 + \sum_{k=2}^{N+1} \frac{1}{k^{\alpha}} \leq 1 + \int_{1}^{N+1} \frac{dx}{x^{\alpha}} = 1 + \frac{(N+1)^{1-\alpha} - 1}{1-\alpha} \leq \frac{(N+1)^{1-\alpha}}{1-\alpha},$$

implying the desired upper bound. To show that the lower bound holds, we can write

$$\sum_{k=0}^{N} \frac{1}{(k+1)^{\alpha}} = \sum_{k=1}^{N+1} \frac{1}{k^{\alpha}} \geq \int_{1}^{N+2} \frac{dx}{x^{\alpha}} \geq \int_{1}^{N+1} \frac{dx}{x^{\alpha}} \geq \frac{(N+1)^{1-\alpha} - 0.5(N+1)^{1-\alpha}}{1-\alpha},$$

where the last inequality is obtained using the assumption that $N \geq 2^{\frac{1}{1-\alpha}} - 1$. Therefore, the desired lower bound holds as well. This completes the proof. ■

A.6 Proof of Corollary 2.4.1

Proof. Let us rewrite Problem (P_1) as the equivalent problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathrm{SOL}(Y, F), \end{array} \tag{A.1}$$

where $Y \triangleq \prod_{i=1}^{d'} Y_i$ and $d' \triangleq 1$ and $Y_1 \triangleq X$. Note that this setting immediately implies that $Y = X$. Now, let us consider Algorithm 2 for solving Problem (A.1) where we assume that $x_0 \in X$ is an arbitrary fixed vector. Since $d' = 1$, Assumption 4.3.2 holds with $\mathrm{Prob}\,(i_k = 1) = 1$ for all $k \geq 0$. This setting implies that Algorithm 2 reduces to a deterministic scheme where the step 5 in Algorithm 2 is equivalent to the following update rule

$$x_{k+1} := \mathcal{P}_X\Big(x_k - \gamma_k \Big(F\left(x_k\right) + \eta_k \tilde{\nabla} f\left(x_k\right)\Big)\Big), \tag{A.2}$$

where we used $Y = Y_1 = X$. Next, we note that from the properties of the Euclidean projection mapping, for any $z \in X$ where $X \triangleq \prod_{i=1}^{d} X_i$, we have that $\mathcal{P}_X(z) = \prod_{i=1}^{d} \mathcal{P}_{X_i}\big(z^{(i)}\big)$. In view of this property, the equation (A.2) compactly represents the d updates given by equation (2.4.21). Therefore, Algorithm 3 is equivalent to Algorithm 2 and thus, all the results in Theorem 2.4.1 will hold with $\mathsf{p}_{min} = 1$. Note that in both equation (2.4.18) and equation (2.4.19), the expectation is eliminated. This completes the proof. ■

A.7 Proof of Lemma 2.5.1

Proof. (a) From the definition of x^* and $x^*_{\eta_k}$ (cf. Definition 2.5.1), we have

$$F(x^*)^T\left(x - x^*\right) \geq 0 \qquad \text{for all } x \in X, \tag{A.3}$$

$$\left(F\left(x^*_{\eta_k}\right) + \eta_k \nabla f\left(x^*_{\eta_k}\right)\right)^T \left(y - x^*_{\eta_k}\right) \geq 0 \qquad \text{for all } y \in X. \tag{A.4}$$

For $x := x^*_{\eta_k}$ and $y := x^*$, adding the resulting two relations together, we obtain

$$\eta_k \nabla f\left(x^*_{\eta_k}\right)^T \left(x^* - x^*_{\eta_k}\right) \geq \left(F\left(x^*\right) - F\left(x^*_{\eta_k}\right)\right)^T \left(x^* - x^*_{\eta_k}\right).$$

From the monotonicity of the mapping F and the preceding relation, we obtain that $\nabla f\left(x^*_{\eta_k}\right)^T \left(x^* - x^*_{\eta_k}\right) \geq 0$. Also, from the strong convexity of f, we have

$$f(x^*) \geq f\left(x^*_{\eta_k}\right) + \nabla f\left(x^*_{\eta_k}\right)^T \left(x^* - x^*_{\eta_k}\right) + \frac{\mu_f}{2}\left\|x^* - x^*_{\eta_k}\right\|^2.$$

From the preceding relations, we obtain

$$f(x^*) \geq f\left(x^*_{\eta_k}\right) + \frac{\mu_f}{2}\left\|x^* - x^*_{\eta_k}\right\|^2 \qquad \text{for all } k \geq 0. \tag{A.5}$$

Thus, $f(x^*) \geq f\left(x^*_{\eta_k}\right)$ for all $k \geq 0$. Recall that from Remark 2.5.1, under Assumption 2.5.1, $x^* \in X$ and $x^*_{\eta_k} \in X$ both exist and are unique. Therefore, $f\left(x^*_{\eta_k}\right)$ is bounded above for all $k \geq 0$. From this statement and invoking the coercive property of f (implied by the strong convexity of f), we can conclude that $\{x^*_{\eta_k}\}$ is a bounded sequence. Therefore, it must have at least one limit point. Let $\{x^*_{\eta_k}\}_{k\in\mathcal{K}}$ be an arbitrary subsequence such that $\lim_{k\to\infty,\ k\in\mathcal{K}} x^*_{\eta_k} = \hat{x}$, where $\lim_{k\to\infty,\ k\in\mathcal{K}}$ denotes the subsequential limit when $k \in \mathcal{K}$ and k goes to infinity. We show that $\hat{x} \in \mathrm{SOL}(X, F)$. Taking the limit from both sides of equation (A.4) with respect to the aforementioned subsequence and using the continuity of F and ∇f, we obtain that for all $y \in X$, $\left(F\left(\hat{x}\right) + \lim_{k\to\infty,\ k\in\mathcal{K}} \eta_k \nabla f\left(\hat{x}\right)\right)^T \left(y - \hat{x}\right) \geq 0$. Note that the mapping $\nabla f\left(\hat{x}\right)$ is bounded. This is because $\hat{x} \in X$ (due to the closedness of X) and that ∇f is continuous on the set X. Therefore, from the preceding inequality and $\lim_{k\to\infty} \eta_k = 0$, we obtain $F\left(\hat{x}\right)^T \left(y - \hat{x}\right) \geq 0$ for all $y \in X$, implying that $\hat{x} \in \mathrm{SOL}(X, F)$ and so, $\hat{x}$ is a feasible solution to Problem (P_1). Next, we show that $\hat{x}$ is the optimal solution to Problem (P_1). From equation (A.5), continuity of f, and neglecting the term $\frac{\mu_f}{2}\left\|x^* - x^*_{\eta_k}\right\|^2$, we obtain $f\left(x^*\right) \geq f\left(\lim_{k\to\infty,\ k\in\mathcal{K}} x^*_{\eta_k}\right) = f(\hat{x})$. Hence, from the uniqueness of x^*, all the limit points of $\{x^*_{\eta_k}\}$ fall in the singleton $\{x^*\}$ and the proof is completed.
(b) If $x^*_{\eta_k} = x^*_{\eta_{k-1}}$, the desired relation holds. Suppose for $k \geq 1$, we have $x^*_{\eta_k} \neq x^*_{\eta_{k-1}}$. From $x^*_{\eta_{k-1}} \in \mathrm{SOL}\left(X, F + \eta_{k-1}\nabla f\right)$ and $x^*_{\eta_k} \in \mathrm{SOL}\left(X, F + \eta_k \nabla f\right)$, we have

$$\begin{aligned}
&\left(F\left(x^*_{\eta_{k-1}}\right) + \eta_{k-1}\nabla f\left(x^*_{\eta_{k-1}}\right)\right)^T \left(x - x^*_{\eta_{k-1}}\right) \geq 0 \qquad \text{for all } x \in X,\\
&\left(F\left(x^*_{\eta_k}\right) + \eta_k \nabla f\left(x^*_{\eta_k}\right)\right)^T \left(y - x^*_{\eta_k}\right) \geq 0 \qquad \text{for all } y \in X.
\end{aligned}$$

Adding the resulting two relations together, for $x := x^*_{\eta_k}$ and $y := x^*_{\eta_{k-1}}$ we have

$$\left(-F\left(x^*_{\eta_k}\right) - \eta_k \nabla f\left(x^*_{\eta_k}\right) + F(x^*_{\eta_{k-1}}) + \eta_{k-1}\nabla f\left(x^*_{\eta_{k-1}}\right)\right)^T \left(x^*_{\eta_k} - x^*_{\eta_{k-1}}\right) \geq 0.$$

The monotonicity of F implies that $\left(F\left(x^*_{\eta_k}\right) - F\left(x^*_{\eta_{k-1}}\right)\right)^T \left(x^*_{\eta_k} - x^*_{\eta_{k-1}}\right) \geq 0$. Adding this relation to the preceding inequality, we have

$$\left(\eta_k \nabla f\left(x^*_{\eta_k}\right) - \eta_{k-1}\nabla f\left(x^*_{\eta_{k-1}}\right)\right)^T \left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right) \geq 0.$$

Adding and subtracting the term $\eta_k \nabla f\left(x^*_{\eta_{k-1}}\right)^T\left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right)$, we obtain

$$\begin{aligned}&\left(\eta_k - \eta_{k-1}\right)\nabla f\left(x^*_{\eta_{k-1}}\right)^T\left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right) \geq \\ &\eta_k\left(\nabla f\left(x^*_{\eta_{k-1}}\right) - \nabla f\left(x^*_{\eta_k}\right)\right)^T\left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right). \end{aligned} \tag{A.6}$$

From the strong convexity of function f, we have

$$\left(\nabla f\left(x^*_{\eta_{k-1}}\right) - \nabla f\left(x^*_{\eta_k}\right)\right)^T\left(x^*_{\eta_{k-1}} - x^*_{\eta_k}\right) \geq \mu_f\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\|^2. \tag{A.7}$$

From equation (A.6) and equation (A.7), and using the Cauchy-Schwarz inequality, we obtain

$$\left|\eta_k - \eta_{k-1}\right|\left\|\nabla f\left(x^*_{\eta_{k-1}}\right)\right\|\left\|x^*_{\eta_{k-1}} - x^*_{\eta_k}\right\| \geq \eta_k\mu_f\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\|^2,$$

Since $x^*_{\eta_k} \neq x^*_{\eta_{k-1}}$, dividing the both sides by $\eta_k\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\|$, we obtain

$$\left|1 - \frac{\eta_{k-1}}{\eta_k}\right|\left\|\nabla f\left(x^*_{\eta_{k-1}}\right)\right\| \geq \mu_f\left\|x^*_{\eta_k} - x^*_{\eta_{k-1}}\right\|. \tag{A.8}$$

From part (a), the trajectory $\{x^*_{\eta_k}\}$ is bounded. Also, for any $k \geq 0$, $x^*_{\eta_k} \in X$ by the definition. Since X is closed, there exists a compact set $S \subset X$ such that $\{x^*_{\eta_k}\} \subset S$. This statement and the continuity of ∇f imply that there exists $\bar{C}_f > 0$ such that $\left\|\nabla f\left(x^*_{\eta_{k-1}}\right)\right\| \leq \bar{C}_f$ for all $k \geq 1$. Thus, from equation (A.8), we obtain the desired inequality. ■

A.8 Proof of Lemma 3.3.3

Proof. Using $\beta \in [0, 1)$ and $\Gamma \geq 1$ we can write

$$\sum_{k=0}^{K}\frac{1}{(k+\Gamma)^\beta} \leq \int_{-1}^{K}\frac{dx}{(x+\Gamma)^\beta} = \frac{(K+\Gamma)^\beta - (\Gamma-1)^\beta}{1-\beta} \leq \frac{(K+\Gamma)^{1-\beta}}{1-\beta}.$$

We can also write

$$\begin{aligned}\sum_{k=0}^{K}\frac{1}{(k+\Gamma)^\beta} &\geq \int_{0}^{K+1}\frac{dx}{(x+\Gamma)^\beta} = \frac{(K+1+\Gamma)^{1-\beta} - \Gamma^{1-\beta}}{1-\beta} \\ &\geq \frac{(K+\Gamma)^{1-\beta} - 0.5(K+\Gamma)^{1-\beta}}{1-\beta},\end{aligned}$$

where the last inequality is implied from $K \geq \left(2^{\frac{1}{1-\beta}} - 1\right)\Gamma$. From the preceding relations we observe that the desired relation holds. ■

A.9 Proof of Lemma 3.5.3

Proof. (i) This result holds directly from Definition 3.5.2 and that $\gamma_0 = \frac{\gamma}{\Gamma^a}$ and $\eta_0 = \frac{\eta}{\Gamma^b}$.
(ii) For any $k \geq 1$ we have

$$1 - \frac{\eta_{k_2}}{\eta_{k_1}} = 1 - \frac{\eta(k_2+\Gamma)^{-b}}{\eta(k_1+\Gamma)^{-b}} = 1 - \left(\frac{k_1+\Gamma}{k_2+\Gamma}\right)^b \leq 1 - \sqrt{\frac{k_1+\Gamma}{k_2+\Gamma}},$$

where the last inequality is due to $b < 0.5$ and that $k_1 \leq k_2$. We obtain

$$1 - \frac{\eta_{k_2}}{\eta_{k_1}} \leq \frac{1 - \frac{k_1+\Gamma}{k_2+\Gamma}}{1 + \sqrt{\frac{k_1+\Gamma}{k_2+\Gamma}}} \leq \frac{k_2 - k_1}{k_2+\Gamma}.$$

(iii) Let us use (ii) for $k_1 := k-1$ and $k_2 := k$. For all $k \geq 1$ we have

$$\begin{aligned}\frac{1}{\gamma_k^3 \eta_k}\left(1 - \frac{\eta_k}{\eta_{k-1}}\right)^2 &\leq \frac{(k+\Gamma)^{3a}(k+\Gamma)^b}{\gamma^3\eta(k+\Gamma)^2} = \frac{1}{\gamma^3\eta(k+\Gamma)^{2-3a-b}} \\ &\leq \frac{1}{\gamma^3\eta\Gamma^{2-3a-b}},\end{aligned}$$

where the last relation is implied by $3a + b < 2$, $\Gamma > 0$, and that $k \geq 1$.
(iv) For all $k \geq 1$ we can write

$$\begin{aligned}\frac{1}{\gamma_k\eta_k\mu_{\min}}\left(\frac{\eta_k\gamma_{k-1}}{\gamma_k\eta_{k-1}} - 1\right) &= \frac{(k+\Gamma)^{a+b}}{\gamma\eta\mu_{\min}}\left(\left(1 + \frac{1}{k+\Gamma-1}\right)^{a-b} - 1\right) \\ &\leq \frac{(k+\Gamma)^{a+b}}{\gamma\eta\mu_{\min}(k+\Gamma-1)},\end{aligned}$$

where the last relation is implied by $a - b < 1$, $k \geq 1$, and $\Gamma \geq 1$. We obtain

$$\begin{aligned}\frac{1}{\gamma_k\eta_k\mu_{\min}}\left(\frac{\eta_k\gamma_{k-1}}{\gamma_k\eta_{k-1}} - 1\right) &\leq \frac{1}{\gamma\eta\mu_{\min}(k+\Gamma)^{1-a-b}}\left(1 + \frac{1}{\Gamma}\right) \\ &\leq \frac{2}{\gamma\eta\mu_{\min}\Gamma^{1-a-b}} \leq 0.5,\end{aligned}$$

where the last two relations are implied by $\Gamma \geq 1$ and $\Gamma^{1-a-b} \geq \frac{4}{\gamma\eta\mu_{\min}}$, respectively. This implies the relation in part (iv). ■

VITA

Harshal D. Kaushik

Candidate for the Degree of

Doctor of Philosophy

Dissertation: ON DISTRIBUTED OPTIMIZATION PROBLEMS WITH VARIATIONAL INEQUALITY CONSTRAINTS: ALGORITHMS, COMPLEXITY ANALYSIS, AND APPLICATIONS

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in December, 2021.

Completed the requirements for the Master of Science in Applied Mechanics at the Indian Institute of Technology (IIT), Madras (Chennai), Tamil Nadu in 2015.

Completed the requirements for the Bachelor of Science in Mechanical Engineering at the University of Pune, Maharashtra in 2012.

Professional Membership:

Society for Industrial and Applied Mathematics (SIAM)
Institute of Operations Research and Management Science (INFORMS)