2017

Distributed nonconvex optimization: Algorithms and convergence analysis

Davood Hajinezhad
Iowa State University

Distributed nonconvex optimization: Algorithms and convergence analysis

by

Davood Hajinezhad

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Mingyi Hong, Co-major Professor
Gary Mirka, Co-major Professor
Sarah Ryan
Sigurdur Olafsson
Zhengdao Wang
Peng Wei

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2017

DEDICATION

This dissertation is for my beloved wife, Elham. Thanks for being always with me, for your love, and endless support in this long trip.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

With immense appreciation, I would like to express my gratitude to the people who helped me to bring this study into success.

First, my major professor, Dr. Mingyi Hong for valuable guidance, boundless knowledge, consistent advices, patience, and continuous support in all aspects of my PhD life. You certainly provided me with the necessary tools that I needed to successfully complete my graduate program and this dissertation.

Additionally I would like to acknowledge my committee members Dr. Mirka, Dr. Ryan, Dr. Olafsson form Industrial and Manufacturing System Engineering department, Dr. Wang from Electrical and Computer Engineering department, and Dr. Wei from Aerospace Engineering department. Your feedbacks definitely improved the quality of my research.

Finally, topmost gratitude to my parents. You are always there for me. Doubtlessly, I was not here without your love and spiritual supports throughout my life.

ABSTRACT

This thesis addresses the problem of distributed optimization and learning over multi-agent networks. Our main focus is to design efficient algorithms for a class of nonconvex problems, defined over networks in which each agent/node only has partial knowledge about the entire problem. Multi-agent nonconvex optimization has gained much attention recently due to its wide applications in big data analysis, sensor networks, signal processing, multi-agent network, resource allocation, communication networks, just to name a few. In this work, we develop a general class of primal-dual algorithms for distributed optimization problems in challenging setups, such as nonconvexity in loss functions, nonsmooth regularizations, and coupling constraints. Further, we consider different setup where each agent can only access the *zeroth-order information* (i.e., the functional values) of its local functions. Rigorous convergence and rate of convergence analysis is provided for the proposed algorithms. Our work represents one of the first attempts to address nonconvex optimization and learning over networks.

CHAPTER 1. INTRODUCTION

This research mainly focuses on designing algorithms for distributed *nonconvex* optimization problems under different network topologies. Distributed nonconvex optimization problem has found a wide range of applications in several areas, including data-intensive optimization [65, 146], signal and information processing [47, 117], multi-agent network resource allocation [134], communication networks [82], just to name a few. In particular, it is a key enabler of many emerging "big data" analytic tasks. In these data-intensive applications, the sheer volume and spatial/temporal disparity of big data render centralized processing and storage a formidable task. This happens, for instance, whenever the volume of data overwhelms the storage capacity of a single computing device. Moreover, collecting sensor-network data, which are observed across a large number of spatially scattered centers/servers/agents, and routing all this local information to centralized processors, under energy, privacy constraints and/or link/hardware failures, is often infeasible or inefficient. Further, with the advent of high performance and parallel computing interfaces it is reasonable to model the problem such that we are able to utilize these interfaces to accelerate the computations.

Typically, distributed optimization problem can be expressed as minimizing the sum of additively separable cost functions plus a regularization function, given below

$$\min_{x\in X} \; g(x) := \sum_{i=1}^{N} f_i(x) + h(x), \tag{1.1}$$

where N denotes the number of agents in the network; $f_i : \mathbb{R}^M \to \mathbb{R}$ represents some cost function related to the agent i, $X \subset \mathbb{R}^M$ is a convex set, and $h(x)$ imposes some regularity such as sparsity to the solution, or denotes the indicator function of a convex set. It is usually assumed that each agent i has complete information on f_i, and it can only communicate with its neighbors. Therefore the key objectives of the individual agents are: 1) to achieve consensus with its neighbors about the optimization variable; 2) to optimize the global objective function $g(x)$.

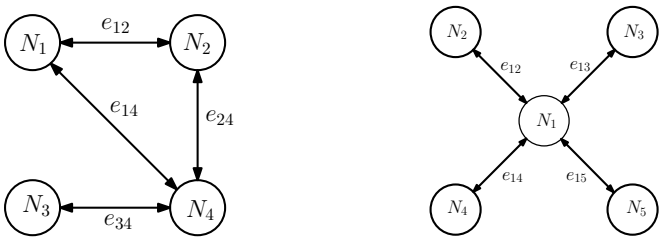


Figure 1.1: Left: Mesh Network (MNet); Right: Star Network (SNet)

We consider two popular network topologies, namely, the mesh network (MNet) (cf. Fig. 1.1 Left) and the star network (SNet) (cf. Fig. 1.1 Right). In the MNet, each node is connected via undirected links to a subset of nodes. Such a network is very popular in a number of applications. For example, in distributed signal processing [47, 117], each node can represent a sensor which has limited communication capability hence can only talk to its neighbors. On the other hand, SNet contains a central controller (i.e., the parent) which is connected to all other nodes (i.e., the children), and there is no connection between the children. Such a network can be used to model parallel computing architecture in which each child represents a computing node, and the parent coordinates the computation of the children [145, 80, 55]. In our work, we consider these different network topologies not only because they are capable of modeling a wide range of applications, but more importantly, their unique characteristics lead to a number of open challenges in designing distributed algorithms.

Extensive research has been done on developing algorithms for distributed optimization (for both MNet and SNet), but these works are mostly restricted to the family of *convex* problems where $f_i(x)$'s are all convex functions (detailed literature review is relegated to the following chapters). Once we go beyond the convexity, the literature is very scant. Therefore, this research is set out to fill such an important gap. Below we briefly describe three applications that motivate distributed nonconvex optimization.

- **Distributed Sparse Principal Component Analysis.** Principal component analysis (PCA) aims to reduce the dimension of multi-variate data set, and has a wide range of applications in science and engineering, see for example [50, 125, 121].

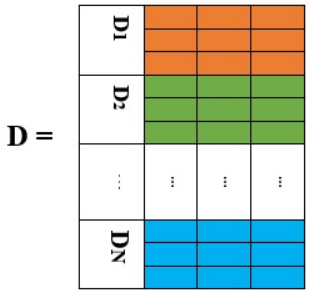


Figure 1.2: Splitting data matrix across the rows

Finding the first sparse principal component (PC) is equivalent to solving the following optimization problem

$$\max \ \|Dx\|_2^2 - \lambda r(x), \quad \text{s.t. } \|x\|_2^2 \leq 1 \tag{1.2}$$

where $D \in \mathbb{R}^{Q\times M}$ is a centered data matrix, and $\|Dx\|_2^2$ represents the explained variance of the first PC [72], $r(x)$ is a sparsity-promoting regularizer, and $\lambda > 0$ is the penalization parameter. In practice, $r(x)$ can take the form of the ℓ_0 norm of x, or its approximations such as the popular ℓ_1 norm, the log sum penalty (LSP) [22] and so on. In this research we consider the more challenging scenario where the data matrix D is not available in a central location, instead it is distributed over a network, and each agent has access to a mini-batch of the data points. In particular, let $D_i \in \mathbb{R}^{Q_i\times M}, i = 1, 2, \cdots N$ denote submatrices that consist non-overlapping rows (or data samples) of D. That is, $D = [D_1; D_2; \cdots; D_N]$; see Figure 1.2. According to this data model, the SPCA problem (1.2) can be reformulated as follows

$$\min \ \sum_{i=1}^{N} -\|D_i x\|_2^2 + \lambda r(x), \quad \text{s.t. } \|x\|_2^2 \leq 1. \tag{1.3}$$

This problem is nonconvex because $-\|D_i x\|_2^2$ in the objective is a concave function. Further, it is easy to see that this problem has the same form as that of (1.1), with $f_i(x) = -\|D_i x\|_2^2$, $h(x) = \lambda r(x)$, and $X = \{x \in \mathbb{R}^M;\ \text{s.t. } \|x\|_2^2 \leq 1\}$.

- **Distributed Nonlinear Regression Problem.** This application is about a distributed regression problem. Consider the MNet, which consists of N agents, and each agent i has Q_i

local observation pairs (z_{ij}, b_{ij}); $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots Q_i$. Suppose that each data point is generated in the following manner:

$$b_{ij} = \frac{1}{1 + \exp(-z_{ij}^\top x)} + \epsilon_{ij},$$

where ϵ_{ij} denotes the additive noise following a zero mean Gaussian distribution. Denote $h_i(x, z_{ij}) = \frac{1}{1+\exp(-z_{ij}^\top x)}$, then we can form the following nonconvex nonlinear least square problem [91]

$$\min_x \sum_{i=1}^{N} \sum_{j=1}^{Q_i} [b_{ij} - h_i(x, z_{ij})]^2 + r(x), \tag{1.4}$$

where again $r(x)$ is a sparsity promoting regularizer. Clearly, problem (1.4) is a special case of (1.1), with $f_i(x) := \sum_{j=1}^{Q_i} [b_{ij} - h_i(x, z_{ij})]^2$.

- **Distributed Target Localization Problem.** Consider a network of N agents collectively aim to locate M target points. Let's $x_i \in \mathbb{R}^d$, $i = 1, 2, \cdots M$ denote the coordinate of the each target location ($d = 2$ or 3). In target localization problem each agent i knows its own location w_i, as well as a noisy measurement d_{ij} of the squared distance to each target points j, $j = 1, 2, \cdots M$. Therefore, this problem seeks to minimize the following nonconvex optimization problem [26]

$$\min_x \sum_{i=1}^{N} \sum_{j=1}^{M} (d_{ij} - \|x_j - w_i\|^2)^2, \tag{1.5}$$

which is another special case of problem (1.1), with $f_i(x) = \sum_{j=1}^{M} (d_{ij} - \|x_j - w_i\|^2)^2$, $h(x) = 0$, and $X = \mathbb{R}^M$.

The rest of this thesis contains three parts, as described below:

- Chapter 2. In this chapter we consider nonconvex optimization problem over the MNet (cf. Fig. 1.1). Typically this problem is modeled as the following

$$\min_{x \in \mathbb{R}^M} \; g(x) := \sum_{i=1}^{N} f_i(x), \tag{1.6}$$

where each f_i, $i \in \{1, \cdots, N\} := [N]$ is a nonconvex cost function. To solve this problem we propose a proximal primal-dual algorithm (Prox-PDA). We show that Prox-PDA converges to the set of stationary solutions (satisfying the first-order optimality condition) in a globally sublinear manner. We also show that Prox-PDA can be extended in several directions to improve its practical performance. To the best of our knowledge, this is the first algorithm that is capable of achieving global sublinear convergence rate for distributed nonconvex optimization. Further, our work reveals an interesting connection between the *primal-dual* based algorithm Prox-PDA and the *primal-only* fast distributed algorithms such as EXTRA [122]. Finally, we generalize the theory for Prox-PDA based algorithms to a challenging distributed matrix factorization problem.

- Chapter 3. This chapter considers the SNet (given in Fig. 1.1 Right). Utilizing SNet we are able to solve the following problem

$$\min_{x \in X} \quad g(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) + f_0(x) + p(x), \tag{1.7}$$

where X is a closed and convex set; for each $i \in \{0, \cdots, N\}$, f_i is a smooth possibly nonconvex function; $p(x)$ is a lower semi-continuous convex but possibly nonsmooth function. Notice that in this problem we are able to deal with nonsmooth term $p(x)$ as well as constraint set X in contrast to the problem (1.6). We propose a class of NonconvEx primal-dual SpliTTing (NESTT) algorithms for this problem. The NESTT is one of the first stochastic algorithms for distributed nonconvex nonsmooth optimization, with provable and nontrivial convergence rates. The main contribution is the following. First, we show that NESTT converges sublinearly to a point belongs to stationary solution set of (1.7). Second, we show that NESTT converges Q-linearly for certain nonconvex ℓ_1 penalized quadratic problems. To the best of our knowledge, this is the first time that linear convergence is established for stochastic and distributed optimization of such type of problems.

- Chapter 4. This chapter focuses on a general class of optimization problem given below

$$\min_{x \in X} \; g(x) := f(x) + h(x), \quad \text{s.t.} \quad Ax = b, \tag{1.8}$$

where $f(x) : \mathbb{R}^N \to \mathbb{R}$ is a continuous smooth function (possibly nonconvex); $A \in \mathbb{R}^{M\times N}$ is a rank deficient matrix; $b \in \mathbb{R}^M$ is a given vector; X is a convex compact set; $h(x) : \mathbb{R}^N \to \mathbb{R}$ is a lower semi-continuous nonsmooth convex function.

Problem (1.8) subsumes a number of applications in the variety of domains, of which distributed composite optimization problem over MNet including nonconvex loss functions and nonsmooth regularizations is a very important yet challenging problem. In what follows we show how distributed optimization problems in different setups can be cast in general formulation (1.8).

The exact consensus problem over networks. Consider a network which consists of N agents who collectively optimize the following problem

$$\min_{y\in\mathbb{R}} \; f(y) + h(y) := \sum_{i=1}^{N} \left[f_i(y) + h_i(y)\right], \tag{1.9}$$

where $f_i(y) : \mathbb{R} \to \mathbb{R}$ is a smooth function, and $h_i(y) : \mathbb{R} \to \mathbb{R}$ is a convex, possibly nonsmooth regularizer (here y is assumed to be scalar for ease of presentation). Note that both f_i and h_i are local to agent i.

To integrate the structure of the network into problem (4.6), we consider MNet which is an undirected, connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with $|\mathcal{V}| = N$ vertices and $|\mathcal{E}| = E$ edges. Each agent can only communicate with its immediate neighbors, and it is responsible for optimizing one component function f_i regularized by h_i. Define the node-edge incidence matrix $A \in \mathbb{R}^{E\times N}$ as following: if $e \in \mathcal{E}$ and it connects vertex i and j with $i > j$, then $A_{ev} = 1$ if $v = i$, $A_{ev} = -1$ if $v = j$ and $A_{ev} = 0$ otherwise. Using this definition, the *signed graph Laplacian matrix* $L_- \in \mathbb{R}^{N\times N}$ is given by

$$L_- := A^T A.$$

Introducing N new variables x_i as the local copy of the global variable y, and define $x := [x_1; \cdots; x_N] \in \mathbb{R}^N$, problem (4.6) can be equivalently expressed as

$$\min_{x\in\mathbb{R}^N} \; f(x) + h(x) := \sum_{i=1}^{N} \big[f_i(x_i) + h_i(x_i)\big], \text{ s.t. } Ax = 0. \tag{1.10}$$

This problem is precisely original problem (1.8) with correspondence $X = \mathbb{R}^N$, $b = 0$, $f(x) := \sum_{i=1}^N f_i(x_i)$, and $h(x) := \sum_{i=1}^N h_i(x_i)$.

The partial consensus problem. In the previous application, the agents are required to reach *exact* consensus, and such constraint is imposed through $Ax = 0$ in (4.7). In practice, however, consensus is rarely achieved exactly, for example due to potential disturbances in network communication; see detailed discussion in [75]. Further, in applications ranging from distributed estimation to rare event detection, the data obtained by the agents, such as harmful algal blooms, network activities, and local temperature, often exhibit distinctive spatial structure [28]. The distributed problem in these settings can be best formulated by using certain partial consensus model in which the local variables of an agent are only required to be close to those of its neighbors. To model such *partial* consensus constraint, we denote ξ_e as the permissible tolerance for $e = (i, j) \in \mathcal{E}$, and replace the strict consensus constraint $x_i - x_j = 0$ with $\|x_i - x_j\|^2 \leq \xi_e$. Further, we define the link variable $z_e = x_i - x_j$, and set $z := \{z_e\}_{e\in\mathcal{E}}$, $Z := \{z \mid \|z_e\|^2 \leq \xi_e \ \forall\ e \in \mathcal{E}\}$. Using these notations, the partial consensus problem can be formulated as

$$\begin{aligned} \min_{x,z} \quad & f(x) + h(x) := \sum_{i=1}^N \big[f_i(x_i) + h_i(x_i)\big] \\ \text{s.t.} \quad & Ax - z = 0,\ z \in Z, \end{aligned} \tag{1.11}$$

which is again a special case of problem (1.8).

Notice that the application of optimization problem (1.8) are not limited to distributed optimization problem. More problems such as *sparse subspace estimation* will be discussed in chapter Chapter 4.

In Chapter 4 we develop an Uzawa type [74] algorithm named PProx-PDA for problem (1.8). One distinctive feature of the PProx-PDA is the use of a novel perturbation scheme for both the primal and dual steps, which is designed to ensure a number of asymptotic convergence and rate of convergence properties (to first-order stationary solutions). Specifically, we show that when certain perturbation parameter remains *constant* across the iterations,

the algorithm converges globally sublinearly to the set of approximate first-order stationary solutions. Further, when the perturbation parameter diminishes to zero with appropriate rate, the algorithm converges to the set of exact first-order stationary solutions. To the best of our knowledge this is the first time that first-order methods with convergence and rate of convergence guarantees are developed for problems in the form of (1.8).

- Chapter 5. This chapter focuses on nonconvex distributed optimization problem under the challenging *zeroth-order* setup. A drawback for the algorithms in previous chapters is that they require at least *first-order* gradient information in order to guarantee global convergence. Unfortunately, in many real-world problems, obtaining such information can be very expensive, if not impossible. For example, in simulation-based optimization [126], the objective function of the problem under consideration can only be evaluated using repeated simulation. In certain scenarios of training deep neural network [76], the relationship between the decision variables and the objective function is too complicated to derive explicit form of the gradient. Further, in bandit optimization [2, 37], a player tries to minimize a sequence of loss functions generated by an adversary, and such loss function can only be observed at those points in which it is realized. In these scenarios, one has to utilize techniques from derivative-free optimization, or optimization using zeroth-order information [127, 27].

  In this chapter we propose zeroth-order primal-dual based algorithms for distributed optimization problems over different network topologies. For MNet, we design an algorithm capable of dealing with nonconvexity and zeroth-order information simultaneously. It is shown that the proposed algorithm converges to the set of stationary solutions of problem (1.6) (with nonconvex but smooth f_i's), in a globally sublinear manner. Further, for SNet we propose a stochastic primal-dual based method, which is able to further utilize the special structure of the network (i.e., the presence of the central controller) and deal with problem (1.7) with nonsmooth objective in zeroth-order setup. Theoretically, we show that this algorithm also converges to the set of stationary solutions in a globally sublinearly manner.

To the best of our knowledge, these algorithms are the first ones for distributed nonconvex optimization that are capable of utilizing zeroth-order information, while possessing global convergence rate guarantees.

CHAPTER 2. PROXIMAL PIMAL-DUAL ALGORITHM FOR DISTRIBUTED NONCONVEX OPTIMIZATION

Abstract

In this paper we consider nonconvex optimization and learning over a network of distributed nodes. We develop a Proximal Primal-Dual Algorithm (Prox-PDA), which enables the network nodes to distributedly and collectively compute the set of first-order stationary solutions in a global sublinear manner [with a rate of $\mathcal{O}(1/r)$, where r is the iteration counter]. To the best of our knowledge, this is the first algorithm that enables distributed nonconvex optimization with global rate guarantees. Our numerical experiments also demonstrate the effectiveness of the proposed algorithm.

2.1 Introduction

We consider the following optimization problem

$$\min_{z\in\mathbb{R}^M} \; g(z) := \sum_{i=1}^{N} f_i(z), \tag{2.1}$$

where each f_i, $i \in \{1, \cdots, N\} := [N]$ is a nonconvex cost function, and we assume that it is smooth and has Lipschitz continuous gradient.

Such *finite sum* problem is of central importance in machine learning and signal/information processing [23, 47]. In particular, in the class of empirical risk minimization (ERM) problem, z represents the feature vectors to be learned, and each f_i can represent: 1) a mini-batch of (possibly nonconvex) loss functions modeling data fidelity [7]; 2) nonconvex activation functions of neural networks [3]; 3) nonconvex utility functions used in applications such as resource allocation [18]. Recently, a number of works in machine learning community have been focused on designing fast

centralized algorithms for solving problem (5.1); e.g., SAG [31], SAGA [118], and SVRG [71] for convex problems, and [111, 3, 57] for nonconvex problems.

In this work, we are interested in designing algorithms that solve problem (5.1) in a distributed manner. In particular, we focus on the scenario where each f_i (or equivalently, each subset of data points in the ERM problem) is available locally at a given computing node $i \in [N]$, and the nodes are connected via a network. Clearly, such distributed optimization and learning scenario is important for machine learning, because in contemporary applications such as document topic modeling and/or social network data analysis, oftentimes data corporas are stored in geographically distributed locations without any central controller managing the entire network of nodes; see [38, 140, 108, 11].

Related Works. Distributed *convex* optimization and learning has been thoroughly investigated in the literature. In [100], the authors propose a distributed subgradient algorithm (DSG), which allows the agents to jointly optimize problem (5.1). Subsequently, many variants of DSG have been proposed, either with special assumptions on the underlying graph, or having additional structures of the problem; see, e.g., [88, 89, 99]. The rate of convergence for DSG is $\mathcal{O}(\log(r)/\sqrt{r})$ under certain diminishing stepsize rules. Recently, a number of algorithms such as the exact first-order algorithm (EXTRA) [122] and DLM [85] have been proposed, which use constant stepsize and achieve faster $\mathcal{O}(1/r)$ rate for convex problems. Recent works that applies distributed optimization algorithms to machine learning applications include [115, 11, 116].

On the other hand, there has been little work for distributed optimization and learning when the objective function involves nonconvex problems. A dual subgradient method has been proposed in [148], which relaxes the exact consensus constraint. In [17] a stochastic projection algorithm using diminishing stepsizes has been proposed. An ADMM based algorithm has been presented in [63] for a special type of problem called *global consensus*, where all distributed nodes are directly connected to a central controller. Utilizing certain convexification decomposition technique the authors of [92] designed an algorithm named NEXT, which converges to the set of stationary solutions when using diminishing stepsizes. To the best of our knowledge, no distributed algorithm

is able to guarantee global convergence rate for problem (5.1), in the scenario where the nodes are distributed in connected a network.

Our Contributions. In this work, we propose a proximal primal-dual algorithm (Prox-PDA) for problem (5.1) over an undirected connected network. We show that Prox-PDA converges to the set of stationary solutions of problem (5.1) (satisfying the first-order optimality condition) in a globally sublinear manner. We also show that Prox-PDA can be extended in several directions to improve its practical performance. To the best of our knowledge, this is the first algorithm that is capable of achieving global sublinear convergence rate for distributed non-convex optimization.

Further, our work reveals an interesting connection between the *primal-dual* based algorithm Prox-PDA and the *primal-only* fast distributed algorithms such as EXTRA [122]. Such new insight into the connection between primal-dual and primal-only algorithms could be of independent interest for the optimization community. Finally, we generalize the theory for Prox-PDA based algorithms to a challenging distributed matrix factorization problem.

System Model

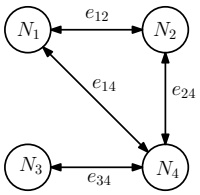
Define a graph $\mathcal{G} := \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ are the node and edge sets; Let $|\mathcal{V}| = N$ and $|\mathcal{E}| = E$. Each node $v \in \mathcal{V}$ represents an agent in the network, and each edge $e_{ij} = (i, j) \in \mathcal{E}$ indicates that node i and j are neighbors; see Fig.5.1(Left). Assume that each node i can only communicate with its direct neighbors, defined as $\mathcal{N}_i := \{j \mid (i, j) \in \mathcal{V}\}$, with $|\mathcal{N}_i| = d_i$. The distributed version of problem (5.1) is given as below

$$\min_{x_i \in \mathbb{R}^M} f(x) := \sum_{i=1}^{N} f_i(x_i), \text{ s.t. } x_i = x_j, \ \forall \ (i, j) \in \mathcal{E}. \tag{2.2}$$

Clearly the above problem is equivalent to (5.1) as long as $\mathcal{G}$ is connected. For notational simplicity, define $x := \{x_i\} \in \mathbb{R}^{NM \times 1}$, and $Q := N \times M$.

To proceed, let us introduce a few useful quantities related to graph $\mathcal{G}$.

• The *incidence matrix* $\tilde{A} \in \mathbb{R}^{E \times N}$ is a matrix with entires $\tilde{A}(k, i) = 1$ and $\tilde{A}(k, j) = -1$ if $k = (i, j) \in \mathcal{E}$ with $j > i$, and all the rest of the entries being zero. For example, for the network



$$\tilde{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Figure 2.2: (Left) An undirected Connected Network, (Right) Incidence Matrix.

in Fig.5.1 (Left); the incidence matrix is given in Fig.5.1 (Right). Define the *extended incidence matrix* as

$$A := \tilde{A} \otimes I_M \in \mathbb{R}^{EM \times Q}. \tag{2.3}$$

• The *Degree matrix* $\tilde{D} \in \mathbb{R}^{N \times N}$ is given by $\tilde{D} := \text{diag}[d_1, \cdots, d_N]$; Let $D := \tilde{D} \otimes I_M \in \mathbb{R}^{Q \times Q}$.

• The signed and the signless Laplacian matrices (denoted as L^- and L^+ respectively), are given below

$$L^- := A^\top A \in \mathbb{R}^{Q \times Q}, \; L^+ := 2D - A^\top A \in \mathbb{R}^{Q \times Q}. \tag{2.4}$$

Using the above notations, one can verify that problem (5.10) can be written in the following compact form:

$$\min_{x \in \mathbb{R}^Q} \; f(x), \quad \text{s.t.} \; \; Ax = 0. \tag{2.5}$$

2.2 The Prox-PDA Algorithm

The proposed algorithm builds upon the classical augmented Lagrangian (AL) method [14, 107]. Let us define the AL function for (5.13) as

$$L_\beta(x, \mu) = f(x) + \langle \mu, Ax \rangle + \frac{\beta}{2} \|Ax\|^2 \tag{2.6}$$

where $\mu \in \mathbb{R}^Q$ is the dual variable; $\beta > 0$ is a penalty parameter. Let $B \in \mathbb{R}^{Q \times Q}$ be some arbitrary matrix to be determined shortly. Then the proposed algorithm is given in the table below (Algorithm 1).

In Prox-PDA, the primal iteration (4.12a) minimizes the augmented Lagrangian plus a proximal term $\frac{\beta}{2}\|x - x^r\|^2_{B^TB}$. We emphasize that the proximal term is critical in both the algorithm

Algorithm 1 The Prox-PDA Algorithm

1: At iteration 0, initialize $\mu^0 = 0$ and $x^0 \in \mathbb{R}^Q$.
2: At each iteration $r+1$, update variables by:

$$x^{r+1} = \arg\min_{x\in\mathbb{R}^Q} \; f(x) + \langle \mu^r, Ax\rangle + \frac{\beta}{2}\|Ax\|^2 + \frac{\beta}{2}\|x - x^r\|^2_{B^TB}; \tag{2.7a}$$

$$\mu^{r+1} = \mu^r + \beta A x^{r+1}. \tag{2.7b}$$

implementation and the analysis. It is used to ensure the following key properties:

(1). The primal problem is strongly convex;

(2). The primal problem is decomposable over different network nodes, hence distributedly implementable.

To see the first point, suppose B^TB is chosen such that $A^TA + B^TB \succeq I_Q$, and that $f(x)$ has Lipschitz gradient. Then by a result in [150][Theorem 2.1], we know that there exists $\beta > 0$ large enough such that the objective function of (4.12a) is strongly convex.

To see the second point, Let $B := |A|$, where the absolute value is taken for each component of A. It can be verified that $B^TB = L^+$, and step (4.12a) becomes

$$\begin{aligned} x^{r+1} &= \arg\min_x \sum_{i=1}^N f_i(x_i) + \langle \mu^r, Ax\rangle + \frac{\beta}{2}x^TL^-x + \frac{\beta}{2}(x-x^r)^TL^+(x-x^r) \\ &= \arg\min_x \sum_{i=1}^N f_i(x_i) + \langle \mu^r, Ax\rangle + \beta x^TDx - \beta x^TL^+x^r \end{aligned}$$

Clearly this problem is *separable* over the nodes, therefore it can be solved completely distributedly.

2.3 The Convergence Analysis

In this section we provide convergence analysis for Algorithm 1. The key in the analysis is the construction of a novel potential function, which decreases at every iteration of the algorithm. We first state our main assumptions below.

[A1.] The function $f(x)$ is differentiable and has Lipschitz continuous gradient, i.e.,

$$\|\nabla f(x) - \nabla f(x)\| \leq L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^Q.$$

Further assume that $A^TA + B^TB \succeq I_Q$.

[A2.] There exists a constant $\delta > 0$ such that

$$\exists \underline{f} > -\infty, \quad \text{s.t. } f(x) + \frac{\delta}{2}\|Ax\|^2 \geq \underline{f},\ \forall\, x \in \mathbb{R}^Q.$$

Without loss of generality we will assume that $\underline{f} = 0$. Below we provide a few nonconvex smooth functions that satisfy our assumptions, all of which are commonly used as activation functions for neural networks.

- **The sigmoid function**. The sigmoid function is given by

  $$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \in [-1,\ 1].$$

  Clearly it satisfies [A2]. We have $\text{sigmoid}'(x) = \frac{e^{-x}}{(1+e^{-x})^2} \in [0,\ 1/4]$, and such boundedness of the first order derivative implies that [A1] is true (by applying the first-order mean value theorem).

- **The** arctan **function**. Note that $\arctan(x) \in [-1, 1]$, so it clearly satisfies [A2]. $\arctan'(x) = \frac{1}{x^2+1} \in [0,\ 1]$ so it is bounded, which implies that [A1] is true.

- **The** tanh **function**. Note that we have

  $$\tanh(x) \in [-1, 1], \quad \tanh'(x) = 1 - \tanh(x)^2 \in [0, 1].$$

  Therefore the function satisfies [A1] – [A2].

- **The** logit **function** as follows

  $$2\text{logit}(x) = \frac{2e^x}{e^x + 1} = 1 + \tanh(x/2).$$

- **The** $\log(1 + x^2)$ **function**. This function has applications in structured matrix factorization [66]. The function itself is obviously nonconvex and lower bounded. Its first order derivative is $\log'(1 + x^2) = \frac{2x}{1+x^2}$ and it is also bounded.

- **The quadratic function x^TQx.** Suppose that Q is a symmetric matrix but not necessarily positive semidefinite, and suppose that x^TQx is strongly convex in the null space of A^TA. Then it can be shown that there exists a δ large enough such that [A2] is true; see e.g., [144, 14].

Other relevant functions include $\sin(x)$, $\mathrm{sinc}(x)$, $\cos(x)$ and so on.

The Analysis Steps

Below we provide the analysis of Prox-PDA. First we provide a bound on the size of the constraint violation using a quantity related to the primal iterates. Let $\sigma_{\min}$ denotes the smallest *non-zero* eigenvalue of A^TA, and we define $w^r := (x^{r+1}-x^r)-(x^r-x^{r-1})$ for notational simplicity. We have the following result.

Lemma 1 *Suppose Assumptions [A1] and [A2] are satisfied. Then the following is true for Prox-PDA.*

$$\frac{1}{\beta}\|\mu^{r+1}-\mu^r\|^2 \le \frac{2L^2}{\beta\sigma_{\min}}\left\|x^r-x^{r+1}\right\|^2+\frac{2\beta}{\sigma_{\min}}\|B^TBw^r\|^2. \tag{2.8}$$

Then we bound the descent of the AL function.

Lemma 2 *Suppose Assumptions [A1] and [A2] are satisfied. Then the following is true for Algorithm 1*

$$L_\beta(x^{r+1},\mu^{r+1})-L_\beta(x^r,\mu^r) \le -\left(\frac{\beta-L}{2}-\frac{2L^2}{\beta\sigma_{\min}}\right)\|x^{r+1}-x^r\|^2+\frac{2\beta\|B^TB\|}{\sigma_{\min}}\|w^r\|^2_{B^TB}. \tag{2.9}$$

A key observation from Lemma 2 is that no matter how large β is, the rhs of (2.9) cannot be made negative. This observation suggests that the augmented Lagrangian alone cannot serve as the potential function for Prox-PDA. In search for an appropriate potential function, we need a new object that is decreasing in the order of $\beta\|w^r\|^2_{B^TB}$.

The following lemma shows that the descent of the sum of the constraint violation and the proximal term has the desired property.

Lemma 3 *Suppose Assumption [A1] is satisfied. Then the following statement is true for the constraint violation and successive difference of the variable* x^{r+1}

$$\begin{aligned}&\frac{\beta}{2}\left(\|Ax^{r+1}\|^2+\|x^{r+1}-x^r\|^2_{B^TB}\right)\\&\leq L\|x^{r+1}-x^r\|^2+\frac{\beta}{2}\left(\|Ax^r\|^2+\|x^r-x^{r-1}\|^2_{B^TB}\right)-\frac{\beta}{2}\left(\|w^r\|^2_{B^TB}+\|A(x^{r+1}-x^r)\|^2\right).\end{aligned}\tag{2.10}$$

It is interesting to observe that the new object, $\beta/2\left(\|Ax^{r+1}\|^2+\|x^{r+1}-x^r\|^2_{B^TB}\right)$, *increases* in $L\|x^{r+1}-x^r\|^2$ and *decreases* in $\beta/2\|w^r\|^2_{B^TB}$, while the AL behaves in an opposite manner (cf. Lemma 2). More importantly, in our new object, the constant in front of $\|x^{r+1}-x^r\|^2$ is *independent* of β. Although neither of these two objects decreases by itself, quite surprisingly, a proper *conic combination* of these two objects decreases at every iteration of Prox-PDA. To precisely state the claim, let us define the *potential function* for Algorithm 1 as

$$P_{c,\beta}(x^{r+1},x^r,\mu^{r+1}):=L_\beta(x^{r+1},\mu^{r+1})+\frac{c\beta}{2}\left(\|Ax^{r+1}\|^2+\|x^{r+1}-x^r\|^2_{B^TB}\right)\tag{2.11}$$

where $c>0$ is some constant to be determined later. We have the following result.

Lemma 4 *Suppose the assumptions made in Lemmas 23 – 3 are satisfied. Then we have the following*

$$\begin{aligned}P_{c,\beta}(x^{r+1},x^r,\mu^{r+1})\leq P_{c,\beta}(x^r,x^{r-1},\mu^r)&-\left(\frac{\beta-L}{2}-\frac{2L^2}{\beta\sigma_{\min}}-cL\right)\|x^{r+1}-x^r\|^2\\&-\left(\frac{c\beta}{2}-\frac{2\beta\|B^TB\|_F}{\sigma_{\min}}\right)\|w^r\|^2_{B^TB}.\end{aligned}\tag{2.12}$$

Below we derive the precise bounds for c and β. First, a sufficient condition for c is given below (note, that $\delta>0$ is defined in Assumption [A2])

$$c\geq\max\left\{\frac{\delta}{L},\frac{4\|B^TB\|_F}{\sigma_{\min}}\right\}.\tag{2.13}$$

Here the term"δ/L" in the max operator is needed for later use. Importantly, such bound on c is *independent* of β. Second, for any given c, we need β to satisfy

$$\frac{\beta-L}{2}-\frac{2L^2}{\beta\sigma_{\min}}-cL>0,$$

which further implies the following lower bond for the penalty term β, which depends on Lipschitz constant L

$$\beta > \frac{L}{2}\left(2c+1+\sqrt{(2c+1)^2+\frac{16L^2}{\sigma_{\min}}}\right). \tag{2.14}$$

Clearly combining the bounds for β and c we see that $\beta > \delta$. We conclude that if both (2.13) and (2.14) are satisfied, then the potential function $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ decreases at every iteration.

Our next step shows that by using the particular choices of c and β in (2.13) and (2.14), the constructed potential function is lower bounded.

Lemma 5 *Suppose [A1] - [A2] are satisfied, and (c, β) are chosen according to (2.13) and (2.14). Then the following statement holds true*

$$\exists\ \underline{P} > -\infty \text{ s.t. } P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) \geq \underline{P},\ \forall\ r > 0.$$

Now we are ready to present the main result of this section. To this end, define $Q(x^{r+1}, \mu^r)$ as the optimality gap of problem (5.13), given by

$$Q(x^{r+1}, \mu^r) := \|\nabla_x L_\beta(x^{r+1}, \mu^r)\|^2 + \|Ax^{r+1}\|^2. \tag{2.15}$$

It is easy to see that $Q(x^{r+1}, \mu^r) \to 0$ implies that any limit point (x^*, μ^*), if it exists, is a KKT point of (5.13) that satisfies the following conditions

$$0 = \nabla f(x^*) + A^T\mu^*, \quad Ax^* = 0. \tag{2.16}$$

In the following we show that the gap $Q(\cdot)$ not only decreases to zero, but does so in a sublinear manner.

Theorem 1 *Suppose Assumption A and the conditions (2.13) and (2.14) are satisfied. Then we have:*

- **(Eventual Consensus).** *We have*

$$\lim_{r\to\infty} \mu^{r+1} - \mu^r \to 0,\ \lim_{r\to\infty} Ax^r \to 0.$$

• **(Convergence to Stationary Points).** *Every limit point of the iterates* $\{x^r, \mu^r\}$ *generated by Algorithm 1 converges to a KKT point of problem* (5.13). *Further,* $Q(x^{r+1}, \mu^r) \to 0$.

• **(Sublinear Convergence Rate).** *For any given* $\varphi > 0$, *let us define* T *to be the first time that the optimality gap reaches below* φ, *i.e.,*

$$T := \arg\min_r Q(x^{r+1}, \mu^r) \leq \varphi.$$

Then for some $\nu > 0$, *we have* $\varphi \leq \frac{\nu}{T-1}$. *That is, the optimality gap* $Q(x^{r+1}, \mu^r)$ *converges sublinearly.*

2.4 Variants of Prox-PDA

In this section, we discuss two important extensions of the Prox-PDA, one allows the x-problem (4.12a) to be solved inexactly, while the second allows the use of increasing penalty parameter ρ.

In many practical applications, exactly minimizing the augmented Lagrangian may not be easy. Therefore, we propose the proximal gradient primal-dual algorithm (Prox-GPDA), whose main steps are given below

$$x^{r+1} = \arg\min_{x \in \mathbb{R}^Q} \langle \nabla f(x^r), x - x^r \rangle + \langle \mu^r, Ax \rangle + \frac{\beta}{2}\|Ax\|^2 + \frac{\beta}{2}\|x - x^r\|^2_{B^TB}; \tag{2.17}$$

$$\mu^{r+1} = \mu^r + \beta A x^{r+1}. \tag{2.18}$$

The analysis of this algorithm follows similar steps as that for Prox-PDA. The major difference is that there are several places in which we need to bound the term $\|\nabla f(x^{r-1}) - \nabla f(x^r)\|$ instead of $\|\nabla f(x^{r+1}) - \nabla f(x^r)\|$. Moreover, the potential function is no longer decreasing at each iteration. For detailed discussion see the supplementary material.

Our second variant do not require to explicitly compute the bound for β given in (2.14). In practice, one may prefer to start with a small penalty parameter and gradually increase it. The main steps are as bellow

$$x^{r+1} = \arg\min_{x \in \mathbb{R}^Q} f(x) + \langle \mu^r, Ax \rangle + \frac{\beta^{r+1}}{2}\|Ax\|^2 + \frac{\beta^{r+1}}{2}\|x - x^r\|^2_{B^TB}; \tag{2.19}$$

$$\mu^{r+1} = \mu^r + \beta^{r+1} A x^{r+1}. \tag{2.20}$$

Note that one can also replace $f(x)$ in (2.19) by $\langle \nabla f(x^r), x - x^r \rangle$ to obtain a similar variant for Prox-GPDA denoted by Prox-GPDA-IP. The key feature of this algorithm is that the primal proximal parameter, the primal penalty parameter, as well as the dual stepsize are all iteration-dependent. It would be challenging to achieve convergence if only a subset of these parameters grow unboundedly.

Throughout this section we will still assume that Assumption A holds true. Further, we will assume that β^r satisfies the following conditions

$$\begin{aligned} &\frac{1}{\beta^r} \to 0, \ \sum_{r=1}^{\infty} \frac{1}{\beta^r} = \infty, \ \beta^{r+1} \geq \beta^r, \\ &\max_r \left(\beta^{r+1} - \beta^r\right) < \kappa, \text{ for some finite } \ \kappa > 0. \end{aligned} \tag{2.21}$$

Also without loss of generality we will assume that

$$B^T B \succ 0, \quad \text{and} \quad \|B^T B\|_F > 1. \tag{2.22}$$

Note that this is always possible, by adding an identity matrix to $B^T B$ if necessary.

The analysis for Prox-PDA-IP is long and technical, therefore we relegate it to supplementary material. Below we provide an outline. The key step is to construct a new potential function, given below

$$P_{\beta^{r+1},c}(x^{r+1}, x^r, \mu^{r+1}) = L_{\beta^{r+1}}(x^{r+1}, \mu^{r+1}) + \frac{c\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2 + \frac{c\beta^{r+1}\beta^r}{2}\|x^r - x^{r+1}\|^2_{B^T B}.$$

The insight here is that in order to achieve the desired descent, in the potential function the coefficients for $\|x^{r+1} - x^r\|^2_{B^T B}$ and $\|Ax^{r+1}\|^2$ should be proportional to $\mathcal{O}\left((\beta^r)^2\right)$. Our proof shows that after some finite number of iterations, the newly constructed potential function starts to descend, and the size of the descent is proportional to the following quantity

$$\frac{\beta^{r+1}}{2}\|x^{r+1} - x^r\|^2 + \frac{(\beta^r)^2}{2}\|w^r\|^2. \tag{2.23}$$

Combining with the fact that the potential function is lower bounded, we can conclude that

$$\sum_{r=1}^{\infty} \frac{\beta^{r+1}}{2}\|x^{r+1} - x^r\|^2 < \infty, \sum_{r=1}^{\infty} \frac{(\beta^r)^2}{2}\|w^r\|^2 < \infty.$$

Using these two inequalities, we can show the desired convergence to the set of stationary solutions of problem (5.13).

We have the following theorem regarding to the convergence of Prox-PDA-IP.

Theorem 2 *Suppose Assumption A and (4.61) are satisfied. Suppose that B is selected such that (2.22) holds true. Then the following hold for Prox-PDA-IP*

- **(Eventual Consensus).** *We have*

$$\lim_{r\to\infty} \mu^{r+1} - \mu^r \to 0, \ \lim_{r\to\infty} Ax^r \to 0, .$$

- **(Convergence to KKT Points).** *Every limit point of the iterates $\{x^r, \mu^r\}$ generated by Prox-PDA-IP converges to a KKT point of problem (5.13). Further, $Q(x^{r+1}, \mu^r) \to 0$.*

2.5 Connections and Discussions

In this section we present an interesting observation which established links between the so-called EXTRA algorithm [122] (developed for distributed, but *convex* optimization) and the Prox-GPDA.

Specifically, the optimality condition of the x-update step (2.17) is given by

$$\nabla f(x^r) + A^T(\mu^r + \beta Ax^{r+1}) + \beta(B^TB(x^{r+1} - x^r)) = 0.$$

Utilizing the fact that $A^TA = L^-$, $B^TB = L^+$ and $L^+ + L^- = 2D$, we have

$$\nabla f(x^r) + A^T\mu^r + 2\beta Dx^{r+1} - \beta L^+x^r = 0.$$

Subtracting the same equation evaluated at the previous iteration, we obtain

$$\nabla f(x^r) - \nabla f(x^{r-1}) + \beta L^-x^r + 2\beta D(x^{r+1} - x^r) - \beta L^+(x^r - x^{r-1}) = 0,$$

where we have used the fact that $A^T(\mu^r - \mu^{r-1}) = \beta A^TAx^r = \beta L^-x^r$. Rearranging terms, we have

$$\begin{aligned} x^{r+1} &= x^r - \frac{1}{2\beta}D^{-1}\left(\nabla f(x^r) - \nabla f(x^{r-1})\right) + \frac{1}{2}D^{-1}(L^+ - L^-)x^r - \frac{1}{2}D^{-1}L^+x^{r-1} \\ &= x^r - \frac{1}{2\beta}D^{-1}\left(\nabla f(x^r) - \nabla f(x^{r-1})\right) + Wx^r - \frac{1}{2}(I+W)x^{r-1} \end{aligned} \tag{2.24}$$

where in the last equality we have defined the *weight matrix* $W := \frac{1}{2}D^{-1}(L^+ - L^-)$, which is a row stochastic matrix.

Iteration (2.24) has the same form as the EXTRA algorithm given in [122], therefore we can conclude that EXTRA is a special case of Prox-GPDA. Moreover, by appealing to our analysis in Section 2.4, it readily follows that iteration (2.24) works for the nonconvex distributed optimization problem as well, as long as the parameter β is selected appropriately.

We remark that each node i can distributedly implement iteration (2.24) by performing the following

$$x_i^{r+1} = x_i^r - \frac{1}{2\beta d_i}\left(\nabla f_i(x_i^r) - \nabla f_i(x_i^{r-1})\right) + \sum_{j\in\mathcal{N}(i)} \frac{1}{d_i}x_j^r - \frac{1}{2}\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{d_i}x_j^{r-1} + x_i^{r-1}\right) \tag{2.25}$$

Clearly, at iteration $r+1$, besides the local gradient information, node i only needs the aggregated information from its neighbors, $\sum_{j\in\mathcal{N}(i)} x_j^r$. Therefore the algorithm is distributedly implementable.

2.6 Distributed Matrix Factorization

In this section we study a variant of the Prox-PDA/Prox-PDA-IP for the following distributed matrix factorization problem [86]

$$\begin{aligned} \min_{X,Y} \quad & \frac{1}{2}\|XY - Z\|_F^2 + \eta\|X\|_F^2 + h(Y) = \sum_{i=1}^N \frac{1}{2}\|Xy_i - z_i\|^2 + \gamma\|X\|_F^2 + h_i(y_i), \\ \text{s.t.} \quad & \|y_i\|^2 \le \tau, \ \forall\ i \end{aligned} \tag{2.26}$$

where $X \in \mathbb{R}^{M\times K}$, $Y \in \mathbb{R}^{K\times N}$; for each i, $y_i \in \mathbb{R}^K$ consists of one column of Y; $Z \in \mathbb{R}^{M\times N}$ is some known matrix; $h(Y) := \sum_{i=1}^N h_i(y_i)$ is some convex but possibly nonsmooth penalization term; $\eta > 0$ is some given constant; for notation simplicity we have defined $\gamma := 1/N\eta$. It is easy to extend the above formulation to the case where Y and Z both have NP columns, and each y_i and z_i consists of P columns of Y and Z respectively. For notational simplicity, in our following discussion we only consider the vector case as given in (2.26).

We assume that $h(Y)$ is lower bounded over dom (h). One application of problem (2.26) is the distributed *sparse dictionary learning* problem where X is the dictionary to be learned, each z_i is a training data sample, and each y_i is the sparse coefficient corresponding to the particular training sample z_i. The constraint $\|y_i\|^2 \leq \tau$ simply says that the size of the coefficient must be bounded.

Consider a distributed scenario where N agents form a graph $\{\mathcal{V}, \mathcal{E}\}$, each having a column of Y. We reformulate problem (2.26) as

$$\begin{aligned} \min_{\{X_i\},\{y_i\}} \ & \sum_{i=1}^{N} \left(\frac{1}{2}\|X_i y_i - z_i\|^2 + h_i(y_i) + \gamma \|X_i\|_F^2 \right) \\ \text{s.t. } & \|y_i\|^2 \leq \tau, \ \forall \ i \quad X_i = X_j, \ \forall \ (i,j) \in \mathcal{E}. \end{aligned}$$

Let us stack all the variables X_i, and define $\boldsymbol{X} := [X_1; X_2; \cdots; X_N] \in \mathbb{R}^{NM \times K}$. Define the block signed incidence matrix as $\boldsymbol{A} = \tilde{A} \otimes I_M \in \mathbb{R}^{EM \times NM}$, where A is the standard graph incidence matrix. Define the block signless incidence matrix $\boldsymbol{B} \in \mathbb{R}^{EM \times NM}$ similarly. If the graph is connected, then the condition $\boldsymbol{AX} = \boldsymbol{0}$ implies network-wide consensus. We formulate the distributed matrix factorization problem as

$$\begin{aligned} \min_{\{X_i\},\{y_i\}} \ & f(\boldsymbol{X}, Y) + h(Y) := \sum_{i=1}^{N} \left(\frac{1}{2}\|X_i y_i - z_i\|^2 + \gamma \|X_i\|_F^2 + h_i(y_i) \right) \\ \text{s.t. } & \|y_i\|^2 \leq \tau, \ \forall \ i \quad \boldsymbol{AX} = \boldsymbol{0}. \end{aligned} \tag{2.27}$$

Clearly the above problem does not satisfy Assumption A, because the objective function is not smooth, and neither $\nabla_{\mathbf{X}} f(\mathbf{X}, Y)$ nor $\nabla_Y f(\mathbf{X}, Y)$ is Lipschitz continuous. The latter fact poses significant difficulty in algorithm development and analysis.

Define the block-signed/signless Laplacians as

$$\boldsymbol{L}^- = \boldsymbol{A}^T \boldsymbol{A}, \ \boldsymbol{L}^+ = \boldsymbol{B}^T \boldsymbol{B}. \tag{2.28}$$

The AL function for the above problem is given by

$$L_\beta(\boldsymbol{X}, Y, \boldsymbol{\Omega}) = \sum_{i=1}^{N} \left(\frac{1}{2}\|X_i y_i - z_i\|^2 + \gamma \|X_i\|_F^2 + h_i(y_i) \right) + \langle \boldsymbol{\Omega}, \boldsymbol{AX} \rangle + \frac{\beta}{2} \langle \boldsymbol{AX}, \boldsymbol{AX} \rangle, \tag{2.29}$$

where $\boldsymbol{\Omega} := \{\Omega_e\} \in \mathbb{R}^{EM\times K}$ is the matrix of the dual variable, with $\Omega_e \in \mathbb{R}^{M\times K}$ being the dual variable for the consensus constraint on link e, i.e, $X_i = X_j$, $e = (i, j)$.

Let us generalize Algorithm 1 for distributed matrix factorization given in Algorithm 2. In Al-

Algorithm 2 Prox-PDA for Distr. Matrix Factorization

1: At iteration 0, initialize $\boldsymbol{\Omega}^0 = \mathbf{0}$, and $\boldsymbol{X}^0, y^0$
2: At each iteration $r+1$, update variables by:

$$\theta_i^r = \|X_i^r y_i^r - z_i\|^2, \quad \forall\ i; \tag{2.30a}$$

$$y_i^{r+1} = \arg\min_{\|y_i\|^2\le\tau} \frac{1}{2}\|X_i^r y_i - z_i\|^2 + h_i(y_i) + \frac{\theta_i^r}{2}\|y_i - y_i^r\|^2, \quad \forall\ i; \tag{2.30b}$$

$$\boldsymbol{X}^{r+1} = \arg\min_{\boldsymbol{X}}\ f(\boldsymbol{X}, Y^{r+1}) + \langle \boldsymbol{\Omega}^r, \boldsymbol{A}\boldsymbol{X}\rangle + \frac{\beta}{2}\langle \boldsymbol{A}\boldsymbol{X}, \boldsymbol{A}\boldsymbol{X}\rangle \tag{2.30c}$$
$$+ \frac{\beta}{2}\langle \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{X}^r), \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{X}^r)\rangle;$$

$$\boldsymbol{\Omega}^{r+1} = \boldsymbol{\Omega}^r + \beta\boldsymbol{A}\boldsymbol{X}^{r+1}. \tag{2.30d}$$

gorithm 2 we have introduced a sequence $\{\theta_i^r \ge 0\}$ which measures the size of the local factorization error. We note that including the proximal term $\frac{\theta_i^r}{2}\|y_i - y_i^r\|^2$ is the key to achieve convergence for Algorithm 2. Again one should note that $\frac{\beta}{2}\langle \boldsymbol{A}\boldsymbol{X}, \boldsymbol{A}\boldsymbol{X}\rangle + \frac{\beta}{2}\langle \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{X}^r), \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{X}^r)\rangle$ is strongly convex in $\boldsymbol{X}$. Let us comment on the distributed implementation of the algorithm. First note that the y subproblem (2.30b) is naturally distributed to each node, that is, only local information is needed to perform the update. Second, the $\boldsymbol{X}$ subproblem (2.30c) can also be decomposed into N subproblems, one for each node. To be more precise, let us examine the terms in (2.30c) one by one. First, the term $f(\boldsymbol{X}, Y^{r+1}) = \sum_{i=1}^N \left(\frac{1}{2}\|X_i y_i^{r+1} - z_i\|^2 + h_i(y_i) + \gamma\|X_i\|_F^2\right)$, hence it is decomposable. Second, the term $\langle \boldsymbol{\Omega}^r, \boldsymbol{A}\boldsymbol{X}\rangle$ can be expressed as

$$\langle \boldsymbol{\Omega}^r, \boldsymbol{A}\boldsymbol{X}\rangle = \sum_{i=1}^N \sum_{e\in U(i)} \langle \Omega_e^r, X_i\rangle - \sum_{e\in H(i)} \langle \Omega_e^r, X_i\rangle$$

where the sets $U(i)$ and $H(i)$ are defined as $U(i) := \{e \mid e = (i,j) \in \mathcal{E}, i \geq j\}$ and $H(i) := \{e \mid e = (i,j) \in \mathcal{E}, j \geq i\}$. Similarly, we have

$$\langle \boldsymbol{B}\boldsymbol{X}^r, \boldsymbol{B}\boldsymbol{X}\rangle = \sum_{i=1}^{N} \left\langle X_i, d_i X_i^r + \sum_{j\in N(i)} X_j^r \right\rangle$$

$$\frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}, \boldsymbol{A}\boldsymbol{X}\rangle + \langle \boldsymbol{B}\boldsymbol{X}, \boldsymbol{B}\boldsymbol{X}\rangle\right) = \beta\langle \boldsymbol{D}\boldsymbol{X}, \boldsymbol{X}\rangle = \beta\sum_{i=1}^{N} d_i \|X_i\|_F^2$$

where $\boldsymbol{D} := \tilde{D} \otimes I_M \in \mathbb{R}^{NM\times NM}$ with $\tilde{D}$ being the degree matrix. It is easy to see that the $\boldsymbol{X}$ subproblem (2.30c) is separable over the distributed agents.

Finally, one can verify that the $\boldsymbol{\Omega}$ update step (2.30d) can be implemented by each edge $e \in \mathcal{E}$ as follows

$$\Omega_e^{r+1} = \Omega_e^r + \beta\left(X_i^{r+1} - X_j^{r+1}\right), \quad e = (i,j), i \geq j.$$

To show convergence rate of the algorithm, we need the following definition

$$Q(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r) := \beta\|\boldsymbol{A}\boldsymbol{X}^{r+1}\|^2 + \|[\boldsymbol{Z}_1^{r+1}; \boldsymbol{Z}_2^{r+1}]\|^2,$$

where we have defined

$$\boldsymbol{Z}_1^{r+1} := \nabla_{\boldsymbol{X}} L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r);$$

$$\boldsymbol{Z}_2^{r+1} := Y^{r+1} - \mathrm{prox}_{h+\iota(\mathcal{Y})}\left[Y^{r+1} - \nabla_Y\left(L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r) - h(Y)\right)\right].$$

In the above expression, the prox operator for a convex lower semi-continuous function $p(\cdot)$ is given by

$$\mathrm{prox}_p(c) = \arg\min_z \; p(z) + \frac{1}{2}\|z - c\|^2. \tag{2.31}$$

We have also used $\mathcal{Y} := \bigcup_i \left\{\|y_i\|^2 \leq \tau\right\}$ to denote the feasible set of Y, and used $\iota(\mathcal{Y})$ to denote the indicator function of such set. Similarly as in Section 5.2.3, we can show that $Q(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r) \to 0$ implies that every limit point of $(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r)$ is a KKT point of problem (2.27).

Next we present the main convergence analysis for Algorithm 2. The proof is long and technical, therefore we relegate it to supplementary material.

Theorem 3 *Consider using Algorithm 2 to solve the distributed matrix factorization problem (2.27). Suppose that $h(Y)$ is lower bounded over dom $h(x)$, and that the penalty parameter β, together with two positive constants c and d, satisfies the following conditions*

$$\begin{aligned} &\frac{\beta+2\gamma}{2}-\frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}}-\frac{cd}{2}>0,\\ &\frac{1}{2}-\frac{8}{\sigma_{\min}\beta}-\frac{c}{d}>0,\ \frac{1}{2}-\frac{8\tau}{\sigma_{\min}\beta}-\frac{c\tau}{d}>0,\\ &\frac{c\beta}{2}-\frac{2\beta\|B^TB\|}{\sigma_{\min}}>0. \end{aligned} \tag{2.32}$$

Then in the limit, consensus will be achieved, i.e.,

$$\lim_{r\to\infty}\|X_i^r-X_j^r\|=0,\quad \forall\ (i,j)\in\mathcal{E}.$$

Further, the sequences $\{\boldsymbol{X}^{r+1}\}$ and $\{\boldsymbol{\Omega}^{r+1}\}$ are both bounded, and every limit point generated by Algorithm 2 is a KKT point of problem (2.26).

Additionally, Algorithm 2 converges sublinearly. Specifically, for any given $\varphi>0$, define T to be the first time that the optimality gap reaches below φ, i.e.,

$$T:=\arg\min_r\ Q(\boldsymbol{X}^{r+1},Y^{r+1},\boldsymbol{\Omega}^r)\le\varphi.$$

Then for some constant $\nu>0$ we have $\varphi\le\frac{\nu}{T-1}$.

We can see that it is always possible to find the tuple $\{\beta,c,d>0\}$ that satisfies (2.32): c can be solely determined by the last inequality; for fixed c, the constant d needs to be chosen large enough such that $1/2-\frac{c}{d}>0$ and $1/2-\frac{c\tau}{d}>0$ are satisfied. After c and d are fixed, one can always choose β large enough to satisfy the first three conditions. In practice, we typically prefer to choose β as small as possible to improve the convergence speed. Therefore empirically one can start with (for some small $\nu>0$): $c=\frac{4\|B^TB\|}{\sigma_{\min}}+\nu,\quad d=\max\{4,2c\tau\}$, and then gradually increase d to find an appropriate β that satisfies the first three conditions.

We remark that Algorithm 2 can be extended to the case with increasing penalty. Due to the space limitation we omit the details here.

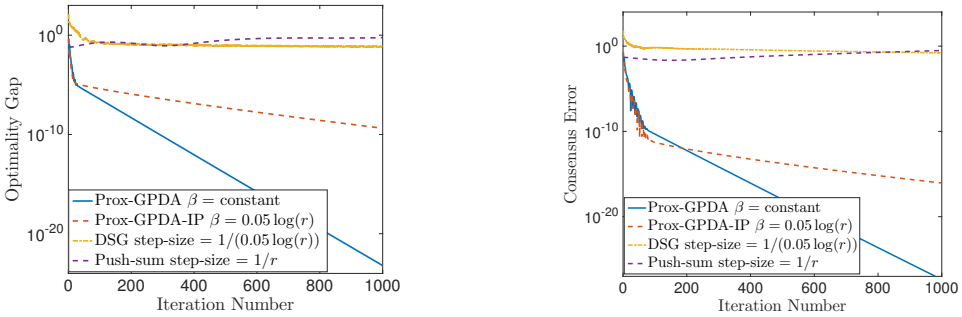


Figure 2.3: Results for the matrix factorization problem.

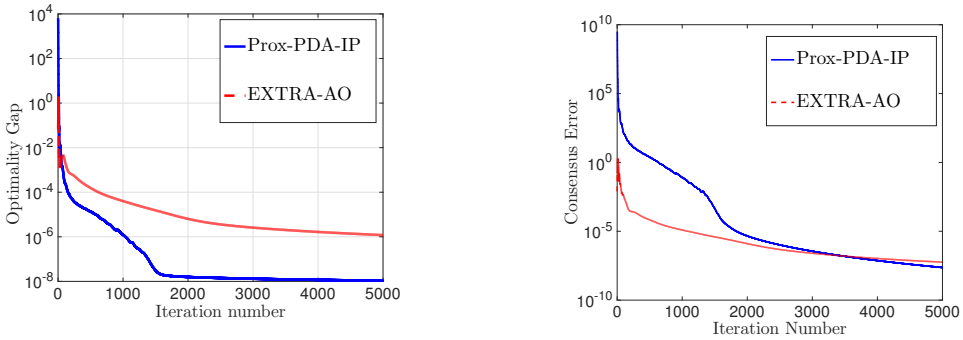


Figure 2.4: Results for the matrix factorization problem.

2.7 Numerical Results

In this section, we demonstrate the performance of the proposed algorithms. All experiments are performed in Matlab (2016b) on a laptop with an Intel Core(TM) i5-4690 CPU (3.50 GHz) and 8GB RAM running Windows 7.

2.7.1 Distributed Binary Classification

In this subsection, we study the problem of binary classification using nonconvex regularizers in the mini-bach setup i.e. each node stores b (batch size) data points, and each component function is given by

$$f_i(x_i) = \frac{1}{Nb}\left[\sum_{j=1}^{b}\log(1+\exp(-y_j x_i^T v_j)) + \sum_{k=1}^{M}\frac{\lambda\alpha x_{i,k}^2}{1+\alpha x_{i,k}^2}\right]$$

where $v_i \in \mathbb{R}^M$ and $y_i \in \{1, -1\}$ are the feature vector and the label for the ith date point [7]. We use the parameter settings of $\lambda = 0.001$, $\alpha = 1$ and $M = 10$. We randomly generated $100,000$ data points and distribute them into $N = 20$ nodes (i.e. $b = 5000$). We use the optimality gap (opt-gap) and constraint violation (con-vio), displayed below, to measure the quality of the solution generated by different algorithms:

$$\text{opt-gap} := \left\| \sum_{i=1}^{N} \nabla f_i(z_i) \right\|^2 + \|Ax\|^2, \ \text{con-vio} = \|Ax\|^2.$$

We compare the the Prox-GPDA, and Prox-GPDA-IP with the distributed subgradient (DSG) method [100] (which is only known to work for convex cases) and the Push-sum algorithm [131]. The performance of all three algorithms in terms of the consensus error and the optimality gap (averaged over 30 problem instances) are presented in Fig. 2.3. The penalty parameter for Prox-GPDA is chosen such that satisfies (2.14), and β^r for Prox-GPDA-IP is set as $0.05\log(r)$, the stepsizes of the DSG algorithm and the Push-sum algorithm are chosen as $1/0.05\log(r)$ and $1/r$, respectively. Note that these parameters are tuned for each algorithm to achieve the best results. It can be observed that the Prox-GPDA with constant stepsize outperforms other algorithms. The Push-sum algorithm does not seem to converge within 1000 iterations.

2.7.2 Distributed Matrix Factorization

In this section we consider the distributed matrix factorization problem (2.26). The training data is constructed by randomly extracting 300 overlapping patches from the 512×512 image of *barbara.png*, each with size 16×16 pixels. Each of the extracted patch is vectorized, resulting a training data set Z of size 256×300. We consider a network of $N = 10$ agents, and the columns of Z are evenly distributed among the agents (each having $P = 30$ columns). We compare Prox-PDA-IP (a variant of Prox-PDA with increasing stepsize) with the EXTRA-AO algorithm proposed in [52]. Note that the EXTRA-AO is also designed for a similar distributed matrix factorization problem and it works well in practice. However, it does not have formal convergence proof. We initialize both algorithms with X being the 2D discrete cosine transform (DCT) matrix. We set $\gamma = 0.05$,

$\tau = 10^5$ and $\beta = 0.001r$, and the results are averaged over 10 problem instances. The stepsizes of the EXTRA-AO is set as $\alpha_{\text{AO}} = 0.03$ and $\beta_{\text{AO}} = 0.002$.

In Fig. 2.4, we compare the performance of the proposed Prox-PDA-IP and the EXTRA-AO versus the number of iterations. It can be observed that our proposed algorithm converges faster than the EXTRA-AO. We have observed that the EXTRA-AO does have reasonably good practical performance, however it lacks formal convergence proof.

2.8 Appendix. Lemma proofs

2.8.1 Proof of Lemma 23

From the optimality condition of the x problem (4.12a) we have

$$\nabla f(x^{r+1}) + A^T(\mu^r + \beta A x^{r+1}) + \beta B^T B(x^{r+1} - x^r) = 0.$$

Applying (2.7b), we have

$$A^T \mu^{r+1} = -\nabla f(x^{r+1}) - \beta B^T B(x^{r+1} - x^r). \tag{2.33}$$

From equation (2.7b) ($\mu^{r+1} = \mu^r + \beta A^T \mu^r$) it is clear the difference of the dual variables lies in the column space of A. Therefore the following is true

$$\sigma_{\min}^{1/2} \|\mu^{r+1} - \mu^r\| \leq \|A^T(\mu^{r+1} - \mu^r)\|.$$

This inequality combined with (2.33) implies that

$$\begin{aligned}\|\mu^{r+1} - \mu^r\| &\leq \frac{1}{\sigma_{\min}^{1/2}} \| - \nabla f(x^{r+1}) - \beta B^T B(x^{r+1} - x^r) - (-\nabla f(x^r) - \beta B^T B(x^r - x^{r-1}))\| \\ &= \frac{1}{\sigma_{\min}^{1/2}} \left\|\nabla f(x^r) - \nabla f(x^{r+1}) - \beta B^T B w^r\right\|.\end{aligned}$$

Squaring both sides and dividing by β, we obtain the desired result. **Q.E.D.**

2.8.2 Proof of Lemma 2

Since $f(x)$ has Lipschitz continuous gradient, and that $A^T A + B^T B \succeq I$ by Assumption [A1], it is known that if $\beta > L$, then the x-problem (4.12a) is strongly convex with modulus $\gamma := \beta - L > 0$;

See [150] [Theorem 2.1]. That is, we have

$$L_\beta(x,\mu^r)+\frac{\beta}{2}\|x-x^r\|^2_{B^TB}-(L_\beta(z,\mu^r)+\frac{\beta}{2}\|z-x^r\|^2_{B^TB})$$
$$\geq \langle \nabla_x L_\beta(z,\mu^r)+\beta(B^TB(z-x^r)),x-z\rangle+\frac{\gamma}{2}\|x-z\|^2,\ \forall\ x,z\in\mathbb{R}^N,\ \forall\ \mu^r. \tag{2.34}$$

Using this property, we have

$$\begin{aligned}
&L_\beta(x^{r+1},\mu^{r+1})-L_\beta(x^r,\mu^r)\\
&=L_\beta(x^{r+1},\mu^{r+1})-L_\beta(x^{r+1},\mu^r)+L_\beta(x^{r+1},\mu^r)-L_\beta(x^r,\mu^r)\\
&\leq L_\beta(x^{r+1},\mu^{r+1})-L_\beta(x^{r+1},\mu^r)+L_\beta(x^{r+1},\mu^r)+\frac{\beta}{2}\|x^{r+1}-x^r\|^2_{B^TB}-L_\beta(x^r,\mu^r)\\
&\overset{(i)}{\leq}\frac{\|\mu^{r+1}-\mu^r\|^2}{\beta}+\langle\nabla_x L_\beta(x^{r+1},\mu^r)+\beta(B^TB(x^{r+1}-x^r)),x^{r+1}-x^r\rangle-\frac{\gamma}{2}\|x^{r+1}-x^r\|^2\\
&\overset{(ii)}{\leq}\frac{\|\mu^{r+1}-\mu^r\|^2}{\beta}-\frac{\gamma}{2}\|x^{r+1}-x^r\|^2\\
&\leq\frac{1}{\sigma_{\min}}\left(\frac{2L^2}{\beta}\left\|x^r-x^{r+1}\right\|^2+2\beta\left\|B^TBw^r\right\|^2\right)-\frac{\gamma}{2}\|x^{r+1}-x^r\|^2\\
&=-\left(\frac{\beta-L}{2}-\frac{2L^2}{\beta\sigma_{\min}}\right)\|x^{r+1}-x^r\|^2+\frac{2\beta}{\sigma_{\min}}\left\|B^TBw^r\right\|^2
\end{aligned}\tag{2.35}$$

where in (i) we have used (2.34) with the identification $z=x^{r+1}$ and $x=x^r$ and the fact that

$$L_\beta(x^{r+1},\mu^{r+1})-L_\beta(x^{r+1},\mu^r)=\langle\mu^{r+1}-\mu^r,Ax^{r+1}\rangle=\frac{1}{\beta}\|\mu^{r+1}-\mu^r\|^2$$

; in (ii) we have used the optimality condition for the x-subproblem (4.12a). The claim is proved. **Q.E.D.**

2.8.3 Proof of Lemma 3

From the optimality condition of the x-subproblem (4.12a) we have

$$\langle\nabla f(x^{r+1})+A^T\mu^r+\beta A^TAx^{r+1}+\beta B^TB(x^{r+1}-x^r),x^{r+1}-x\rangle\leq 0,\ \forall\ x\in\mathbb{R}^Q.$$

If we shift r to $r-1$, we get

$$\langle\nabla f(x^r)+A^T\mu^{r-1}+\beta A^TAx^r+\beta B^TB(x^r-x^{r-1}),x^r-x\rangle\leq 0,\ \forall\ x\in\mathbb{R}^Q.$$

Plugging $x = x^r$ into the first inequality and $x = x^{r+1}$ into the second, adding the resulting inequalities and utilizing the μ-update step (2.7b) we obtain

$$\langle \nabla f(x^{r+1}) - \nabla f(x^r) + A^T(\mu^{r+1} - \mu^r) + \beta B^T B w^r, x^{r+1} - x^r \rangle \leq 0.$$

Rearranging, we have

$$\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r \rangle \leq -\langle \nabla f(x^{r+1}) - \nabla f(x^r) + \beta B^T B w^r, x^{r+1} - x^r \rangle. \tag{2.36}$$

Let us bound the lhs and the rhs of (2.36) separately.

First the lhs of (2.36) can be expressed as

$$\begin{aligned}
\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r \rangle &= \langle \beta A^T A x^{r+1}, x^{r+1} - x^r \rangle \\
&= \langle \beta A x^{r+1}, A x^{r+1} - A x^r \rangle \\
&= \beta \|A x^{r+1}\|^2 - \beta \langle A x^{r+1}, A x^r \rangle \\
&= \frac{\beta}{2} \left(\|A x^{r+1}\|^2 - \|A x^r\|^2 + \|A(x^{r+1} - x^r)\|^2 \right). \qquad (2.37)
\end{aligned}$$

Second we have the following bound for the rhs of (2.36)

$$\begin{aligned}
& - \langle \nabla f(x^{r+1}) - \nabla f(x^r) + \beta B^T B w^r, x^{r+1} - x^r \rangle \\
& \leq L \|x^{r+1} - x^r\|^2 - \beta \langle B^T B w^r, x^{r+1} - x^r \rangle \\
& = L \|x^{r+1} - x^r\|^2 + \frac{\beta}{2} \left(\|x^r - x^{r-1}\|^2_{B^T B} - \|x^{r+1} - x^r\|^2_{B^T B} - \|w^r\|^2_{B^T B} \right). \qquad (2.38)
\end{aligned}$$

Combining the above two bounds, we have

$$\begin{aligned}
\frac{\beta}{2} \left(\|A x^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^T B} \right) \leq L \|x^{r+1} - x^r\|^2 + \frac{\beta}{2} \left(\|x^r - x^{r-1}\|^2_{B^T B} + \|A x^r\|^2 \right) \\
- \frac{\beta}{2} \left(\|w^r\|^2_{B^T B} + \|A(x^{r+1} - x^r)\|^2 \right).
\end{aligned}$$

The desired claim is proved. **Q.E.D.**

2.8.4 Proof of Lemma 4

Multiplying both sides of (2.10) by the constant c and then add them to (2.9), we obtain

$$
\begin{aligned}
&L_\beta(x^{r+1},\mu^{r+1}) + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^TB}\right)\\
&\le L_\beta(x^r,\mu^r) + cL\|x^{r+1}-x^r\|^2 + \frac{c\beta}{2}\left(\|x^r-x^{r-1}\|^2_{B^TB} + \|Ax^r\|^2\right)\\
&-\left(\frac{\beta-L}{2} - \frac{2L^2}{\beta\sigma_{\min}}\right)\|x^{r+1}-x^r\|^2 + \frac{2\beta}{\sigma_{\min}}\left\|B^TBw^r\right\|^2 - \frac{c\beta}{2}\left(\|w^r\|^2_{B^TB} + \|A(x^{r+1}-x^r)\|^2\right)\\
&\le L_\beta(x^r,\mu^r) + \frac{c\beta}{2}\left(\|x^r-x^{r-1}\|^2_{B^TB} + \|Ax^r\|^2\right)\\
&\quad-\left(\frac{\beta-L}{2} - \frac{2L^2}{\beta\sigma_{\min}} - cL\right)\|x^{r+1}-x^r\|^2 - \left(\frac{c\beta}{2} - \frac{2\beta\|B^TB\|_F}{\sigma_{\min}}\right)\|w^r\|^2_{B^TB}.
\end{aligned}
$$

The desired result is proved. **Q.E.D.**

2.8.5 Proof of Lemma 25

To prove this we need to utilize the boundedness assumption in [A2].

First, we can express the augmented Lagrangian function as following

$$
\begin{aligned}
L_\beta(x^{r+1},\mu^{r+1}) &= f(x^{r+1}) + \langle\mu^{r+1},Ax^{r+1}\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2\\
&= f(x^{r+1}) + \frac{1}{\beta}\langle\mu^{r+1},\mu^{r+1}-\mu^r\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2\\
&= f(x^{r+1}) + \frac{1}{2\beta}\left(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1}-\mu^r\|^2\right) + \frac{\beta}{2}\|Ax^{r+1}\|^2.
\end{aligned}
$$

Therefore, summing over $r = 1\cdots,T$, we obtain

$$
\sum_{r=1}^{T} L_\beta(x^{r+1},\mu^{r+1}) = \sum_{r=1}^{T}\left(f(x^{r+1}) + \frac{\beta}{2}\|Ax^{r+1}\|^2 + \frac{1}{2\beta}\|\mu^{r+1}-\mu^r\|^2\right) + \frac{1}{2\beta}\left(\|\mu^{T+1}\|^2 - \|\mu^1\|^2\right).
$$

Suppose Assumption [A2] is satisfied and β is chosen according to (2.13) and (2.14), then clearly the above sum is lower bounded since

$$
f(x) + \frac{\beta}{2}\|Ax\|^2 \ge f(x) + \frac{\delta}{2}\|Ax\|^2 \ge 0,\ \forall\ x\in\mathbb{R}^Q.
$$

This fact implies that the sum of the potential function is also lower bounded (note, the remaining terms in the potential function are all nonnegative), that is

$$
\sum_{r=1}^{T} P_{c,\beta}(x^{r+1},x^r,\mu^{r+1}) > -\infty,\quad \forall\ T>0.
$$

Note that if c and β are chosen according to (2.13) and (2.14), then $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ is non-increasing. Combined with the lower boundedness of the sum of the potential function, we can conclude that the following is true

$$P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) > -\infty, \quad \forall\ r > 0. \tag{2.39}$$

This completes the proof. **Q.E.D.**

2.8.6 Proof of Theorm 1

First we prove part (1). Combining Lemmas 4 and 25, we conclude that $\|x^{r+1} - x^r\|^2 \to 0$. Then according to (5.32), in the limit we have $\mu^{r+1} \to \mu^r$, or equivalently $Ax^r \to 0$. That is, the constraint violation will be satisfied in the limit.

Then we prove part (2). From the optimality condition of x-update step (4.12a) we have

$$\nabla f(x^{r+1}) + A^T\mu^r + \beta A^T(Ax^{r+1}) + \beta B^T B(x^{r+1} - x^r) = 0.$$

Then we argue that $\{\mu^r\}$ is a bounded sequence if $\nabla f(x^{r+1})$ is bounded. Indeed the fact that $\|x^{r+1} - x^r\|^2 \to 0$ and $Ax^{r+1} \to 0$ imply that both $(x^{r+1} - x^r)$ and Ax^{r+1} are bounded. Then the boundedness of μ^r follows from the assumption that $\nabla f(x)$ is bounded for any $x \in \mathbb{R}^Q$, and that μ^r lies in the column space of A.

Then we argue that $\{x^r\}$ is bounded if $f(x) + \frac{\beta}{2}\|Ax\|^2$ is coercive. Note that the potential function can be expressed as

$$\begin{aligned} P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) &= f(x^{r+1}) + \langle \mu^{r+1}, Ax^{r+1}\rangle + \frac{\beta}{2}\|Ax^{r+1}\|^2 + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^TB}\right) \\ &= f(x^{r+1}) + \frac{1}{2\beta}(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1} - \mu^r\|^2) + \frac{\beta}{2}\|Ax^{r+1}\|^2 \\ &\quad + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^TB}\right) \end{aligned}$$

and by our analysis in Lemma 25 we know that it is decreasing thus *upper bounded*. Suppose that $\{x^r\}$ is unbounded and let $\mathcal{K}$ denote an infinite subset of iteration index in which $\lim_{r\in\mathcal{K}} x^r = \infty$. Passing limit to $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ over $\mathcal{K}$, and using the fact that $x^{r+1} \to x^r$, $\mu^{r+1} \to \mu^r$, we have

$$\lim_{r\in\mathcal{K}} P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) = \lim_{r\in\mathcal{K}} f(x^{r+1}) + \frac{c\beta + \beta}{2}\|Ax^{r+1}\| = \infty$$

where the last equality comes from the coerciveness assumption. This is a contradiction to the fact that the potential function $P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1})$ is upper bounded. This concludes the proof for the second part of the result.

Then we prove part (3). Let $\mathcal{K}$ denote any converging infinite iteration index such that $\{(\mu^r, x^r)\}_{r\in\mathcal{K}}$ converges to the limit point (μ^*, x^*). Passing limit in $\mathcal{K}$, and using the fact that $\|x^{r+1} - x^r\| \to 0$, we have

$$\nabla f(x^*) + A^T\mu^* + \beta A^T A x^* = 0.$$

Combined with the fact that $Ax^* = 0$, we conclude that (μ^*, x^*) is indeed a stationary point of the original problem (5.13), satisfying (5.15).

Additionally, even if the sequence $\{x^{r+1}, \mu^{r+1}\}$ does not have a limit point, from part (1) we still have $\|\mu^{r+1} - \mu^r\| \to 0$ and $\|x^r - x^{r+1}\| \to 0$. Hence

$$\lim_{r\to\infty} \nabla_x L_\beta(x^{r+1}, \mu^r) = \lim_{r\to\infty} \nabla f(x^{r+1}) \overset{\text{(i)}}{=} \lim_{r\to\infty} -\beta B^T B(x^{r+1} - x^r) = 0$$

where (i) is from the optimality condition of the x-subproblem (4.12a). Therefore we have $Q(x^{r+1}, \mu^r) \to 0$.

Finally we prove part (4). Our first step is to bound the size of the gradient of the augmented Lagrangian. From the optimality condition of the x-problem (4.12a), we have

$$\begin{aligned}
\|\nabla_x L_\beta(x^r, \mu^{r-1})\|^2 &= \|\nabla_x L_\beta(x^{r+1}, \mu^r) + \beta B^T B(x^{r+1} - x^r) - \nabla_x L_\beta(x^r, \mu^{r-1})\|^2 \\
&= \|\nabla f(x^{r+1}) - \nabla f(x^r) + A^T(\mu^{r+1} - \mu^r) + \beta B^T B(x^{r+1} - x^r)\|^2 \\
&\le 3L^2\|x^{r+1} - x^r\|^2 + 3\|\mu^{r+1} - \mu^r\|^2\|A^T A\| + 3\beta^2\|B^T B(x^{r+1} - x^r)\|^2.
\end{aligned}$$

By utilizing the estimate (5.32), we see that there must exist a constant $\xi > 0$ such that the following is true

$$Q(x^r, \mu^{r-1}) = \|\nabla_x L_\beta(x^r, \mu^{r-1})\|^2 + \beta\|Ax^r\|^2 \le \xi \left\|x^r - x^{r+1}\right\|^2 + \xi \left\|B^T B w^r\right\|^2.$$

From the descent estimate (2.9) we see that there must exist a constant $\nu > 0$ such that

$$P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) - P_{c,\beta}(x^r, x^{r-1}, \mu^r) \le -\nu\|x^{r+1} - x^r\|^2 - \nu \left\|B^T B w^r\right\|^2.$$

Matching the above two bounds, we have

$$Q(x^r, \mu^{r-1}) \leq \frac{\nu}{\xi} \left(P_{c,\beta}(x^r, x^{r-1}, \mu^r) - P_{c,\beta}(x^{r+1}, x^r, \mu^{r+1}) \right).$$

Summing over r, and let T denote the first time that $Q(x^r, \mu^{r-1})$ reaches below φ, we obtain

$$\begin{aligned} \varphi \leq \frac{1}{T-1} \sum_{r=1}^{T-1} Q(x^r, \mu^{r-1}) &\leq \frac{1}{T-1} \frac{\nu}{\xi} \left(P_{c,\beta}(x^1, x^0, \mu^1) - P_{c,\beta}(x^T, x^{T-1}, \mu^T) \right) \\ &\leq \frac{1}{T-1} \frac{\nu}{\xi} \left(P_{c,\beta}(x^1, x^0, \mu^1) - \underline{P} \right) := \frac{\nu}{T-1}. \end{aligned}$$

We conclude that the convergence in term of the optimality gap function $Q(x^{r+1}, \mu^r)$ is sublinear. **Q.E.D.**

2.8.7 The Analysis Outline for Prox-GPDA

First, following the derivation leading to (5.32) we obtain

$$\frac{1}{\beta} \|\mu^{r+1} - \mu^r\|^2 \leq \frac{2L^2}{\beta \sigma_{\min}} \left\| x^r - x^{r-1} \right\|^2 + \frac{2\beta}{\sigma_{\min}} \left\| B^T B w^r \right\|^2. \tag{2.40}$$

Note that the first term is now related to the square of the difference between the *previous* two iterations.

Following the proof steps in Lemma 2, the descent of the augmented Lagrangian is given by

$$\begin{aligned} &L_\beta(x^{r+1}, \mu^{r+1}) - L_\beta(x^r, \mu^r) \\ &\leq -\frac{\beta - L}{2} \|x^{r+1} - x^r\|^2 + \frac{2\beta}{\sigma_{\min}} \left\| B^T B w^r \right\|^2 + \frac{2L^2}{\beta \sigma_{\min}} \left\| x^r - x^{r-1} \right\|^2. \end{aligned} \tag{2.41}$$

In the third step we have the following estimate

$$\begin{aligned} &\frac{\beta}{2} \left(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|^2_{B^T B} \right) \\ &\leq \frac{L}{2} \|x^{r-1} - x^r\|^2 + \frac{L}{2} \|x^{r+1} - x^r\|^2 + \frac{\beta}{2} \left(\|x^r - x^{r-1}\|^2_{B^T B} + \|Ax^r\|^2 \right) \\ &\quad - \frac{\beta}{2} \left(\|w^r\|^2_{B^T B} + \|A(x^{r+1} - x^r)\|^2 \right). \end{aligned} \tag{2.42}$$

Note that the first two terms come from the following estimate

$$\begin{aligned} -\langle x^{r+1} - x^r, \nabla f(x^r) - \nabla f(x^{r-1}) \rangle &\leq \frac{L}{2} \|x^{r+1} - x^r\|^2 + \frac{1}{2L} \|\nabla f(x^r) - \nabla f(x^{r-1})\|^2 \\ &\leq \frac{L}{2} \|x^{r+1} - x^r\|^2 + \frac{L}{2} \|x^r - x^{r-1}\|^2, \end{aligned}$$

where the first inequality is the application of Young's inequality.

In the fourth step we have the following overall descent estimate

$$\begin{aligned}
&L_\beta(x^{r+1},\mu^{r+1}) + \frac{c\beta}{2}\left(\|Ax^{r+1}\|^2 + \|x^{r+1}-x^r\|^2_{B^TB}\right)\\
&\le L_\beta(x^r,\mu^r) + \frac{c\beta}{2}\left(\|x^r-x^{r-1}\|^2_{B^TB} + \|Ax^r\|^2\right) - \left(\frac{\beta-L}{2} - \frac{cL}{2}\right)\|x^{r+1}-x^r\|^2\\
&\quad + \left(\frac{2L^2}{\beta\sigma_{\min}} + \frac{cL}{2}\right)\|x^{r-1}-x^r\|^2 - \left(\frac{c\beta}{2} - \frac{2\beta\|B^TB\|}{\sigma_{\min}}\right)\|w^r\|^2_{B^TB}\,. \qquad (2.43)
\end{aligned}$$

Note that there is a slight difference between this descent estimate and our previous estimate (4.28), because now there is a positive term in the rhs, which involves $\|x^r - x^{r-1}\|^2$. Therefore the potential function is difficult to decrease by itself. Fortunately, such extra term can be bounded by the descent of the *previous* iteration. We can take the summation over all the iterations and obtain

$$\begin{aligned}
&L_\beta(x^{T+1},\mu^{T+1}) + \frac{c\beta}{2}\left(\|Ax^{T+1}\|^2 + \|x^{T+1}-x^T\|^2_{B^TB}\right)\\
&\le L_\beta(x^1,\mu^1) + \frac{c\beta}{2}\left(\|x^1-x^0\|^2_{B^TB} + \|Ax^1\|^2\right) + \left(\frac{2L^2}{\beta\sigma_{\min}} + cL\right)\|x^0-x^1\|^2\\
&\quad - \sum_{r=1}^{T-1}\left(\frac{\beta-L}{2} - \frac{2L^2}{\beta\sigma_{\min}} - cL\right)\|x^{r+1}-x^r\|^2 - \sum_{r=1}^{T}\left(\frac{c\beta}{2} - \frac{2\beta\|B^TB\|}{\sigma_{\min}}\right)\|w^r\|^2_{B^TB}\,.
\end{aligned}$$

Clearly as long as the potential function is lower bounded, we have $x^{r+1}\to x^r$ and $x^{r+1}-x^r \to x^r - x^{r-1}$. The rest of the proof follows similar steps leading to Theorem 1, hence is omitted.

2.8.8 Proof of Convergence for Prox-PDA-IP

In this part we present the convergence analysis for Prox-PDA-IP algorithm which main steps are given in (2.19) and (2.20). Our analysis consists of a series of steps.

Step 1. Our first step is again to bound the size of the successive difference of $\{\mu^r\}$. To this end, write down the optimality condition for the x-update (2.19) as

$$A^T\mu^{r+1} = -\nabla f(x^{r+1}) - \beta^{r+1}B^TB(x^{r+1}-x^r). \qquad (2.44)$$

Subtracting the previous iteration, we obtain

$$A^T(\mu^{r+1}-\mu^r) = -(\nabla f(x^{r+1}) - \nabla f(x^r)) - \beta^r B^TB\,(w^r) - (\beta^{r+1}-\beta^r)B^TB(x^{r+1}-x^r). \qquad (2.45)$$

Therefore, using the fact that $\mu^{r+1}-\mu^r \in \mathrm{col}(A)$, we have

$$\begin{aligned}&\frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2\\&\leq \frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^r)^2\|B^TB\|\right)\|x^{r+1}-x^r\|^2+\frac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}}\left\|B^TB\left(w^r\right)\right\|^2.\end{aligned}\tag{2.46}$$

Also from the optimality condition we have the following relation

$$x^{r+1}=x^r-\frac{1}{\beta^{r+1}}(B^TB)^{-1}\left(\nabla f(x^{r+1})+A^T\mu^{r+1}\right):=x^r-\frac{1}{\beta^{r+1}}v^{r+1},\tag{2.47}$$

where we have defined the primal update direction v^{r+1} as

$$v^{r+1}=(B^TB)^{-1}\left(\nabla f(x^{r+1})+A^T\mu^{r+1}\right).$$

Step 2. In the second step we analyze the descent of the augmented Lagrangian. We have the following estimate

$$\begin{aligned}&L_{\beta^{r+1}}(x^{r+1},\mu^{r+1})-L_{\beta^r}(x^r,\mu^r)\\&=L_{\beta^{r+1}}(x^{r+1},\mu^{r+1})-L_{\beta^{r+1}}(x^{r+1},\mu^r)+L_{\beta^{r+1}}(x^{r+1},\mu^r)-L_{\beta^{r+1}}(x^r,\mu^r)+L_{\beta^{r+1}}(x^r,\mu^r)-L_{\beta^r}(x^r,\mu^r)\\&\overset{\text{(i)}}{\leq}\frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2+\frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2-\frac{\beta^{r+1}-L}{2}\|x^{r+1}-x^r\|^2\\&\overset{\text{(ii)}}{\leq}-\left(\frac{\beta^{r+1}-L}{2}-\frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^r)^2\|B^TB\|\right)\right)\|x^{r+1}-x^r\|^2+\frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2\\&\quad+\frac{3(\beta^r)^2}{\beta^{r+1}\sigma_{\min}}\left\|B^TB\left(w^r\right)\right\|^2\end{aligned}\tag{2.48}$$

where in (i) we have used the optimality of the x-subproblem (cf. the derivation in (5.66)), and the fact that

$$L_{\beta^{r+1}}(x^r,\mu^r)-L_{\beta^r}(x^r,\mu^r)=\frac{\beta^{r+1}-\beta^r}{2}\|Ax^r\|^2=\frac{\beta^{r+1}-\beta^r}{2(\beta^r)^2}\|\mu^r-\mu^{r-1}\|^2;\tag{2.49}$$

in (ii) we have applied (2.46).

Step 3. In the third step, we construct the remaining part of the potential function. We have the following two inequalities from the optimality condition of the x-update (2.19)

$$\left\langle\nabla f(x^{r+1})+A^T\mu^{r+1}+\beta^{r+1}B^TB(x^{r+1}-x^r),x^{r+1}-x\right\rangle\leq 0,\ \forall\ x\in\mathbb{R}^Q$$

$$\left\langle\nabla f(x^r)+A^T\mu^r+\beta^rB^TB(x^r-x^{r-1}),x^r-x\right\rangle\leq 0,\ \forall\ x\in\mathbb{R}^Q.$$

Plugging $x = x^r$ and $x = x^{r+1}$ to these two equations and adding them together, we obtain

$$\begin{aligned}
&\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r\rangle \\
&\leq -\langle \nabla f(x^{r+1}) - \nabla f(x^r), x^{r+1} - x^r\rangle - \langle B^T B(\beta^{r+1}(x^{r+1} - x^r) - \beta^r(x^r - x^{r-1})), x^{r+1} - x^r\rangle.
\end{aligned}$$

The lhs of the above inequality can be expressed as

$$\begin{aligned}
&\langle A^T(\mu^{r+1} - \mu^r), x^{r+1} - x^r\rangle \\
&= \frac{\beta^{r+1}}{2}\left(\|Ax^{r+1}\|^2 - \|Ax^r\|^2 + \|A(x^{r+1} - x^r)\|^2\right) \\
&= \frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 - \frac{\beta^r}{2}\|Ax^r\|^2 + \frac{\beta^{r+1}}{2}\|A(x^{r+1} - x^r)\|^2 + \frac{\beta^r - \beta^{r+1}}{2}\|Ax^r\|^2,
\end{aligned}$$

while its rhs can be bounded as

$$\begin{aligned}
&- \langle \nabla f(x^{r+1}) - \nabla f(x^r), x^{r+1} - x^r\rangle - \langle B^T B(\beta^{r+1}(x^{r+1} - x^r) - \beta^r(x^r - x^{r-1})), x^{r+1} - x^r\rangle \\
&\leq L\|x^{r+1} - x^r\|^2 - (\beta^{r+1} - \beta^r)\|x^{r+1} - x^r\|^2_{B^TB} \\
&\quad + \frac{\beta^r}{2}\left(\|x^r - x^{r-1}\|^2_{B^TB} - \|x^r - x^{r+1}\|^2_{B^TB} - \|w^r\|^2_{B^TB}\right) \\
&= L\|x^{r+1} - x^r\|^2 - \frac{\beta^{r+1} - \beta^r}{2}\|x^{r+1} - x^r\|^2_{B^TB} \\
&\quad + \frac{\beta^r}{2}\|x^r - x^{r-1}\|^2_{B^TB} - \frac{\beta^{r+1}}{2}\|x^r - x^{r+1}\|^2_{B^TB} - \frac{\beta^r}{2}\|w^r\|^2_{B^TB} \\
&\overset{(4.61)}{\leq} L\|x^{r+1} - x^r\|^2 + \frac{\beta^r}{2}\|x^r - x^{r-1}\|^2_{B^TB} - \frac{\beta^{r+1}}{2}\|x^r - x^{r+1}\|^2_{B^TB} - \frac{\beta^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

Therefore, combining the above three inequalities we obtain

$$\begin{aligned}
&\frac{\beta^{r+1}}{2}\|Ax^{r+1}\|^2 + \frac{\beta^{r+1}}{2}\|x^r - x^{r+1}\|^2_{B^TB} \\
&\leq \frac{\beta^r}{2}\|Ax^r\|^2 + \frac{\beta^r}{2}\|x^r - x^{r-1}\|^2_{B^TB} + \frac{\beta^{r+1} - \beta^r}{2(\beta^r)^2}\|\mu^{r-1} - \mu^r\|^2 + L\|x^{r+1} - x^r\|^2 - \frac{\beta^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

Multiplying both sides by β^r, we obtain

$$\begin{aligned}
&\frac{\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2+\frac{\beta^{r+1}\beta^r}{2}\|x^r-x^{r+1}\|^2_{B^TB}\\
&\leq \frac{\beta^r\beta^{r-1}}{2}\|Ax^r\|^2+\frac{\beta^r\beta^{r-1}}{2}\|x^r-x^{r-1}\|^2_{B^TB}+\frac{\beta^{r+1}-\beta^r}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2+\beta^rL\|x^{r+1}-x^r\|^2\\
&\quad-\frac{(\beta^r)^2}{2}\|w^r\|^2_{B^TB}+\frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|Ax^r\|^2+\frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\
&=\frac{\beta^r\beta^{r-1}}{2}\|Ax^r\|^2+\frac{\beta^r\beta^{r-1}}{2}\|x^r-x^{r-1}\|^2_{B^TB}+\frac{\beta^{r+1}-\beta^{r-1}}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2+\beta^rL\|x^{r+1}-x^r\|^2\\
&\quad-\frac{(\beta^r)^2}{2}\|w^r\|^2_{B^TB}+\frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB}
\end{aligned}\tag{2.50}$$

where in the last equality we have merged the terms $\frac{\beta^{r+1}-\beta^r}{2\beta^r}\|\mu^{r-1}-\mu^r\|^2$ and $\frac{\beta^r(\beta^r-\beta^{r-1})}{2}\|Ax^r\|^2$.

Step 4. In this step we construct and estimate the descent of the potential function. For some given $c>0$, let us define the potential function as

$$P_{\beta^{r+1},c}(x^{r+1},x^r,\mu^{r+1})=L_{\beta^{r+1}}(x^{r+1},\mu^{r+1})+\frac{c\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2+\frac{c\beta^{r+1}\beta^r}{2}\|x^r-x^{r+1}\|^2_{B^TB}.$$

Note that this potential function has some major differences compared with the one we used before; cf. (2.11). In particular, the second and the third terms are now quadratic, rather than linear, in the penalty parameters. This new construction is the key to our following analysis.

Then combining the estimate in (2.50) and (2.48), we obtain

$$\begin{aligned}
&P_{\beta^{r+1},c}(x^{r+1},x^r,\mu^{r+1})-P_{\beta^r,c}(x^r,x^{r-1},\mu^r)\\
&\leq-\left(\frac{\beta^{r+1}-L}{2}-\frac{3}{\beta^{r+1}\sigma_{\min}}\left(L^2+(\beta^{r+1}-\beta^r)^2\|B^TB\|\right)-c\beta^rL\right)\|x^{r+1}-x^r\|^2\\
&\quad+\frac{\beta^{r+1}-\beta^{r-1}}{2\beta^r}(\frac{1}{\beta^r}+c)\|\mu^r-\mu^{r-1}\|^2+\frac{c\beta^r(\beta^r-\beta^{r-1})}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\
&\quad-\left(\frac{c(\beta^r)^2}{2}-\frac{3(\beta^r)^2\|B^TB\|}{\beta^{r+1}\sigma_{\min}}\right)\|w^r\|^2_{B^TB}
\end{aligned}\tag{2.51}$$

where in the inequality we have also used the fact that $\beta^r\geq\beta^{r-1}$.

Taking the sum of r from t to T (for some $T > t > 1$) and utilize again the estimate in (2.46), we have

$$\begin{aligned}
&P_{\beta^{T+1},c}(x^{T+1}, x^T, \mu^{T+1}) - P_{\beta^t,c}(x^t, x^{t-1}, \mu^t) \\
&\leq \sum_{r=t}^{T} -\Bigg(\frac{\beta^{r+1} - L}{2} - \frac{3 + 3(1/\beta^r + c)(\beta^{r+1} - \beta^{r-1})/2\beta^r}{\beta^{r+1}\sigma_{\min}} \left(L^2 + (\beta^{r+1} - \beta^{r-1})^2 \|B^T B\| \right) \\
&\quad - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1} - \beta^r)\|B^T B\|}{2} \Bigg) \|x^{r+1} - x^r\|^2 \\
&\quad - \left(\frac{c(\beta^r)^2}{2} - \frac{(3 + 3(1/\beta^r + c)(\beta^{r+1} - \beta^{r-1})/2\beta^r)(\beta^r)^2 \|B^T B\|}{\beta^{r+1}\sigma_{\min}} \right) \|w^r\|^2_{B^T B} \\
&\quad + \frac{c\beta^t(\beta^t - \beta^{t-1})}{2} \|x^t - x^{t-1}\|^2_{B^T B} + \frac{\beta^{t+1} - \beta^{t-1}}{2\beta^t}(1/\beta^t + c)\|\mu^t - \mu^{t-1}\|^2. \qquad (2.52)
\end{aligned}$$

First, note that for any $c \in (0, 1)$, the coefficient in front of $\|w^r\|^2_{B^T B}$ becomes negative for sufficiently large (but finite) t. This is because $\{\beta^r\} \to \infty$, and that the first term in the parenthesis scales in $\mathcal{O}((\beta^r)^2)$ while the second term scales in $\mathcal{O}(\beta^r)$. For the first term to be negative, we need $c > 0$ to be *small enough* such that the following is true for large enough r

$$\frac{\beta^{r+1} - L}{2} - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1} - \beta^r)\|B^T B\|}{2} > \frac{\beta^{r+1}}{24}.$$

Suppose that r is large enough such that $(\beta^{r+1} - L)/2 > \beta^{r+1}/3$, or equivalently $\beta^{r+1} > 3L$. Also choose $c = \min\{1/(4L), 1/(12\kappa\|B^T B\|)\}$, where κ is given in (4.61). Then we have

$$\frac{\beta^{r+1} - L}{2} - c\beta^r L - \frac{c\beta^{r+1}(\beta^{r+1} - \beta^r)\|B^T B\|}{2} > \frac{\beta^{r+1}}{3} - \frac{\beta^{r+1}}{4} - \frac{\beta^{r+1}}{24} = \frac{\beta^{r+1}}{24}. \qquad (2.53)$$

For this given c, we can also show that the following is true for sufficiently large r

$$\begin{aligned}
&\frac{3 + 3(1/\beta^r + c)(\beta^{r+1} - \beta^{r-1})/2\beta^r}{\beta^{r+1}\sigma_{\min}} \left(L^2 + (\beta^{r+1} - \beta^r)^2 \|B^T B\| \right) \leq \frac{\beta^{r+1}}{48} \\
&\left(\frac{c(\beta^r)^2}{2} - \frac{(3 + 3(1/\beta^r + c)(\beta^{r+1} - \beta^{r-1})/2\beta^r)(\beta^r)^2 \|B^T B\|}{\beta^{r+1}\sigma_{\min}} \right) \geq \frac{c(\beta^r)^2}{48}.
\end{aligned}$$

In conclusion we have that for sufficiently large but finite t_0, we have

$$\begin{aligned}
&P_{\beta^{T+1},c}(x^{T+1}, x^T, \mu^{T+1}) - P_{\beta^{t_0},c}(x^{t_0-1}, x^{t_0}, \mu^{t_0}) \\
&\leq \sum_{r=t_0}^{T} \left(-\frac{\beta^{r+1}}{48} \|x^{r+1} - x^r\|^2 - \frac{c(\beta^r)^2}{48} \|w^r\|^2_{B^T B} \right) \\
&\quad + \frac{c\beta^{t_0}(\beta^{t_0} - \beta^{t_0-1})}{2} \|x^{t_0} - x^{t_0-1}\|^2_{B^T B} + \frac{\beta^{t_0+1} - \beta^{t_0-1}}{2\beta^{t_0}}(1/\beta^{t_0} + c)\|\mu^{t_0} - \mu^{t_0-1}\|^2. \qquad (2.54)
\end{aligned}$$

Therefore we conclude that if $\{\beta^{r+1}\}$ satisfies (4.61), and for $c > 0$ sufficiently small, there exits a finite $t_0 > 0$ such that for all $T > t_0$, the first two terms of the rhs of (2.52) are negative.

Step 5. Next we show that the potential function must be lower bounded. Observe that the augmented Lagrangian is given by

$$
\begin{aligned}
&L_{\beta^{r+1}}(x^{r+1}, \mu^{r+1}) \\
&= f(x^{r+1}) + \langle \mu^{r+1}, Ax^{r+1} \rangle + \frac{\beta^{r+1}}{2} \|Ax^{r+1}\|^2 \\
&= f(x^{r+1}) + \frac{1}{2\beta^{r+1}} \left(\|\mu^{r+1}\|^2 - \|\mu^r\|^2 + \|\mu^{r+1} - \mu^r\|^2 \right) + \frac{\beta^{r+1}}{2} \|Ax^{r+1}\|^2 \\
&= f(x^{r+1}) + \frac{1}{2\beta^{r+1}} \|\mu^{r+1}\|^2 - \frac{1}{2\beta^r} \|\mu^r\|^2 + \frac{1}{2\beta^{r+1}} \|\mu^{r+1} - \mu^r\|^2 + \left(\frac{1}{2\beta^r} - \frac{1}{2\beta^{r+1}} \right) \|\mu^r\|^2 + \frac{\beta^{r+1}}{2} \|Ax^{r+1}\|^2 \\
&\geq f(x^{r+1}) + \frac{1}{2\beta^{r+1}} \|\mu^{r+1}\|^2 - \frac{1}{2\beta^r} \|\mu^r\|^2 + \frac{1}{2\beta^{r+1}} \|\mu^{r+1} - \mu^r\|^2 + \frac{\beta^{r+1}}{2} \|Ax^{r+1}\|^2
\end{aligned}
$$

where we have used the fact that $\beta^{r+1} \geq \beta^r$. Note that t_0 in (2.54) is a finite number hence $\frac{1}{2\beta^{t_0}} \|\mu^{t_0}\|^2$ is finite, and utilize Assumption [A2], we conclude that

$$
\sum_{r=t_0}^{\infty} L_{\beta^{r+1}}(x^{r+1}, \mu^{r+1}) > -\infty. \tag{2.55}
$$

By noting that the remaining terms of the potential function are all nonnegative, we have

$$
\sum_{r=1}^{\infty} P_{\beta^{r+1},c}(x^{r+1}, x^r, \mu^{r+1}) > -\infty. \tag{2.56}
$$

Combining (2.56) and the bound (2.54) (which is true for a finite $t_0 > 0$), we conclude that the potential function $P_{\beta^{r+1},c}(x^{r+1}, x^r, \mu^{r+1})$ is lower bounded for all r.

Step 6. In this step we show that the successive differences of various quantities converge.

The lower boundedness of the potential function combined with the bound (2.54) (which is true for a finite $t_0 > 0$) implies that

$$
\sum_{r=1}^{\infty} \beta^{r+1} \|x^{r+1} - x^r\|^2 < \infty, \tag{2.57a}
$$

$$
\sum_{r=1}^{\infty} (\beta^r)^2 \|w^r\|^2_{B^T B} < \infty. \tag{2.57b}
$$

Therefore, we have

$$\beta^{r+1}\|x^{r+1}-x^r\|^2 \to 0, \tag{2.58a}$$

$$(\beta^r)^2\|w^r\|^2_{B^TB} \to 0. \tag{2.58b}$$

These two facts applied to (2.45), combined with $\mu^{r+1}-\mu^r \in \mathrm{col}(A)$, indicate that the following is true

$$\mu^{r+1}-\mu^r \to 0. \tag{2.59}$$

Also (2.54) implies that the potential function is *upper bounded* as well, and this indicates that

$$\frac{c\beta^{r+1}\beta^r}{2}\|Ax^{r+1}\|^2 \text{ is bounded}, \quad \frac{c\beta^{r+1}\beta^r}{2}\|x^r-x^{r+1}\|^2 \text{ is bounded}. \tag{2.60}$$

The second of the above inequality implies that $\beta^{r+1}B^TB(x^{r+1}-x^r)$ is bounded. If we further assume that $\nabla f(x)$ is bounded, and use (2.44), we can conclude that $\{\mu^r\}$ is bounded.

Step 7. Next we show that every limit point of (x^r,μ^r) converges to a stationary solution of problem (5.13). Let us pass a subsequence $\mathcal{K}$ to (x^r,μ^r) and denote (x^*,μ^*) as its limit point. For notational simplicity, in the following the index r all belongs to the set $\mathcal{K}$.

From relation (2.57a) we have that any given $\epsilon>0$, there exists t large enough such that the following is true

$$\sum_{r=t-1}^{\infty}\beta^{r+1}\|x^{r+1}-x^r\|^2 \le \frac{\epsilon}{c\kappa 16}. \tag{2.61}$$

Utilizing (2.47), we have that the following is true

$$\sum_{r=1}^{\infty}\frac{1}{\beta^{r+1}}\|v^{r+1}\|^2<\infty, \quad \lim_{t\to\infty}\sum_{r=t}^{\infty}(\beta^r)^2\|w^r\|^2_{B^TB}=0. \tag{2.62}$$

The first relation implies that $\liminf_{r\to\infty}\|v^{r+1}\|=0$. Applying these relations to (2.46), we have

$$\sum_{r=1}^{\infty}\frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2<\infty.$$

This implies that for any given $\epsilon>0$, $c>0$, there exists an index t sufficiently large such that

$$\sum_{r=t-1}^{\infty}\frac{1}{\beta^{r+1}}\|\mu^{r+1}-\mu^r\|^2<\frac{\epsilon^2}{4096L\|B^TB\|_F\kappa(1+c)}. \tag{2.63}$$

Applying this inequality and (2.61) to (2.54), we have that for large enough t and for any $T > t$ the following is true

$$P_{\beta^{T+1},c}(x^{T+1}, x^T, \mu^{T+1}) - P_{\beta^t,c}(x^t, x^{t-1}, \mu^t) \leq -\sum_{r=t}^{T} \left(\frac{\beta^{r+1}}{48} \|x^{r+1} - x^r\|^2 \right) + \frac{\epsilon^2}{4096L\|B^T B\|}. \tag{2.64}$$

Next we modify a classical argument in [15][Proposition 3.5] to show that

$$\lim_{r\to\infty} \|v^{r+1}\| \to 0.$$

We already know from the first relation in (2.62) that $\liminf_{r\to\infty} \|v^{r+1}\| = 0$. Suppose that $\|v^{r+1}\|$ does not converge to 0, then we must have $\limsup_{r\to\infty} \|v^{r+1}\| > 0$. Hence there exists an $\epsilon > 0$ such that $\|v^{r+1}\| < \epsilon/2$ for infinitely many r, and $\|v^{r+1}\| > \epsilon$ for infinitely many r. Then there exists an infinite subset of iteration indices $\mathcal{R}$ such that for each $r \in \mathcal{R}$, there exits a $t(r)$ such that

$$\begin{aligned} &\|v^r\| < \epsilon/2, \quad , \|v^{t(r)}\| > \epsilon, \\ &\epsilon/2 < \|v^t\| \leq \epsilon, \quad \forall\ r < t < t(r). \end{aligned}$$

Using the fact that $\lim_{r\in\mathcal{K}} \mu^r = \mu^*$, we have that for r large enough, the following is true for all $t \geq 0$

$$\|\mu^r - \mu^{r+t}\| \leq \frac{\epsilon}{8} \frac{1}{\|(B^T B)^{-1}\| \|A^T A\|}. \tag{2.65}$$

Without loss of generality we can assume that this relation holds for all $r \in \mathcal{R}$. Note that the following is true

$$\begin{aligned} \frac{\epsilon}{2} \leq \|v^{t(r)}\| - \|v^r\| \leq \|v^{t(r)} - v^r\| &= \left\| (B^T B)^{-1} \sum_{t=r}^{t(r)-1} \left(\nabla f(x^{t+1}) - \nabla f(x^t) + A^T(\mu^{t+1} - \mu^t) \right) \right\| \\ &\leq \|(B^T B)^{-1}\| \left(\sum_{t=r}^{t(r)-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\| + \|A^T A\| \|\mu^{t(r)} - \mu^r\| \right) \\ &\overset{(2.47)}{\leq} \|(B^T B)^{-1}\| \left(\sum_{t=r}^{t(r)-1} \frac{L}{\beta^{t+1}} \|v^{t+1}\| + \|A^T A\| \|\mu^{t(r)} - \mu^r\| \right) \\ &\leq \epsilon L \|(B^T B)^{-1}\| \sum_{t=r}^{t(r)-1} \frac{1}{\beta^{t+1}} + \frac{\epsilon}{8} \end{aligned} \tag{2.66}$$

where in the last inequality we have used (2.65) and the fact that for all $t \in (r+1, t(r))$, we have $\|v^t\| < \epsilon$. This implies that

$$\frac{3}{8L\|(B^TB)^{-1}\|} \leq \sum_{t=r}^{t(r)-1} \frac{1}{\beta^{t+1}}. \tag{2.67}$$

Using the descent of the potential function (2.64) we have, for $r \in \mathcal{R}$ and r large enough

$$\begin{aligned}
&P_{\beta^{t(r)},c}(x^{t(r)}, x^{t(r)-1}, \mu^{t(r)}) - P_{\beta^r,c}(x^r, x^{r-1}, \mu^r) \\
&\leq -\sum_{t=r}^{t(r)-1} \frac{1}{48\beta^{t+1}}\|v^{t+1}\|^2 + \frac{\epsilon^2}{4096L\|B^TB\|} \\
&\overset{\text{(i)}}{\leq} -\left(\frac{\epsilon}{4}\right)^2 \sum_{t=r}^{t(r)-1} \frac{1}{48\beta^{t+1}} + \frac{\epsilon^2}{4096L\|B^TB\|} \\
&\overset{\text{(ii)}}{\leq} -\frac{\epsilon^2}{2048L\|B^TB\|} + \frac{\epsilon^2}{4096L\|B^TB\|} \\
&\leq -\frac{\epsilon^2}{4096L\|B^TB\|}
\end{aligned} \tag{2.68}$$

where in (i) we have used the fact that for all $r \in \mathcal{R}$, $\|v^{r+i}\| \geq \frac{\epsilon}{2}$ for $i = 1, \cdots, t(r)$; in (ii) we have used (2.67). However we know that the potential function is convergent, i.e.,

$$\lim_{r\to\infty} P_{\beta^{t(r)},c}(x^{t(r)}, x^{t(r)-1}, \mu^{t(r)}) \to P_{\beta^r,c}(x^r, x^{r-1}, \mu^r) = 0$$

which contradicts to (4.108). Therefore we conclude that $\|v^{r+1}\| \to 0$.

Finally, combining $\|v^{r+1}\| \to 0$ with the convergence of $\mu^{r+1} - \mu^r$ (cf. (2.59)), we conclude that every limit point of $\{x^r, \mu^r\}$ satisfies

$$\nabla f(x^*) + A^T\mu^* = 0, \quad Ax^* = 0.$$

Therefore it is a stationary solution for problem (5.13). This completes the proof.

2.8.9 Proof of Convergence for Algorithm 2

To make the derivation compact, define the following matrix

$$\begin{aligned}
\boldsymbol{M}^{r+1} &:= \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1}) \\
&= \left[((X_1^{r+1}y_1^{r+1}) - z_1)(y_1^{r+1})^T + 2\gamma X_1^{r+1}; \cdots; ((X_N^{r+1}y_N^{r+1}) - z_N)(y_N^{r+1})^T + 2\gamma X_N^{r+1}\right].
\end{aligned} \tag{2.69}$$

The proof consists of six steps.

Step 1. First we note that the optimality condition for the $\boldsymbol{X}$-subproblem (2.30c) is given by

$$\boldsymbol{A}^T\boldsymbol{\Omega}^{r+1} = -\boldsymbol{M}^{r+1} - \beta\langle \boldsymbol{B}^T\boldsymbol{B}, (\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\rangle. \tag{2.70}$$

By utilizing the fact that $\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r$ lies in the column space of $\boldsymbol{A}$, and the eigenvalues of A^TA equal to those of $\boldsymbol{A}^T\boldsymbol{A}$, we have the following bound

$$\|\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r\|_F^2 \le \frac{2}{\sigma_{\min}}\left(\|\boldsymbol{M}^{r+1} - \boldsymbol{M}^r\|_F^2 + \beta^2\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2\right).$$

Next let us analyze the first term in the rhs of the above inequality. The following identity holds true

$$\begin{aligned}
&\|\boldsymbol{M}^{r+1} - \boldsymbol{M}^r\|_F^2 \\
&= \sum_{i=1}^N \|(X_i^{r+1}y_i^{r+1} - z_i)(y_i^{r+1})^T - (X_i^r y_i^r - z_i)(y_i^r)^T + 2\gamma(X_i^{r+1} - X_i^r)\|_F^2 \\
&\le \sum_{i=1}^N 4\|X_i^{r+1} - X_i^r\|_F^2\|y_i^{r+1}(y_i^{r+1})^T\|^2 + 4\|X_i^r y_i^r - z_i\|^2\|y_i^{r+1} - y_i^r\|^2 \\
&\quad + 4\|X_i^r(y_i^{r+1} - y_i^r)\|^2\|y_i^{r+1}\|^2 + 16\gamma^2\|X_i^{r+1} - X_i^r\|_F^2 \\
&\le \sum_{i=1}^N 4(\tau^2 + 4\gamma^2)\|X_i^{r+1} - X_i^r\|_F^2 + 4\theta_i^r\|y_i^{r+1} - y_i^r\|^2 + 4\tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2
\end{aligned} \tag{2.71}$$

where in the last inequality we have defined the constant θ_i^r as

$$\theta_i^r := \|X_i^r y_i^r - z_i\|^2. \tag{2.72}$$

Therefore, combining the above two inequalities, we obtain

$$\begin{aligned}
\frac{1}{\beta}\|\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r\|_F^2 \le \frac{8}{\beta\sigma_{\min}}\sum_{i=1}^N \left((\tau^2 + 4\gamma^2)\|X_i^{r+1} - X_i^r\|_F^2 + \theta_i^r\|y_i^{r+1} - y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1} - y_i^r)\|^2\right) \\
+ \frac{2\beta}{\sigma_{\min}}\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2
\end{aligned} \tag{2.73}$$

Step 2. Next let us analyze the descent of the augmented Lagrangian. First we have

$$\begin{aligned}
&L_\beta(\boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^r) - L_\beta(\boldsymbol{X}^r, Y^r, \boldsymbol{\Omega}^r) \\
&= \sum_{i=1}^N \left(\frac{1}{2}\|X_i^r y_i^{r+1} - z_i\|^2 + h_i(y_i^{r+1}) - \frac{1}{2}\|X_i^r y_i^r - z_i\|^2 - h_i(y_i^r) \right) \\
&\le \sum_{i=1}^N \left(\frac{1}{2}\|X_i^r y_i^{r+1} - z_i\|^2 + h_i(y_i^{r+1}) + \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2 - \frac{1}{2}\|X_i^r y_i^r - z_i\|^2 - h_i(y_i^r) \right) \\
&\le \sum_{i=1}^N \Bigg(\langle (X_i^r)^T (X_i^r y_i^{r+1} - z_i) + \theta_i^r (y_i^{r+1} - y_i^r), y_i^{r+1} - y_i^r \rangle - \frac{1}{2}\|X_i^r (y_i^{r+1} - y_i^r)\|^2 \\
&\quad - \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2 + \langle \zeta_i^{r+1}, y_i^{r+1} - y_i^r \rangle \Bigg) \\
&\le -\sum_{i=1}^N \left(\frac{1}{2}\|X_i^r (y_i^{r+1} - y_i^r)\|^2 + \frac{\theta_i^r}{2}\|y_i^{r+1} - y_i^r\|^2 \right)
\end{aligned} \tag{2.74}$$

where in the second to the last equality we have used the convexity of h_i, and $\zeta_i^{r+1} \in \partial h_i(y_i^{r+1})$; the last inequality uses the optimality condition of the y-step (2.30b). Similarly, we can show that

$$L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^r) - L_\beta(\boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^r) \le -\frac{\beta + 2\gamma}{2}\|\boldsymbol{X}^{r+1} - \boldsymbol{X}^r\|_F^2 \tag{2.75}$$

where we have utilized the fact that $\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{B}^T\boldsymbol{B} = 2\mathbf{D} \succeq \mathbf{I}_{NM}$. Therefore, combining the estimate (2.73), we obtain

$$\begin{aligned}
&L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) - L_\beta(\boldsymbol{X}^r, Y^r, \boldsymbol{\Omega}^r) \\
&\le -\left(\frac{\beta + 2\gamma}{2} - \frac{8(\tau^2 + 4\gamma^2)}{\beta\sigma_{\min}} \right) \sum_{i=1}^N \|X_i^{r+1} - X_i^r\|_F^2 - \sum_{i=1}^N \left(\frac{\theta_i^r}{2} - \frac{8\theta_i^r}{\beta\sigma_{\min}} \right) \|y_i^{r+1} - y_i^r\|^2 \\
&\quad - \left(\frac{1}{2} - \frac{8\tau}{\sigma_{\min}\beta} \right) \sum_{i=1}^N \|X_i^r (y_i^{r+1} - y_i^r)\|^2 + \frac{2\beta}{\sigma_{\min}} \|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\|_F^2.
\end{aligned} \tag{2.76}$$

Step 3. This step follows Lemma 3 in the analysis of Algorithm 1. In particular, after writing down the optimality condition of the $\boldsymbol{X}^{r+1}$ and $\boldsymbol{X}^r$ step, we can obtain

$$\begin{aligned}
&\langle \boldsymbol{A}^T(\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r), \boldsymbol{X}^{r+1} - \boldsymbol{X}^r \rangle \\
&\le -\left\langle \boldsymbol{M}^{r+1} - \boldsymbol{M}^r + \beta\boldsymbol{B}^T\boldsymbol{B}\left[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})\right], \boldsymbol{X}^{r+1} - \boldsymbol{X}^r \right\rangle.
\end{aligned}$$

Then it is easy to show that the above inequality implies the following

$$\begin{aligned}
&\frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\right)\\
&\le \frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r}, \boldsymbol{A}\boldsymbol{X}^{r}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r}-\boldsymbol{X}^{r-1}), \boldsymbol{X}^{r}-\boldsymbol{X}^{r-1}\rangle\right) - \frac{\beta}{2}\langle \boldsymbol{A}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{A}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)\rangle\\
&\quad - \langle \boldsymbol{M}^{r+1}-\boldsymbol{M}^r, \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle - \frac{\beta}{2}\|\boldsymbol{B}[(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)-(\boldsymbol{X}^r-\boldsymbol{X}^{r-1})]\|_F^2.
\end{aligned}$$

Note the following fact

$$\begin{aligned}
&-\langle \boldsymbol{M}^{r+1}-\boldsymbol{M}^r, \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\\
&= -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\\
&= -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{r+1}, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) + \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\\
&\overset{\text{(i)}}{\le} -\langle \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\\
&\overset{\text{(ii)}}{\le} \frac{1}{2d}\|\nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^{r+1}) - \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^r, Y^r)\|_F^2 + \frac{d}{2}\|\boldsymbol{X}^{r+1}-\boldsymbol{X}^r\|_F^2\\
&\overset{\text{(iii)}}{\le} \frac{1}{d}\sum_{i=1}^N \left(\theta_i^r\|y_i^{r+1}-y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1}-y_i^r)\|^2\right) + \frac{d}{2}\|\boldsymbol{X}^{r+1}-\boldsymbol{X}^r\|_F^2
\end{aligned} \tag{2.77}$$

where in (i) we utilize the convexity of $f(\boldsymbol{X}, Y)$ wrt $\boldsymbol{X}$ for any fixed y; in (ii) we use the Cauchy-Swartz inequality, where $d > 0$ is a constant (to be determined later); (iii) is true due to a similar calculation as in (2.71).

Overall we have

$$\begin{aligned}
&\frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\right)\\
&\le \frac{\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r}, \boldsymbol{A}\boldsymbol{X}^{r}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r}-\boldsymbol{X}^{r-1}), \boldsymbol{X}^{r}-\boldsymbol{X}^{r-1}\rangle\right) - \frac{\beta}{2}\langle \boldsymbol{A}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{A}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)\rangle\\
&\quad + \frac{1}{d}\sum_{i=1}^N \left(\theta_i^r\|y_i^{r+1}-y_i^r\|^2 + \tau\|X_i^r(y_i^{r+1}-y_i^r)\|^2\right) + \frac{d}{2}\|\boldsymbol{X}^{r+1}-\boldsymbol{X}^r\|_F^2\\
&\quad - \frac{\beta}{2}\|\boldsymbol{B}[(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r)-(\boldsymbol{X}^r-\boldsymbol{X}^{r-1})]\|_F^2
\end{aligned} \tag{2.78}$$

Step 4. Let us define the potential function as

$$\begin{aligned}
&P_{\beta,c}(\boldsymbol{X}^{r+1}, \boldsymbol{X}^r, Y^{r+1}, \boldsymbol{\Omega}^{r+1})\\
&:= L_\beta(\boldsymbol{X}^{r+1}, Y^{r+1}, \boldsymbol{\Omega}^{r+1}) + \frac{c\beta}{2}\left(\langle \boldsymbol{A}\boldsymbol{X}^{r+1}, \boldsymbol{A}\boldsymbol{X}^{r+1}\rangle + \langle \boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1}-\boldsymbol{X}^r), \boldsymbol{X}^{r+1}-\boldsymbol{X}^r\rangle\right).
\end{aligned} \tag{2.79}$$

Then utilize the bounds (2.76) and (2.78), we obtain

$$\begin{aligned}
&P_{\beta,c}(\boldsymbol{X}^{r+1},\boldsymbol{X}^{r},Y^{r+1},\boldsymbol{\Omega}^{r+1})-P_{\beta,c}(\boldsymbol{X}^{r},\boldsymbol{X}^{r-1},Y^{r},\boldsymbol{\Omega}^{r})\\
\leq\quad & -\left(\frac{\beta+2\gamma}{2}-\frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}}-\frac{cd}{2}\right)\sum_{i=1}^{N}\|X_i^{r+1}-X_i^{r}\|_F^2\\
&-\sum_{i=1}^{N}\left(\frac{\theta_i^r}{2}-\frac{8\theta_i^r}{\beta\sigma_{\min}}-\frac{c\theta_i^r}{d}\right)\|y_i^{r+1}-y_i^r\|^2-\left(\frac{1}{2}-\frac{8\tau}{\sigma_{\min}\beta}-\frac{c\tau}{d}\right)\sum_{i=1}^{N}\|X_i^r(y_i^{r+1}-y_i^r)\|^2\\
&-\left(\frac{c\beta}{2}-\frac{2\beta\|B^TB\|}{\sigma_{\min}}\right)\|\boldsymbol{B}[(\boldsymbol{X}^{r+1}-\boldsymbol{X}^{r})-(\boldsymbol{X}^{r}-\boldsymbol{X}^{r-1})]\|_F^2.
\end{aligned}$$

Therefore the following are the condition that guarantees the descent of the potential function

$$\begin{aligned}
\frac{\beta+2\gamma}{2}-\frac{8(\tau^2+4\gamma^2)}{\beta\sigma_{\min}}-\frac{cd}{2}>0,\quad \frac{1}{2}-\frac{8}{\sigma_{\min}\beta}-\frac{c}{d}>0\\
\frac{1}{2}-\frac{8\tau}{\sigma_{\min}\beta}-\frac{c\tau}{d}>0,\quad \frac{c\beta}{2}-\frac{2\beta\|B^TB\|}{\sigma_{\min}}>0.
\end{aligned}\tag{2.80}$$

To see that it is always possible to find the tuple (β,c,d), first let us set c such that the last inequality is satisfied

$$c>\frac{4\|B^TB\|}{\sigma_{\min}}.\tag{2.81}$$

Second, let us pick any d such that the following is true

$$d>\max\{2c\tau,2c\}.$$

Then clearly it is possible to make β large enough such that all the four conditions in (2.80) are satisfied.

Step 5. We need to prove that the potential function is lower bounded. We lower bound the augmented Lagrangian as follows

$$\begin{aligned}
&L_\beta(\boldsymbol{X}^{r+1},Y^{r+1},\boldsymbol{\Omega}^{r+1})\\
&=\sum_{i=1}^{N}\left(\frac{1}{2}\|X_i^{r+1}y_i^{r+1}-z_i\|^2+\gamma\|X_i^{r+1}\|_F^2+h_i(y_i^{r+1})\right)+\langle\boldsymbol{\Omega}^{r+1},\boldsymbol{A}\boldsymbol{X}^{r+1}\rangle+\frac{\beta}{2}\langle\boldsymbol{A}\boldsymbol{X}^{r+1},\boldsymbol{A}\boldsymbol{X}^{r+1}\rangle\\
&=\sum_{i=1}^{N}\left(\frac{1}{2}\|X_i^{r+1}y_i^{r+1}-z_i\|^2+\gamma\|X_i^{r+1}\|_F^2+h_i(y_i^{r+1})\right)+\frac{\beta}{2}\langle\boldsymbol{A}\boldsymbol{X}^{r+1},\boldsymbol{A}\boldsymbol{X}^{r+1}\rangle\\
&\quad+\frac{1}{2\beta}\left(\|\boldsymbol{\Omega}^{r+1}-\boldsymbol{\Omega}^{r}\|_F^2+\|\boldsymbol{\Omega}^{r+1}\|_F^2-\|\boldsymbol{\Omega}^{r}\|_F^2\right).
\end{aligned}\tag{2.82}$$

Then by the same argument leading to (2.39), we conclude that as long as h_i is lower bounded over its domain, then the potential function will be lower bounded.

Step 6. Combining the results in Step 5 and Step 4, we conclude the following

$$\sum_{i=1}^{N} \|X_i^{r+1} - X_i^r\|_F^2 \to 0, \quad \sum_{i=1}^{N} \|y_i^{r+1} - y_i^r\|^2 \to 0 \tag{2.83a}$$

$$\sum_{i=1}^{N} \|X_i^r(y_i^{r+1} - y_i^r)\|^2 \to 0, \quad \left\|\boldsymbol{B}^T\boldsymbol{B}[(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r) - (\boldsymbol{X}^r - \boldsymbol{X}^{r-1})]\right\|_F \to 0. \tag{2.83b}$$

Then utilizing (2.73), we have

$$\boldsymbol{\Omega}^{r+1} - \boldsymbol{\Omega}^r \to \mathbf{0}, \text{ or equivalently } \boldsymbol{A}\boldsymbol{X}^{r+1} \to 0.$$

That is, in the limit the network-wide consensus is achieved. Next we show that the primal and dual iterates are bounded.

Note that the potential function is both lower and upper bounded. Combined with (2.83) we must have that the augmented Lagrangian is both upper and lower bounded. Using the expression (2.82), the assumption that $h_i(y_i)$ is lower bounded, and the fact that y_i is bounded, we have that in the limit, the following term is bounded

$$\sum_{i=1}^{N} \frac{1}{2}\|X_i^{r+1}y_i^{r+1} - z_i\|^2 + \gamma\|X_i^{r+1}\|_F^2.$$

This implies that the primal variable sequence $\{X_i^{r+1}\}$ are bounded for all i. To show the boundedness of the dual sequence, note that $\boldsymbol{\Omega}^{r+1} \in \text{col}(\boldsymbol{A})$ (due to the initialization that $\boldsymbol{\Omega}^0 = \mathbf{0}$). Therefore using (2.70) we have

$$\sigma_{\min}(\boldsymbol{A}^T\boldsymbol{A})\|\boldsymbol{\Omega}^{r+1}\|_F^2 \leq 2\|\boldsymbol{M}^{r+1}\|_F^2 + 2\beta\|\boldsymbol{B}^T\boldsymbol{B}(\boldsymbol{X}^{r+1} - \boldsymbol{X}^r)\|_F^2$$

Note that from the expression of $\boldsymbol{M}$ in (2.69), we see that $\{\boldsymbol{M}^{r+1}\}$ is bounded because both $\boldsymbol{X}^{r+1}$ and Y^{r+1} are bounded. Similarly, the second term on the rhs of the above inequality is bounded because $\boldsymbol{X}^{r+1} \to \boldsymbol{X}^r$. These two facts imply that $\{\boldsymbol{\Omega}^{r+1}\}$ is bounded as well.

Arguing the convergence to stationary point as well as the convergence rate follows exactly the same steps as in the proof of Theorem 1.

CHAPTER 3. A NONCONVEX PRIMAL-DUAL SPLITTING METHOD FOR DISTRIBUTED AND STOCHASTIC OPTIMIZATION

Abstract

We study a stochastic and distributed algorithm for nonconvex problems whose objective consists of a sum of N nonconvex L_i/N-smooth functions, plus a nonsmooth regularizer. The proposed NonconvEx primal-dual SpliTTing (NESTT) algorithm splits the problem into N subproblems, and utilizes an augmented Lagrangian based primal-dual scheme to solve it in a distributed and stochastic manner. With a special non-uniform sampling, a version of NESTT achieves ϵ-stationary solution using $\mathcal{O}((\sum_{i=1}^N \sqrt{L_i/N})^2/\epsilon)$ gradient evaluations, which can be up to $\mathcal{O}(N)$ times better than the (proximal) gradient descent methods. It also achieves Q-linear convergence rate for nonconvex ℓ_1 penalized quadratic problems with polyhedral constraints. Further, we reveal a fundamental connection between *primal-dual* based methods and a few *primal only* methods such as IAG/SAG/SAGA.

3.1 Introduction

Consider the following nonconvex and nonsmooth constrained optimization problem

$$\min_{z\in Z} \quad f(z) := \frac{1}{N}\sum_{i=1}^{N} g_i(z) + g_0(z) + p(z), \tag{3.1}$$

where $Z \subseteq \mathbb{R}^d$; for each $i \in \{0, \cdots, N\}$, $g_i : \mathbb{R}^d \to \mathbb{R}$ is a smooth possibly nonconvex function which has L_i-Lipschitz continuous gradient; $p(z) : \mathbb{R}^d \to \mathbb{R}$ is a lower semi-continuous convex but possibly nonsmooth function. Define $g(z) := \frac{1}{N}\sum_{i=1}^N g_i(z)$ for notational simplicity.

Problem (4.1) is quite general. It arises frequently in applications such as machine learning and signal processing; see a recent survey [23]. In particular, each smooth functions $\{g_i\}_{i=1}^N$ can represent: 1) a mini-batch of loss functions modeling data fidelity, such as the ℓ_2 loss, the logistic

loss, etc; 2) nonconvex activation functions for neural networks, such as the logit or the tanh functions; 3) nonconvex utility functions used in signal processing and resource allocation, see [18]. The smooth function g_0 can represent smooth nonconvex regularizers such as the non-quadratic penalties [7], or the smooth part of the SCAD or MCP regularizers (which is a concave function) [137]. The convex function p can take the following form: 1) nonsmooth convex regularizers such as ℓ_1 and ℓ_2 functions; 2) an indicator function for convex and closed feasible set Z, denoted as $\iota_Z(\cdot)$; 3) convex functions without global Lipschitz continuous gradient, such as $p(z) = z^4$ or $p(z) = 1/z + \iota_{z\geq 0}(z)$.

In this work we solve (4.1) in a stochastic and distributed manner. We consider the setting in which N distributed agents each having the knowledge of one smooth function $\{g_i\}_{i=1}^N$, and they are connected to a cluster center which handles g_0 and p. At any given time, a randomly selected agent is activated and performs computation to optimize its local objective. Such distributed computation model has been popular in large-scale machine learning and signal processing [20]. Such model is also closely related to the (centralized) stochastic *finite-sum* optimization problem [78, 31, 71, 111, 4, 118], in which each time the iterate is updated based on the gradient information of a random component function. One of the key differences between these two problem types is that in the distributed setting there can be disagreement between local copies of the optimization variable z, while in the centralized setting only one copy of z is maintained.

Our Contributions. We propose a class of NonconvEx primal-dual SpliTTing (NESTT) algorithms for problem (4.1). We split $z \in \mathbb{R}^d$ into local copies of $x_i \in \mathbb{R}^d$, while enforcing the equality constraints $x_i = z$ for all i. That is, we consider the following reformulation of (4.1)

$$\min_{x,z\in\mathbb{R}^d} \quad \ell(x,z) := \frac{1}{N}\sum_{i=1}^N g_i(x_i) + g_0(z) + h(z), \quad \text{s.t. } x_i = z,\ i = 1,\cdots,N, \tag{3.2}$$

where $h(z) := \iota_Z(z) + p(z)$, $x := [x_1;\cdots;x_N]$. Our algorithm uses the Lagrangian relaxation of the equality constraints, and at each iteration a (possibly non-uniformly) randomly selected primal variable is optimized, followed by an approximate dual ascent step. Note that such splitting scheme has been popular in the convex setting [20], but not so when the problem becomes nonconvex.

The NESTT is one of the first stochastic algorithms for distributed nonconvex nonsmooth optimization, with provable and nontrivial convergence rates. Our main contribution is given below. First, in terms of some primal and dual optimality gaps, NESTT converges sublinearly to a point belongs to stationary solution set of (3.2). Second, NESTT converges Q-linearly for certain nonconvex ℓ_1 penalized quadratic problems. To the best of our knowledge, this is the first time that linear convergence is established for stochastic and distributed optimization of such type of problems. Third, we show that a gradient-based NESTT with *non-uniform sampling* achieves an ϵ-stationary solution of (4.1) using $\mathcal{O}((\sum_{i=1}^N \sqrt{L_i/N})^2/\epsilon)$ gradient evaluations. Compared with the classical gradient descent, which in the worst case requires $\mathcal{O}(\sum_{i=1}^N L_i/\epsilon)$ gradient evaluation to achieve ϵ-stationarity, our obtained rate can be up to $\mathcal{O}(N)$ times better in the case where the L_i's are not equal.

Our work also reveals a fundamental connection between *primal-dual* based algorithms and the *primal only* average-gradient based algorithm such as SAGA/SAG/IAG [31, 118, 19]. With the key observation that the dual variables in NESTT serve as the "memory" of the past gradients, one can specialize NESTT to SAGA/SAG/IAG. Therefore, NESTT naturally generalizes these algorithms to the nonconvex nonsmooth setting. It is our hope that by bridging the primal-dual splitting algorithms and primal-only algorithms (in *both* the convex and nonconvex setting), there can be significant further research developments benefiting both algorithm classes.

Related Work. Many stochastic algorithms have been designed for (3.2) when it is convex. In these algorithms the component functions g_i's are randomly sampled and optimized. Popular algorithms include the SAG/SAGA [31, 118], the SDCA [120], the SVRG [71], the RPDG [78] and so on. When the problem becomes nonconvex, the well-known incremental based algorithm can be used [128, 13], but these methods generally lack convergence rate guarantees. The SGD based method has been studied in [44], with $\mathcal{O}(1/\epsilon^2)$ convergence rate. Recent works [4] and [111] develop algorithms based on SVRG and SAGA for a special case of (4.1) where the entire problem is smooth and unconstrained. To the best of our knowledge there has been no stochastic algorithms with provable, and non-trivial, convergence rate guarantees for solving problem (4.1).

On the other hand, distributed stochastic algorithms for solving problem (4.1) in the nonconvex setting has been proposed in [63], in which each time a randomly picked subset of agents update their local variables. However there has been no convergence rate analysis for such distributed stochastic scheme. There has been some recent distributed algorithms designed for (4.1) [92], but again without global convergence rate guarantee.

Preliminaries. The augmented Lagrangian function for problem (4.1) is given by:

$$L(x,z;\lambda)=\sum_{i=1}^{N}\left(\frac{1}{N}g_i(x_i)+\langle\lambda_i,x_i-z\rangle+\frac{\eta_i}{2}\|x_i-z\|^2\right)+g_0(z)+h(z), \tag{3.3}$$

where $\lambda:=\{\lambda_i\}_{i=1}^N$ is the set of dual variables, and $\eta:=\{\eta_i>0\}_{i=1}^N$ are penalty parameters.

We make the following assumptions about problem (4.1) and the function (3.3).

A-(a) The function $f(z)$ is bounded from below over $Z\cap\mathrm{int}(\mathrm{dom}\ f)$: $\underline{f}:=\min_{z\in Z}f(z)>-\infty$. $p(z)$ is a convex lower semi-continuous function; Z is a closed convex set.

A-(b) The g_i's and g have Lipschitz continuous gradients, i.e.,

$$\|\nabla g(y)-\nabla g(z)\|\le L\|y-z\|,\text{and}\quad \|\nabla g_i(y)-\nabla g_i(z)\|\le L_i\|y-z\|,\ \forall\ y,z$$

Clearly $L\le 1/N\sum_{i=1}^N L_i$, and the equality can be achieved in the worst case. For simplicity of analysis we will further assume that $L_0\le\frac{1}{N}\sum_{i=1}^N L_i$.

A-(c) Each η_i in (3.3) satisfies $\eta_i>L_i/N$; if g_0 is nonconvex, then $\sum_{i=1}^N\eta_i>3L_0$.

Assumption A-(c) implies that $L(x,z;\lambda)$ is *strongly convex* w.r.t. each x_i and z, with modulus $\gamma_i:=\eta_i-L_i/N$ and $\gamma_z=\sum_{i=1}^N\eta_i-L_0$, respectively [149, Theorem 2.1].

We then define the *prox-gradient* (`pGRAD`) for (4.1), which will serve as a measure of stationarity. It can be checked that the `pGRAD` vanishes at the set of stationary solutions of (4.1) [110].

Definition 1 *The proximal gradient of problem (4.1) is given by (for any $\gamma>0$)*

$$\tilde{\nabla}f_\gamma(z):=\gamma\left(z-prox^\gamma_{p+\iota_Z}[z-1/\gamma\nabla(g(z)+g_0(z))]\right),\ \textit{with}\quad prox^\gamma_{p+\iota_Z}[u]:=\operatorname*{argmin}_{u\in Z}\ p(u)+\frac{\gamma}{2}\|z-u\|^2.$$

Algorithm 3 NESTT-G Algorithm

1: **for** $r = 1$ **to** R **do**
2: Pick $i_r \in \{1, 2, \cdots, N\}$ with probability p_{i_r} and update (x, λ)

$$x_{i_r}^{r+1} = \arg\min_{x_{i_r}} V_{i_r}\left(x_{i_r}, z^r, \lambda_{i_r}^r\right); \tag{3.4}$$

$$\lambda_{i_r}^{r+1} = \lambda_{i_r}^r + \alpha_{i_r}\eta_{i_r}\left(x_{i_r}^{r+1} - z^r\right); \tag{3.5}$$

$$\lambda_j^{r+1} = \lambda_j^r, \quad x_j^{r+1} = z^r, \quad \forall\, j \neq i_r; \tag{3.6}$$

Update z: $$z^{r+1} = \arg\min_{z \in Z} L(\{x_i^{r+1}\}, z; \lambda^r). \tag{3.7}$$

3: **end for**
4: **Output:** (z^m, x^m, λ^m) where m randomly picked from $\{1, 2, \cdots, R\}$.

3.2 The NESTT-G Algorithm

Algorithm Description. We present a primal-dual splitting scheme for the reformulated problem (3.2). The algorithm is referred to as the NESTT with Gradient step (NESTT-G) since each agent only requires to know the gradient of each component function. To proceed, let us define the following function (for some constants $\{\alpha_i > 0\}_{i=1}^N$):

$$V_i(x_i, z; \lambda_i) = \frac{1}{N} g_i(z) + \frac{1}{N}\langle \nabla g_i(z), x_i - z\rangle + \langle \lambda_i, x_i - z\rangle + \frac{\alpha_i \eta_i}{2}\|x_i - z\|^2.$$

Note that $V_i(\cdot)$ is related to $L(\cdot)$ in the following way: it is a quadratic approximation (approximated at the point z) of $L(x, y; \lambda)$ w.r.t. x_i. The parameters $\alpha := \{\alpha_i\}_{i=1}^N$ give some freedom to the algorithm design, and they are critical in improving convergence rates as well as in establishing connection between NESTT-G with a few primal only stochastic optimization schemes.

The algorithm proceeds as follows. Before each iteration begins the cluster center broadcasts z to everyone. At iteration $r+1$ a randomly selected agent $i_r \in \{1, 2, \cdots N\}$ is picked, who minimizes $V_{i_r}(\cdot)$ w.r.t. its local variable x_{i_r}, followed by a dual ascent step for λ_{i_r}. The rest of the agents update their local variables by simply setting them to z. The cluster center then minimizes $L(x, z; \lambda)$ with respect to z. See Algorithm 1 for details. We remark that NESTT-G is related to the popular ADMM method for *convex* optimization [20]. However our particular update schedule (randomly picking (x_i, λ_i) plus deterministic updating z), combined with the special x-step (minimizing an

approximation of $L(\cdot)$ evaluated at a different block variable z) is not known before. These features are critical in our following rate analysis.

Convergence Analysis. To proceed, let us define $r(j)$ as the last iteration in which the jth block is picked before iteration $r+1$. i.e. $r(j) := \max\{t \mid t < r+1, j = i(t)\}$. Define $y_j^r := z^{r(j)}$ if $j \neq i_r$, and $y_{i_r}^r = z^r$. Define the filtration $\mathcal{F}^r$ as the σ-field generated by $\{i(t)\}_{t=1}^{r-1}$.

A few important observations are in order. Combining the (x, z) updates (3.4) – (3.7), we have

$$x_q^{r+1} = z^r - \frac{1}{\alpha_q \eta_q}(\lambda_q^r + \frac{1}{N}\nabla g_q(z^r)),\ \frac{1}{N}\nabla g_q(z^r) + \lambda_q^r + \alpha_q\eta_q(x_q^{r+1} - z^r) = 0,\ with\ q = i_r \tag{3.8a}$$

$$\lambda_{i_r}^{r+1} = -\frac{1}{N}\nabla g_{i_r}(z^r),\ \lambda_j^{r+1} = -\frac{1}{N}\nabla g_j(z^{r(j)}),\ \forall\ j \neq i_r,\ \Rightarrow \lambda_i^{r+1} = -\frac{1}{N}\nabla g_i(y_i^r),\ \forall\ i \tag{3.8b}$$

$$x_j^{r+1} \overset{(3.6)}{=} z^r \overset{(3.8b)}{=} z^r - \frac{1}{\alpha_j\eta_j}(\lambda_j^r + \frac{1}{N}\nabla g_j(z^{r(j)})),\ \forall\ j \neq i_r. \tag{3.8c}$$

The key here is that the dual variables serve as the "memory" for the past gradients of g_i's. To proceed, we first construct a *potential function* using an *upper bound* of $L(x, y; \lambda)$. Note that

$$\frac{1}{N}g_j(x_j^{r+1}) + \langle \lambda_j^r, x_j^{r+1} - z^r\rangle + \frac{\eta_j}{2}\|x_j^{r+1} - z^r\|^2 = \frac{1}{N}g_j(z^r),\ \forall\ j \neq i_r \tag{3.9}$$

$$\begin{aligned}
&\frac{1}{N}g_{i_r}(x_{i_r}^{r+1}) + \langle \lambda_{i_r}^r, x_{i_r}^{r+1} - z^r\rangle + \frac{\eta_i}{2}\|x_{i_r}^{r+1} - z^r\|^2 \\
&\overset{\text{(i)}}{\leq} \frac{1}{N}g_{i_r}(z^r) + \frac{\eta_{i_r} + L_{i_r}/N}{2}\|x_{i_r}^{r+1} - z^r\|^2 \\
&\overset{\text{(ii)}}{=} \frac{1}{N}g_{i_r}(z^r) + \frac{\eta_{i_r} + L_{i_r}/N}{2(\alpha_{i_r}\eta_{i_r})^2}\|1/N(\nabla g_{i_r}(y_{i_r}^{r-1}) - \nabla g_{i_r}(z^r))\|^2
\end{aligned} \tag{3.10}$$

where (i) uses (3.8b) and applies the descent lemma on the function $1/N g_i(\cdot)$; in (ii) we have used (3.5) and (3.8b). Since each i is picked with probability p_i, we have

$$\begin{aligned}
&\mathbb{E}_{i_r}[L(x^{r+1}, z^r; \lambda^r) \mid \mathcal{F}^r] \\
&\leq \sum_{i=1}^N \frac{1}{N}g_i(z^r) + \sum_{i=1}^N \frac{p_i(\eta_i + L_i/N)}{2(\alpha_i\eta_i)^2}\|1/N(\nabla g_i(y_i^{r-1}) - \nabla g_i(z^r))\|^2 + g_0(z^r) + h(z^r) \\
&\leq \sum_{i=1}^N \frac{1}{N}g_i(z^r) + \sum_{i=1}^N \frac{3p_i\eta_i}{(\alpha_i\eta_i)^2}\|1/N(\nabla g_i(y_i^{r-1}) - \nabla g_i(z^r))\|^2 + g_0(z^r) + h(z^r) := Q^r,
\end{aligned}$$

where in the last inequality we have used Assumption [A-(c)]. In the following, we will use $\mathbb{E}_{\mathcal{F}^r}[Q^r]$ as the potential function, and show that it decreases at each iteration.

Lemma 6 *Suppose Assumption A holds, and pick*

$$\alpha_i = p_i = \beta\eta_i, \text{ where } \beta := \frac{1}{\sum_{i=1}^N \eta_i}, \quad \text{and} \quad \eta_i \geq \frac{9L_i}{Np_i}, \quad i = 1, \cdots N. \tag{3.11}$$

Then the following descent estimate holds true for NESTT-G

$$\mathbb{E}[Q^r - Q^{r-1}|\mathcal{F}^{r-1}] \leq -\frac{\sum_{i=1}^N \eta_i}{8}\mathbb{E}_{z^r}\|z^r - z^{r-1}\|^2 - \sum_{i=1}^N \frac{1}{2\eta_i}\|\frac{1}{N}(\nabla g_i(z^{r-1}) - \nabla g_i(y_i^{r-2}))\|^2. \tag{3.12}$$

Sublinear Convergence. Define the optimality gap as the following:

$$\mathbb{E}[G^r] := \mathbb{E}\left[\|\tilde{\nabla}_{1/\beta} f(z^r)\|^2\right] = \frac{1}{\beta^2}\mathbb{E}\left[\|z^r - \text{prox}_h^{1/\beta}[z^r - \beta\nabla(g(z^r) + g_0(z^r))]\|^2\right]. \tag{3.13}$$

Note that when $h, g_0 \equiv 0$, $\mathbb{E}[G^r]$ reduces to $E[\|\nabla g(z^r)\|^2]$. We have the following result.

Theorem 4 *Suppose Assumption A holds, and pick (for $i = 1, \cdots, N$)*

$$\alpha_i = p_i = \frac{\sqrt{L_i/N}}{\sum_{i=1}^N \sqrt{L_i/N}}, \ \eta_i = 3\left(\sum_{i=1}^N \sqrt{L_i/N}\right)\sqrt{L_i/N}, \ \beta = \frac{1}{3(\sum_{i=1}^N \sqrt{L_i/N})^2}. \tag{3.14}$$

Then every limit point generated by NESTT-G is a stationary solution of problem (3.2). Further,

$$1)\ \mathbb{E}[G^m] \leq \frac{80}{3}\Big(\sum_{i=1}^N \sqrt{L_i/N}\Big)^2 \frac{\mathbb{E}[Q^1 - Q^{R+1}]}{R};$$

$$2)\ \mathbb{E}[G^m] + \mathbb{E}\left[\sum_{i=1}^N 3\eta_i^2 \left\|x_i^m - z^{m-1}\right\|^2\right] \leq \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \frac{\mathbb{E}[Q^1 - Q^{R+1}]}{R}.$$

Note that Part (1) is useful in the *centralized* finite-sum minimization setting, as it shows the sublinear convergence of NESTT-G, measured only by the primal optimality gap evaluated at z^r. Meanwhile, part (2) is useful in the *distributed* setting, as it also shows that the expected constraint violation, which measures the consensus among agents, shrinks in the same order. We also comment that the above result suggests that to achieve an ϵ-stationary solution, the NESTT-G requires about $\mathcal{O}\left(\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2/\epsilon\right)$ number of gradient evaluations (for simplicity we have ignored an additive N factor for evaluating the gradient of the entire function at the initial step of the algorithm).

It is interesting to observe that our choice of p_i is proportional to the *square root* of the Lipschitz constant of each component function, rather than to L_i. Because of such choice of the sampling

probability, the derived convergence rate has a mild dependency on N and L_i's. Compared with the conventional gradient-based methods, our scaling can be up to N times better. Detailed discussion and comparison will be given in Section 3.4.

Note that similar sublinear convergence rates can be obtained for the case $\alpha_i = 1$ for all i (with different scaling constants). However due to space limitation, we will not present those results here.

Linear Convergence. In this section we show that the NESTT-G is capable of linear convergence for a family of nonconvex quadratic problems, which has important applications, for example in high-dimensional statistical learning [90]. To proceed, we will assume the following.

B-(a) Each function $g_i(z)$ is a quadratic function of the form $g_i(z) = 1/2 z^T A_i z + \langle b, z \rangle$, where A_i is a symmetric matrix but not necessarily positive semidefinite;

B-(b) The feasible set Z is a closed compact polyhedral set;

B-(c) The nonsmooth function $p(z) = \mu \|z\|_1$, for some $\mu \geq 0$.

Our linear convergence result is based upon certain error bound condition around the stationary solutions set, which has been shown in [94] for smooth quadratic problems and has been extended to including ℓ_1 penalty in [132, Theorem 4]. Due to space limitation the statement of the condition will be given in the supplemental material, along with the proof of the following result.

Theorem 5 *Suppose that Assumptions A, B are satisfied. Then the sequence* $\{\mathbb{E}[Q^{r+1}]\}_{r=1}^{\infty}$ *converges Q-linearly* [1] *to some* $Q^* = f(z^*)$*, where* z^* *is a stationary solution for problem* (4.1)*. That is, there exists a finite* $\bar{r} > 0$*,* $\rho \in (0,1)$ *such that for all* $r \geq \bar{r}$*,* $\mathbb{E}[Q^{r+1} - Q^*] \leq \rho \mathbb{E}[Q^r - Q^*]$.

Linear convergence of this type for problems satisfying Assumption B has been shown for (deterministic) proximal gradient based methods [132, Theorem 2, 3]. To the best of our knowledge, this is the first result that shows the same linear convergence for a stochastic and distributed algorithm.

[1] *A sequence* $\{x^r\}$ *is said to converge Q-linearly to some* $\bar{x}$ *if* $\limsup_r \|x^{r+1} - \bar{x}\| / \|x^r - \bar{x}\| \leq \rho$*, where* $\rho \in (0,1)$ *is some constant; cf [132] and references therein.*

Algorithm 4 NESTT-E Algorithm

1: **for** $r = 1$ **to** R **do**

2: Update z by minimizing the augmented Lagrangian:

$$z^{r+1} = \arg\min_{z} \; L(x^r, z; \lambda^r). \tag{3.15}$$

3: Randomly pick $i_r \in \{1, 2, \cdots N\}$ with probability p_{i_r}:

$$x_{i_r}^{r+1} = \operatorname*{argmin}_{x_{i_r}} U_{i_r}(x_{i_r}, z^{r+1}; \lambda_{i_r}^r); \tag{3.16}$$

$$\lambda_{i_r}^{r+1} = \lambda_{i_r}^r + \alpha_{i_r}\eta_{i_r}\left(x_{i_r}^{r+1} - z^{r+1}\right); \tag{3.17}$$

$$x_j^{r+1} = x_j^r, \quad \lambda_j^{r+1} = \lambda_j^r \quad \forall\, j \neq i_r. \tag{3.18}$$

4: **end for**

5: **Output:** (z^m, x^m, λ^m) where m randomly picked from $\{1, 2, \cdots, R\}$.

3.3 The NESTT-E Algorithm

Algorithm Description. In this section, we present a variant of NESTT-G, which is named NESTT with Exact minimization (NESTT-E). Our motivation is the following. First, in NESTT-G every agent should update its local variable at every iteration [cf. (3.4) or (3.6)]. In practice this may not be possible, for example at any given time a few agents can be in the *sleeping mode* so they cannot perform (3.6). Second, in the distributed setting it has been generally observed (e.g., see [24, Section V]) that performing exact minimization (whenever possible) instead of taking the gradient steps for local problems can significantly speed up the algorithm. The NESTT-E algorithm to be presented in this section is designed to address these issues. To proceed, let us define a new function as follows:

$$U(x, z; \lambda) := \sum_{i=1}^{N} U_i(x_i, z; \lambda_i) := \sum_{i=1}^{N} \left(\frac{1}{N} g_i(x_i) + \langle \lambda_i, x_i - z \rangle + \frac{\alpha_i \eta_i}{2} \|x_i - z\|^2 \right).$$

Note that if $\alpha_i = 1$ for all i, then the $L(x, z; \lambda) = U(x, z; \lambda) + p(z) + h(z)$. The algorithm details are presented in Algorithm 2.

Convergence Analysis. We begin analyzing NESTT-E. The proof technique is quite different from that for NESTT-G, and it is based upon using the expected value of the *Augmented Lagrangian* function as the potential function; see [63, 56, 54]. For the ease of description we define the following

quantities:

$$w := (x, z, \lambda), \quad \beta := \frac{1}{\sum_{i=1}^N \eta_i}, \quad c_i := \frac{L_i^2}{\alpha_i \eta_i N^2} - \frac{\gamma_i}{2} + \frac{1-\alpha_i}{\alpha_i}\frac{L_i}{N}, \quad \alpha := \{\alpha_i\}_{i=1}^N.$$

To measure the optimality of NESTT-E, define the *prox-gradient* of $L(x, z; \lambda)$ as:

$$\tilde{\nabla} L(w) = \left[(z - \text{prox}_h[z - \nabla_z(L(w) - h(z))]); \nabla_{x_1} L(w); \cdots; \nabla_{x_N} L(w)\right] \in \mathbb{R}^{(N+1)d}. \tag{3.19}$$

We define the optimality gap by adding to $\|\tilde{\nabla} L(w)\|^2$ the size of the constraint violation [63]:

$$H(w^r) := \|\tilde{\nabla} L(w^r)\|^2 + \sum_{i=1}^N \frac{L_i^2}{N^2}\|x_i^r - z^r\|^2.$$

It can be verified that $H(w^r) \to 0$ implies that w^r reaches a stationary solution for problem (3.2). We have the following theorem regarding the convergence properties of NESTT-E.

Theorem 6 *Suppose Assumption A holds, and that (η_i, α_i) are chosen such that $c_i < 0$. Then for some constant $\underline{f}$, we have*

$$\mathbb{E}[L(w^r)] \geq \mathbb{E}[L(w^{r+1})] \geq \underline{f} > -\infty, \quad \forall\ r \geq 0.$$

Further, almost surely every limit point of $\{w^r\}$ is a stationary solution of problem (3.2). Finally, for some function of α denoted as $C(\alpha) = \sigma_1(\alpha)/\sigma_2(\alpha)$, we have the following:

$$\mathbb{E}[H(w^m)] \leq \frac{C(\alpha)\mathbb{E}[L(w^1) - L(w^{R+1})]}{R}, \tag{3.20}$$

where $\sigma_1 := \max(\hat{\sigma}_1(\alpha), \tilde{\sigma}_1)$ and $\sigma_2 := \max(\hat{\sigma}_2(\alpha), \tilde{\sigma}_2)$, and these constants are given by

$$\hat{\sigma}_1(\alpha) = \max_i \left\{4\left(\frac{L_i^2}{N^2} + \eta_i^2 + \left(\frac{1}{\alpha_i} - 1\right)^2 \frac{L_i^2}{N^2}\right) + 3\left(\frac{L_i^4}{\alpha_i \eta_i^2 N^4} + \frac{L_i^2}{N^2}\right)\right\},$$

$$\tilde{\sigma}_1 = \sum_{i=1}^N 4\eta_i^2 + (2 + \sum_{i=1}^N \eta_i + L_0)^2 + 3\sum_{i=1}^N \frac{L_i^2}{N^2},$$

$$\hat{\sigma}_2(\alpha) = \max_i \left\{p_i\left(\frac{\gamma_i}{2} - \frac{L_i^2}{N^2 \alpha_i \eta_i} - \frac{1-\alpha_i}{\alpha_i}\frac{L_i}{N}\right)\right\}, \quad \tilde{\sigma}_2 = \frac{\sum_{i=1}^N \eta_i - L_0}{2}.$$

We remark that the above result shows the sublinear convergence of NESTT-E to the set of stationary solutions. Note that $\gamma_i = \eta_i - L_i/N$, to satisfy $c_i < 0$, a simple derivation yields

$$\eta_i > \frac{L_i\left((2-\alpha_i) + \sqrt{(\alpha_i - 2)^2 + 8\alpha_i}\right)}{2N\alpha_i}.$$

Further, the above result characterizes the dependency of the rates on various parameters of the algorithm. For example, to see the effect of α on the convergence rate, let us set $p_i = \frac{L_i}{\sum_{i=1}^N L_i}$, and $\eta_i = 3L_i/N$, and assume $L_0 = 0$, then consider two different choices of α: $\widehat{\alpha}_i = 1$, $\forall\ i$ and $\widetilde{\alpha}_i = 4$, $\forall\ i$. One can easily check that applying these different choices leads to following results:

$$C(\widehat{\alpha}) = 49\sum_{i=1}^N L_i/N, \qquad C(\widetilde{\alpha}) = 28\sum_{i=1}^N L_i/N.$$

The key observation is that increasing α_i's reduces the constant in front of the rate. Hence, we expect that in practice larger α_i's will yield faster convergence.

3.4 Connections and Comparisons with Existing Works

In this section we compare NESTT-G/E with a few existing algorithms in the literature. First, we present a somewhat surprising observation, that NESTT-G takes the same form as some well-known algorithms for *convex* finite-sum problems. To formally state such relation, we show in the following result that NESTT-G in fact admits a compact *primal-only* characterization.

Proposition 1 *The NESTT-G can be written into the following compact form:*

$$z^{r+1} = \arg\min_z\ h(z) + g_0(z) + \frac{1}{2\beta}\|z - u^{r+1}\|^2 \tag{3.21a}$$

$$\text{with}\quad u^{r+1} := z^r - \beta\Big(\frac{1}{N\alpha_{i_r}}(\nabla g_{i_r}(z^r) - \nabla g_{i_r}(y_{i_r}^{r-1})) + \frac{1}{N}\sum_{i=1}^N \nabla g_i(y_i^{r-1})\Big). \tag{3.21b}$$

Based on this observation, the following comments are in order.

(1) Suppose $h \equiv 0$, $g_0 \equiv 0$ and $\alpha_i = 1$, $p_i = 1/N$ for all i. Then (3.21) takes the same form as the SAG presented in [118]. Further, when the component functions g_i's are picked *cyclically* in a Gauss-Seidel manner, the iteration (3.21) takes the same form as the IAG algorithm [19].

(2) Suppose $h \neq 0$ and $g_0 \neq 0$, and $\alpha_i = p_i = 1/N$ for all i. Then (3.21) is the same as the SAGA algorithm [31], which is design for optimizing convex nonsmooth finite sum problems.

Note that SAG/SAGA/IAG are all designed for convex problems. Through the lens of primal-dual splitting, our work shows that they can be generalized to nonconvex nonsmooth problems as well.

Secondly, NESTT-E is related to the proximal version of the nonconvex ADMM [64, Algorithm 2]. However, the introduction of α_i's is new, which can significantly improve the practical performance but complicates the analysis. Further, there has been no counterpart of the sublinear and linear convergence rate analysis for the stochastic version of [64, Algorithm 2].

Thirdly, we note that a recent paper [111] has shown that SAGA works for smooth and unconstrained nonconvex problem. Suppose that $h \equiv 0$, $g_0 \neq 0$, $L_i = L_j$, $\forall\ i, j$ and $\alpha_i = p_i = 1/N$, the authors show that SAGA achieves ϵ-stationarity using $\mathcal{O}(N^{2/3}(\sum_{i=1}^N L_i/N)/\epsilon)$ gradient evaluations. Compared with GD, which achieves ϵ-stationarity using $\mathcal{O}(\sum_{i=1}^N L_i/\epsilon)$ gradient evaluations in the worse case (in the sense that $\sum_{i=1}^N L_i/N = L$), the rate in [111] is $\mathcal{O}(N^{1/3})$ times better. However, the algorithm in [111] is different from NESTT-G in two aspects: 1) it does not generalize to the nonsmooth constrained problem (4.1); 2) it samples two component functions at each iteration, while NESTT-G only samples once. Further, the analysis and the scaling are derived for the case of uniform L_i's, therefore it is not clear how the algorithm and the rates can be adapted for the non-uniform case. On the other hand, our NESTT works for the general nonsmooth constrained setting. The non-uniform sampling used in NESTT-G is well-suited for problems with non-uniform L_i's, and our scaling can be up to N times better than GD (or its proximal version) in the worst case. Note that problems with non-uniform L_i's for the component functions are common in applications such as sparse optimization and signal processing. For example in LASSO problem the data matrix is often normalized by feature (or "column-normalized" [103]), therefore the ℓ_2 norm of each row of the data matrix (which corresponds to the Lipschitz constant for each component function) can be dramatically different.

In Table 3.1 we list the comparison of the number of gradient evaluations for NESTT-G and GD, in the worst case (in the sense that $\sum_{i=1}^N L_i/N = L$). For simplicity, we omitted an additive constant of $\mathcal{O}(N)$ for computing the initial gradients.

3.5 Numerical Results

In this section we evaluate the performance of NESTT. Consider the high dimensional regression problem with noisy observation [90], where M observations are generated by $y = X\nu + \epsilon$. Here $y \in \mathbb{R}^M$ is the observed data sample; $X \in \mathbb{R}^{M\times P}$ is the covariate matrix; $\nu \in \mathbb{R}^P$ is the ground truth, and $\epsilon \in \mathbb{R}^M$ is the noise. Suppose that the covariate matrix is not perfectly known, i.e., we observe $A = X + W$ where $W \in \mathbb{R}^{M\times P}$ is the noise matrix with known covariance matrix Σ_W. Let us define $\hat{\Gamma} := 1/M(A^\top A) - \Sigma_W$, and $\hat{\gamma} := 1/M(A^\top y)$. To estimate the ground truth ν, let us consider the following (nonconvex) optimization problem posed in [90, problem (2.4)] (where $R > 0$ controls sparsity):

$$\min_z \; z^\top \hat{\Gamma} z - \hat{\gamma} z \quad \text{s.t.} \quad \|z\|_1 \le R. \tag{3.22}$$

Due to the existence of noise, $\hat{\Gamma}$ is not positive semidefinite hence the problem is not convex. Note that this problem satisfies Assumption A– B, then by Theorem 5 NESTT-G converges Q-linearly.

To test the performance of the proposed algorithm, we generate the problem following similar setups as [90]. Let $X = (X_1; \cdots, X_N) \in \mathbb{R}^{M\times P}$ with $\sum_i N_i = M$ and each $X_i \in \mathbb{R}^{N_i\times P}$ corresponds to N_i data points, and it is generated from i.i.d Gaussian. Here N_i represents the size of each mini-batch of samples. Generate the observations $y_i = X_i \times \nu^* + \epsilon_i \in \mathbb{R}^{N_i}$, where ν^* is a K-sparse vector to be estimated, and $\epsilon_i \in \mathbb{R}^{N_i}$ is the random noise. Let $W = [W_1; \cdots; W_N]$, with $W_i \in \mathbb{R}^{N_i\times P}$ generated with i.i.d Gaussian. Therefore we have $z^\top \hat{\Gamma} z = \frac{1}{N}\sum_{i=1}^N \frac{N}{M} z^\top \left(X_i^\top X_i - W_i^\top W_i\right) z$. We set $M = 100,000$, $P = 5000$, $N = 50$, $K = 22 \approx \sqrt{P}$,and $R = \|\nu^*\|_1$. We implement NESTT-G/E, the SGD, and the nonconvex SAGA proposed in [111] with stepsize $\beta = \frac{1}{3L_{\max}N^{2/3}}$ (with $L_{\max} := \max_i L_i$). Note that the SAGA proposed in [111] *only* works for the unconstrained problems with uniform L_i, therefore when applied to (3.22) it is *not* guaranteed to converge. Here we only include it for comparison purposes.

In Fig. 3.2 we compare different algorithms in terms of the gap $\|\tilde{\nabla}_{1/\beta} f(z^r)\|^2$. In the left figure we consider the problem with $N_i = N_j$ for all i, j, and we show performance of the proposed algorithms with uniform sampling (i.e., the probability of picking ith block is $p_i = 1/N$). On the

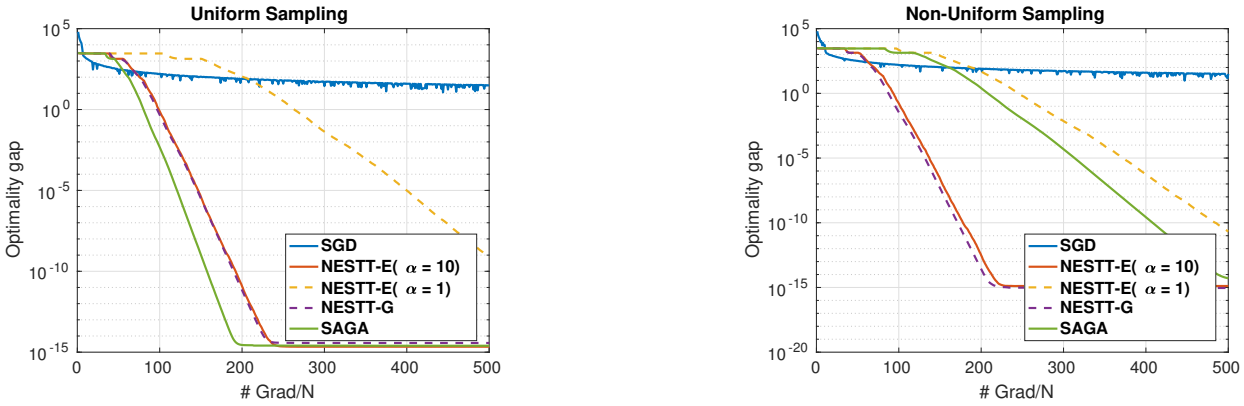


Figure 3.2: Comparison of NESTT-G/E, SAGA, SGD on problem (3.22)

right one we consider problems in which approximately half of the component functions have twice the size of L_i's as the rest, and consider the non-uniform sampling ($p_i = \sqrt{L_i/N}/\sum_{i=1}^{N}\sqrt{L_i/N}$). Clearly in both cases the proposed algorithms perform quite well. Furthermore, it is clear that the NESTT-E performs well with large $\alpha := \{\alpha_i\}_{i=1}^{N}$, which confirms our theoretical rate analysis. Also it is worth mentioning that when the N_i's are non-uniform, the proposed algorithms [NESTT-G and NESTT-E (with $\alpha = 10$)] significantly outperform SAGA and SGD. In Table 3.2 we further compare different algorithms when changing the number of component functions (i.e., the number of mini-batches N) while the rest of the setup is as above. We run each algorithm with 100 passes over the dataset. Similarly as before, our algorithms perform well, while SAGA seems to be sensitive to the uniformity of the size of the mini-batch [note that there is no convergence guarantee for SAGA applied to the nonconvex constrained problem (3.22)].

3.6 Appendix. Proofs

Some Key Properties of NESTT-G

To facilitate the following derivation, in this section we collect some key properties of NESTT-G.

First, from the optimality condition of the x update we have

$$x_{i_r}^{r+1} = z^r - \frac{1}{\alpha_{i_r}\eta_{i_r}}\left(\lambda_{i_r}^r + \frac{1}{N}\nabla g_{i_r}(z^r)\right), \tag{3.23a}$$

$$x_j^{r+1} \overset{(3.6)}{=} z^r \overset{(3.8b)}{=} z^r - \frac{1}{\alpha_j\eta_j}(\lambda_j^r + \frac{1}{N}\nabla g_j(z^{r(j)})),\ \forall\ j \neq i_r. \tag{3.23b}$$

Then using the update scheme of the λ we can further obtain

$$\lambda_{i_r}^{r+1} = -\frac{1}{N}\nabla g_{i_r}(z^r), \tag{3.24a}$$

$$\lambda_j^{r+1} = -\frac{1}{N}\nabla g_j(z^{r(j)}),\ \forall\ j \neq i_r. \tag{3.24b}$$

Therefore, using the definition of y_i^r we have the following compact forms

$$\lambda_i^{r+1} = -\frac{1}{N}\nabla g_i(y_i^r),\ i = 1, \cdots, N. \tag{3.25}$$

$$x_i^{r+1} = z^r - \frac{1}{\alpha_i\eta_i}\left(\lambda_i^r + \frac{1}{N}\nabla g_i(y_i^r)\right),\ i = 1, \cdots, N. \tag{3.26}$$

Second, let us look at the optimality condition for the z update. The z-update (3.7) is given by

$$\begin{aligned} z^{r+1} &= \arg\min_z\ L(\{x_i^{r+1}\}, z; \lambda^r) \\ &= \arg\min_z \sum_{i=1}^N \left(\langle \lambda_i^r, x_i^{r+1} - z\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - z\|^2\right) + g_0(z) + h(z). \end{aligned} \tag{3.27}$$

Note that this problem is strongly convex because we have assumed that $\sum_{i=1}\eta_i > 3L_0$; cf. Assumption [A-(c)].

Let us define

$$\begin{aligned}
u^{r+1} &:= \frac{\sum_{i=1}^N \eta_i x_i^{r+1} + \sum_{i=1}^N \lambda_i^r}{\sum_{i=1}^N \eta_i} \\
&= \frac{\sum_{i=1}^N \eta_i z^r - \eta_{i_r}(z^r - x_{i_r}^{r+1})}{\sum_{i=1}^N \eta_i} + \frac{\sum_{i=1}^N \lambda_i^r}{\sum_{i=1}^N \eta_i} \\
&\overset{(3.23a)}{=} \frac{\sum_{i=1}^N \eta_i z^r - \frac{\eta_{i_r}}{\alpha_{i_r}\eta_{i_r}}(\lambda_{i_r}^r + 1/N\nabla g_{i_r}(z^r))}{\sum_{i=1}^N \eta_i} + \frac{\sum_{i=1}^N \lambda_i^r}{\sum_{i=1}^N \eta_i} \\
&\overset{(5.100)}{=} z^r - \frac{\frac{1}{\alpha_{i_r}}(-\nabla g_{i_r}(y_{i_r}^{r-1}) + \nabla g_{i_r}(z^r))}{N\sum_{i=1}^N \eta_i} - \frac{\sum_{i=1}^N \nabla g_i(y_i^{r-1})}{N\sum_{i=1}^N \eta_i} \\
&\overset{\text{(i)}}{=} z^r - \frac{\beta}{N\alpha_{i_r}}(-\nabla g_{i_r}(y_{i_r}^{r-1}) + \nabla g_{i_r}(z^r)) - \frac{\beta\sum_{i=1}^N \nabla g_i(y_i^{r-1})}{N} \qquad (3.28) \\
&\overset{\text{(ii)}}{:=} z^r - \beta v_{i_r}^{r+1} \qquad (3.29)
\end{aligned}$$

where in (i) we have defined $\beta := 1/\sum_{i=1}^N \eta_i$; in (ii) we have defined

$$v_{i_r}^{r+1} := \frac{1}{N}\sum_{i=1}^N \nabla g_i(y_i^{r-1}) + \frac{1}{\alpha_{i_r}}\left(-\frac{1}{N}\nabla g_{i_r}(y_{i_r}^{r-1}) + \frac{1}{N}\nabla g_{i_r}(z^r)\right). \qquad (3.30)$$

Clearly if we pick $\alpha_i = p_i$ for all i, then we have

$$\mathbb{E}_{i_r}[u^{r+1} \mid \mathcal{F}^r] = z^r - \frac{\beta}{N}\sum_{i=1}^N \nabla g_i(z^r). \qquad (3.31)$$

Using the definition of u^{r+1}, it is easy to check that

$$z^{r+1} = \arg\min_z \frac{1}{2\beta}\|z - u^{r+1}\|^2 + h(z) + g_0(z) = \text{prox}_h^{1/\beta}[u^{r+1} - \beta\nabla g_0(z^{r+1})]. \qquad (3.32)$$

The optimality condition for the z subproblem is given by:

$$z^{r+1} - u^{r+1} + \beta\nabla g_0(z^{r+1}) + \beta\xi^{r+1} = 0 \qquad (3.33)$$

where, $\xi^{r+1} \in \partial h(z^{r+1})$ is a subgradient of $h(z^{r+1})$. Using the definition of v_{i_r} in (5.107), we obtain

$$z^{r+1} = z^r - \beta(v_{i_r}^{r+1} + \nabla g_0(z^{r+1}) + \xi^{r+1}). \qquad (3.34)$$

Third, if $\alpha_i = p_i$, then we have:

$$\begin{aligned}
&\mathbb{E}_{i_r}\left[\left\|-\frac{\lambda_{i_r}^r + 1/N\nabla g_{i_r}(z^r)}{\alpha_{i_r}} + \frac{1}{N}\sum_{i=1}^N \nabla g_i(z^r) - \sum_{i=1}^N \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2\right] \\
&\overset{(a)}{=} \mathrm{Var}\left[-\frac{\lambda_{i_r}^r + 1/N\nabla g_{i_r}(z^r)}{\alpha_{i_r}}\right] \\
&\overset{(b)}{\leq} \sum_{i=1}^N \frac{1}{\alpha_i}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2,
\end{aligned} \tag{3.35}$$

where (a) is true because whenever $\alpha_i = p_i$ for all i, then

$$\mathbb{E}_{i_r}\left[-\frac{\lambda_{i_r}^r + 1/N\nabla g_{i_r}(z^r)}{\alpha_{i_r}}\right] = \frac{1}{N}\sum_{i=1}^N \nabla g_i(z^r) - \sum_{i=1}^N \frac{1}{N}\nabla g_i(y_i^{r-1});$$

The inequality in (b) is true because for a random variable x we have $\mathrm{Var}(x) \leq \mathbb{E}[x^2]$.

3.6.1 Proof of Lemma 6

Step 1). Using the definition of potential function Q^r, we have:

$$\begin{aligned}
&\mathbb{E}[Q^r - Q^{r-1} \mid \mathcal{F}^{r-1}] \\
&= \mathbb{E}\left[\sum_{i=1}^N \frac{1}{N}\left(g_i(z^r) - g_i(z^{r-1})\right) + g_0(z^r) - g_0(z^{r-1}) + h(z^r) - h(z^{r-1}) \mid \mathcal{F}^{r-1}\right] \\
&+ \mathbb{E}\left[\sum_{i=1}^N \frac{3p_i}{\alpha_i^2\eta_i}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2 - \frac{3p_i}{\alpha_i^2\eta_i}\left\|\frac{1}{N}\nabla g_i(z^{r-1}) - \frac{1}{N}\nabla g_i(y_i^{r-2})\right\|^2 \mid \mathcal{F}^{r-1}\right].
\end{aligned} \tag{3.36}$$

Step 2). The first term in (5.116) can be bounded as follows (omitting the subscript $\mathcal{F}^r$).

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\frac{1}{N}\left(g_i(z^r)-g_i(z^{r-1})\right)+g_0(z^r)-g_0(z^{r-1})+h(z^r)-h(z^{r-1})\mid\mathcal{F}^{r-1}\right]\\
&\overset{\text{(i)}}{\leq}\mathbb{E}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\langle\nabla g_i(z^{r-1}),z^r-z^{r-1}\rangle+\langle\nabla g_0(z^{r-1}),z^r-z^{r-1}\rangle\\
&\qquad+\langle\xi^r,z^r-z^{r-1}\rangle+\frac{\sum_{i=1}^{N}L_i/N+L_0}{2}\|z^r-z^{r-1}\|^2\mid\mathcal{F}^{r-1}\Bigg]\\
&\overset{\text{(ii)}}{=}\mathbb{E}\left[\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(z^{r-1})+\xi^r+\nabla g_0(z^r)+\frac{1}{\beta}(z^r-z^{r-1}),z^r-z^{r-1}\right\rangle\mid\mathcal{F}^{r-1}\right]\\
&\qquad-\left(\frac{1}{\beta}-\frac{\sum_{i=1}^{N}L_i/N+3L_0}{2}\right)\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2\\
&\overset{(3.34)}{=}\mathbb{E}\left[\left\langle\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(z^{r-1})-v^r_{i(r-1)},z^r-z^{r-1}\right\rangle\mid\mathcal{F}^{r-1}\right]\\
&\quad-\left(\frac{1}{\beta}-\frac{\sum_{i=1}^{N}L_i/N+3L_0}{2}\right)\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2\\
&\overset{\text{(iii)}}{\leq}\frac{1}{2\ell_1}\mathbb{E}\left[\left\|1/N\sum_{i=1}^{N}\nabla g_i(z^{r-1})-v^r_{i(r-1)}\right\|^2\mid\mathcal{F}^{r-1}\right]+\frac{\ell_1}{2}\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2\\
&\qquad-\left(\frac{1}{\beta}-\frac{\sum_{i=1}^{N}L_i/N+3L_0}{2}\right)\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2
\end{aligned}
\tag{3.37}
$$

where in (i) we have used the Lipschitz continuity of the gradients of g_i's as well as the convexity of h; in (ii) we have used the fact that

$$
\langle\nabla g_0(z^{r-1}),z^r-z^{r-1}\rangle\leq\langle\nabla g_0(z^r),z^r-z^{r-1}\rangle+L_0\|z^r-z^{r-1}\|^2; \tag{3.38}
$$

in (iii) we have applied the Young's inequality for some $\ell_1>0$.

Choosing $\ell_1=\frac{1}{2\beta}$, we have:

$$
\begin{aligned}
&\frac{1}{2\ell_1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(z^{r-1})-v^r_{i(r-1)}\right\|^2\\
&\overset{(5.107)}{=}\beta\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla g_i(z^{r-1})-\frac{\lambda^{r-1}_{i(r-1)}+1/N\nabla g_{i(r-1)}(z^{r-1})}{\alpha_{i(r-1)}}-\sum_{i=1}^{N}\frac{1}{N}\nabla g_i(y_i^{r-2})\right\|^2\right]\\
&\overset{(3.35)}{\leq}\beta\sum_{i=1}^{N}\frac{1}{\alpha_i}\left\|\frac{1}{N}\nabla g_i(z^{r-1})-\frac{1}{N}\nabla g_i(y_i^{r-2})\right\|^2.
\end{aligned}
$$

Overall we have the following bound for the first term in (5.116):

$$\mathbb{E}\left[\sum_{i=1}^{N}\frac{1}{N}\left(g_i(z^r)-g_i(z^{r-1})\right)+g_0(z^r)-g_0(z^{r-1})+h(z^r)-h(z^{r-1})\mid\mathcal{F}^{r-1}\right] \tag{3.39}$$
$$\leq\sum_{i=1}^{N}\frac{\beta}{\alpha_i}\left\|\frac{1}{N}\nabla g_i(z^{r-1})-\frac{1}{N}\nabla g_i(y_i^{r-2})\right\|^2-\left(\frac{3}{4\beta}-\frac{\sum_{i=1}^{N}L_i/N+3L_0}{2}\right)\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2.$$

Step 3). We bound the second term in (5.116) in the following way:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla g_i(z^r)-\nabla g_i(y_i^{r-1})\|^2\mid\mathcal{F}^{r-1}\right]\\
&=\mathbb{E}\left[\|\nabla g_i(z^r)-\nabla g_i(y_i^{r-1})+\nabla g_i(z^{r-1})-\nabla g_i(z^{r-1})\|^2\mid\mathcal{F}^{r-1}\right]\\
&\overset{\text{(i)}}{\leq}(1+\xi_i)\mathbb{E}_{z^r}\|\nabla g_i(z^r)-\nabla g_i(z^{r-1})\|^2+\left(1+\frac{1}{\xi_i}\right)\mathbb{E}_{y_i^{r-1}}\|\nabla g_i(y_i^{r-1})-\nabla g_i(z^{r-1})\|^2\\
&\overset{\text{(ii)}}{=}(1+\xi_i)\mathbb{E}_{z^r}\|\nabla g_i(z^r)-\nabla g_i(z^{r-1})\|^2+(1-p_i)\left(1+\frac{1}{\xi_i}\right)\|\nabla g_i(y_i^{r-2})-\nabla g_i(z^{r-1})\|^2
\end{aligned} \tag{3.40}$$

where in (i) we have used the fact that the randomness of z^{r-1} comes from i_{r-2}, so fixing $\mathcal{F}^{r-1}$, z^{r-1} is deterministic; we have also applied the following inequality:

$$(a+b)^2\leq(1+\xi)a^2+(1+\frac{1}{\xi})b^2\quad\forall\,\xi>0.$$

The equality (ii) is true because the randomness of y_i^{r-1} comes from i_{r-1}, and for each i there is a probability p_i such that x_i^r is updated, so that $\nabla g_i(y_i^{r-1})=\nabla g_i(z^{r-1})$, otherwise x_i is not updated so that $\nabla g_i(y_i^{r-1})=\nabla g_i(y_i^{r-2})$.

Step 4). Applying (3.40) and set $\alpha_i=p_i$, the second part of (5.116) can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\frac{3p_i}{\alpha_i^2\eta_i}\left\|\frac{1}{N}\nabla g_i(z^r)-\frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2-\frac{3p_i}{\alpha_i^2\eta_i}\left\|\frac{1}{N}\nabla g_i(z^{r-1})-\frac{1}{N}\nabla g_i(y_i^{r-2})\right\|^2\mid\mathcal{F}^{r-1}\right]\\
&\leq\sum_{i=1}^{N}\frac{3L_i^2}{\alpha_i\eta_iN^2}\left(1+\xi_i\right)\mathbb{E}_{z^r}\|z^r-z^{r-1}\|^2\\
&+\frac{3}{\alpha_i\eta_i}\left((1-p_i)(1+\frac{1}{\xi_i})-1\right)\left\|\frac{1}{N}\nabla g_i(y_i^{r-2})-\frac{1}{N}\nabla g_i(z^{r-1})\right\|^2.
\end{aligned} \tag{3.41}$$

Combining (5.117) and (5.120) eventually we have

$$\begin{aligned}
&\mathbb{E}[Q^r - Q^{r-1} \mid \mathcal{F}^r] \\
&\leq \sum_{i=1}^{N} \left\{ \frac{\beta}{\alpha_i} + \frac{3}{\alpha_i \eta_i} \left((1-p_i)(1+\frac{1}{\xi_i}) - 1 \right) \right\} \left\| \frac{1}{N} \nabla g_i(z^{r-1}) - \frac{1}{N} \nabla g_i(y_i^{r-2}) \right\|^2 \\
&+ \left\{ -\frac{3}{4\beta} + \frac{\sum_{i=1}^{N} L_i/N + 3L_0}{2} + \sum_{i=1}^{N} \frac{3L_i^2}{\alpha_i \eta_i N^2} (1+\xi_i) \right\} \mathbb{E}_{z^r} \|z^r - z^{r-1}\|^2. \qquad (3.42)
\end{aligned}$$

Let us define $\{\tilde{c}_i\}$ and $\hat{c}$ as following:

$$\begin{aligned}
\tilde{c}_i &= \frac{\beta}{\alpha_i} + \frac{3}{\alpha_i \eta_i} \left((1-p_i)(1+\frac{1}{\xi_i}) - 1 \right) \\
\hat{c} &= -\frac{3}{4\beta} + \frac{\sum_{i=1}^{N} L_i/N + 3L_0}{2} + \sum_{i=1}^{N} \frac{3L_i^2}{\alpha_i \eta_i N^2} (1+\xi_i).
\end{aligned}$$

In order to prove the lemma it is enough to show that $\tilde{c}_i < -\frac{1}{2\eta_i}\ \forall\ i$, and $\hat{c} < -\sum_{i=1}^{N} \frac{\eta_i}{8}$. Let us pick

$$\alpha_i = p_i,\ \xi_i = \frac{2}{p_i},\ p_i = \frac{\eta_i}{\sum_{i=1}^{N} \eta_i}. \qquad (3.43)$$

Recall that $\beta = \frac{1}{\sum_{i=1}^{N} \eta_i}$.These values yield the following

$$\tilde{c}_i = \frac{1}{\eta_i} - \frac{3}{\eta_i} \left(\frac{p_i + 1}{2} \right) \leq \frac{1}{\eta_i} - \frac{3}{2\eta_i} = -\frac{1}{2\eta_i} < 0.$$

To show that $\hat{c} \leq -\sum_{i=1}^{N} \frac{\eta_i}{8}$ let us assume that $\eta_i = d_i L_i$ for some $d_i > 0$. Note that by assumption we have

$$\sum_{i=1}^{N} \eta_i \geq 3L_0.$$

Therefore we have the following expression for $\hat{c}$:

$$\hat{c} \leq -\sum_{i=1}^{N} \frac{1}{4} d_i L_i + \frac{L_i}{2N} + \frac{3L_i}{p_i d_i N^2} \left(1 + \frac{2}{p_i} \right) < \sum_{i=1}^{N} \frac{L_i}{d_i} \left(-\frac{1}{4} d_i^2 + \frac{d_i}{2N} + \frac{9}{p_i^2 N^2} \right).$$

As a result, to have $\hat{c} < -\sum_{i=1}^{N} \frac{\eta_i}{8}$, we need

$$\frac{L_i}{d_i} \left(\frac{1}{4} d_i^2 - \frac{d_i}{2N} - \frac{9}{p_i^2 N^2} \right) \geq \frac{d_i L_i}{8}, \quad \forall\ i. \qquad (3.44)$$

Or equivalently

$$\frac{1}{8}d_i^2 - \frac{d_i}{2N} - \frac{9}{p_i^2 N^2} \geq 0, \quad \forall\, i. \tag{3.45}$$

By finding the root of the above quadratic inequality, we need $d_i \geq \frac{9}{Np_i}$, which is equivalent to choosing the following parameters

$$\eta_i \geq \frac{9L_i}{Np_i}. \tag{3.46}$$

The lemma is proved. **Q.E.D.**

3.6.2 Proof of Theorem 4

First, using the fact that $f(z)$ is lower bounded [cf. Assumption A-(a)], it is easy to verify that $\{Q^r\}$ is a bounded sequence. Denote its lower bound to be $\underline{Q}$. From Lemma 6, it is clear that $\{Q^r - \underline{Q}\}$ is a nonnegative supermartingale. Apply the Supermartigale Convergence Theorem [15, Proposition 4.2] we conclude that $\{Q^r\}$ converges almost surely (a.s.), and that

$$\left\|\nabla g_i(z^{r-1}) - \nabla g_i(y_i^{r-2})\right\|^2 \to 0, \quad \mathbb{E}_{z^r}\|z^r - z^{r-1}\| \to 0, \quad \text{a.s.}, \quad \forall\, i. \tag{3.47}$$

The first inequality implies that $\|\lambda_{i_r}^r - \lambda_{i_r}^{r-1}\| \to 0$. Combining this with equation (3.5) yields $\|x_{i_r}^r - z^{r-1}\| \to 0$, which further implies that $\|z^r - z^{r-1}\| \to 0$. By utilizing (3.8b) – (3.8c), we can conclude that

$$\|x_i^r - x_i^{r-1}\| \to 0, \quad \|\lambda_i^r - \lambda_i^{r-1}\| \to 0, \quad \text{a.s.}, \quad \forall\, i. \tag{3.48}$$

That is, almost surely the successive differences of all the primal and dual variables go to zero. Then it is easy to show that every limit point of the sequence (x^r, z^r, λ^r) converge to a stationary solution of problem (3.2) (for example, see the argument in [64, Theorem 2.1]. Here we omit the full proof.

Part 1). We bound the gap in the following way (where the expectation is taking over the nature history of the algorithm):

$$\begin{aligned}
&\mathbb{E}\left[\|z^r - \mathrm{prox}_h^{1/\beta}[z^r - \beta\nabla(g(z^r) + g_0(z^r))]\|^2\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\|z^r - z^{r+1} + \mathrm{prox}_h^{1/\beta}[u^{r+1} - \beta\nabla g_0(z^{r+1})] - \mathrm{prox}_h^{1/\beta}[z^r - \beta\nabla(g(z^r) + g_0(z^r))]\|^2\right] \\
&\overset{(b)}{\leq} 3\mathbb{E}\|z^r - z^{r+1}\|^2 + 3\mathbb{E}\|u^{r+1} - z^r + \beta\nabla g(z^r)\|^2 + 3L_0^2\beta^2\|z^{r+1} - z^r\|^2 \\
&\overset{(c)}{\leq} \frac{10}{3}\mathbb{E}\|z^r - z^{r+1}\|^2 + 3\beta^2\mathbb{E}\left[\|\nabla g(z^r) - \frac{\lambda_{i_r}^r + 1/N\nabla g_{i_r}(z^r)}{\alpha_{i_r}} - \sum_{i=1}^N 1/N\nabla g_i(y_i^{r-1})\|^2\right] \\
&\overset{(3.35)}{\leq} \frac{10}{3}\mathbb{E}\|z^r - z^{r+1}\|^2 + 3\beta^2\sum_{i=1}^N \frac{1}{\alpha_i}\mathbb{E}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2 \\
&\leq \frac{10}{3}\mathbb{E}\|z^r - z^{r+1}\|^2 + 3\sum_{i=1}^N \frac{\beta}{\eta_i}\mathbb{E}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2
\end{aligned} \tag{3.49}$$

where (a) is due to (3.32); (b) is true due to the nonexpansivness of the prox operator, and the Cauchy-Swartz inequality; in (c) we have used the definition of u in (5.108) and the fact that $3L_0 \leq \sum_{i=1}^N \eta_i = \frac{1}{\beta}$ [cf. Assumption A-(c)]. In the last inequality we have applied (3.43), which implies that

$$\frac{\beta}{\alpha_i} = \frac{1}{p_i\sum_{j=1}^N \eta_j} = \frac{1}{\eta_i}. \tag{3.50}$$

Note that η_i's has to satisfy (3.46). Let us follow (3.11) and choose

$$\eta_i = \frac{9L_i}{p_i N} = \frac{9\sum_{j=1}^N \eta_j}{N\eta_i}L_i.$$

We have

$$\eta_i = \sqrt{9L_i/N\sum_{j=1}^N \eta_j} = \sqrt{9L_i/N}\sqrt{\sum_{j=1}^N \eta_j} \tag{3.51}$$

Summing i from 1 to N we have

$$\sqrt{\sum_{i=1}^N \eta_i} = \sum_{i=1}^N \sqrt{9L_i/N} \tag{3.52}$$

Then we conclude that

$$\frac{1}{\beta} = \sum_{i=1}^N \eta_i = \left(\sum_{i=1}^N \sqrt{9L_i/N}\right)^2. \tag{3.53}$$

So plugging the expression of β into (3.50) and (3.51), we conclude

$$\alpha_i = p_i = \frac{\sqrt{L_i/N}}{\sum_{i=1}^N \sqrt{L_i/N}}, \quad \eta_i = \sqrt{9L_i/N}\sum_{j=1}^N \sqrt{9L_j/N}. \tag{3.54}$$

After plugging in the above inequity into (3.13), we obtain:

$$\begin{aligned}\mathbb{E}[G^r] &\overset{(5.126)}{\leq} \frac{10}{3\beta^2}\mathbb{E}\|z^r - z^{r+1}\|^2 + \sum_{i=1}^N \frac{3}{\beta\eta_i}\mathbb{E}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2 \\ &\overset{(3.12)}{\leq} \frac{80}{3\beta}\mathbb{E}[Q^r - Q^{r+1}] = \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \mathbb{E}[Q^r - Q^{r+1}]\end{aligned} \tag{3.55}$$

If we sum both sides over $r = 1, \cdots, R$, we obtain:

$$\sum_{r=1}^R \mathbb{E}[G^r] \leq \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \mathbb{E}[Q^1 - Q^{R+1}].$$

Using the definition of z^m, we have

$$\mathbb{E}[G^m] = \mathbb{E}_{\mathcal{F}^r}\left[\mathbb{E}_m[G^m \mid \mathcal{F}^r]\right] = 1/R\sum_{r=1}^R \mathbb{E}_{\mathcal{F}^r}[G^r].$$

Therefore, we can finally conclude that:

$$\mathbb{E}[G^m] \leq \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \frac{\mathbb{E}[Q^1 - Q^{R+1}]}{R} \tag{3.56}$$

which proves the first part.

Part 2). In order to prove the second part let us recycle inequality in (3.55) and write

$$\begin{aligned}&\mathbb{E}\left[G^r + \sum_{i=1}^N \frac{3}{\beta\eta_i}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2\right] \\ &\leq \frac{10}{3\beta^2}\mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^N \frac{6}{\beta\eta_i}\mathbb{E}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2 \\ &\leq \frac{80}{3\beta}\mathbb{E}[Q^r - Q^{r+1}] = 48\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \mathbb{E}[Q^r - Q^{r+1}].\end{aligned}$$

Also note that

$$\mathbb{E}_{x^r}\left[\left\|x_i^{r+1} - z^r\right\|^2 \mid \mathcal{F}^r\right] = \sum_{i=1}^N \frac{1}{\alpha_i\eta_i^2}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2 \tag{3.57}$$

Combining the above two inequalities, we conclude

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}^r}[G^r] + \mathbb{E}_{\mathcal{F}^r}\left[\sum_{i=1}^N 3\eta_i^2 \left\|x_i^{r+1} - z^r\right\|^2\right] \\
&= \mathbb{E}_{\mathcal{F}^r}[G^r] + \mathbb{E}_{\mathcal{F}^r}\left[\sum_{i=1}^N \frac{3\eta_i\alpha_i}{\beta} \left\|x_i^{r+1} - z^r\right\|^2\right] \\
&= \mathbb{E}\left[G^r + \sum_{i=1}^N \frac{3}{\beta\eta_i}\left\|\frac{1}{N}\nabla g_i(z^r) - \frac{1}{N}\nabla g_i(y_i^{r-1})\right\|^2\right] \\
&\leq \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \mathbb{E}_{\mathcal{F}^r}[Q^r - Q^{r+1}] \qquad (3.58)
\end{aligned}$$

where in the first equality we have used the relation $\frac{\alpha_i}{\beta} = \eta_i$ [cf. (3.50)]. Using a similar argument as in first part, we conclude that

$$\mathbb{E}[G^m] + \mathbb{E}\left[\sum_{i=1}^N 3\eta_i^2 \left\|x_i^m - z^{m-1}\right\|^2\right] \leq \frac{80}{3}\left(\sum_{i=1}^N \sqrt{L_i/N}\right)^2 \frac{\mathbb{E}[Q^1 - Q^{R+1}]}{R}. \qquad (3.59)$$

This completes the proof. **Q.E.D.**

3.6.3 Proof of Theorem 5

We first need the following lemma, which characterizes certain error bound condition around the stationary solution set.

Lemma 7 *Suppose Assumptions A and B hold. Let Z^* denotes the set of stationary solutions of problem (4.1), and $dist(z, Z^*) := \min_{u\in Z^*}\|z-u\|$. Then we have the following*

1. *(**Error Bound Condition**) For any $\xi \geq \min_z f(z)$, exists a positive scalar τ such that the following error bound holds*

$$dist(z, Z^*) \leq \tau\|\tilde{\nabla}_{1/\beta} f(z)\| \qquad (3.60)$$

for all $z \in (Z \cap \text{dom } h)$ and $z \in \{z : f(z) \leq \xi\}$.

2. *(**Separation of Isocost Surfaces**) There exists a scalar $\delta > 0$ such that*

$$\|z - v\| \geq \delta \quad \text{whenever} \quad z \in Z^*, v \in Z^*, f(z) \neq f(v). \qquad (3.61)$$

The first statement holds true largely due to [132, Theorem 4], and the second statement holds true due to [93, Lemma 3.1]; see detailed discussion after [132, Assumption 2]. Here the only difference with the statement [132, Theorem 4] is that the error bound condition (3.60) holds true *globally*. This is by the assumption that Z is a compact set. Below we provide a brief argument.

From [R3, Theorem 4], we know that when Assumption B is satisfied, we have that for any $\xi \geq \min_z f(z)$, there exists scalars τ and ϵ such that the following error bound holds

$$\text{dist}\,(z, Z^*) \leq \tau \|\tilde{\nabla}_{1/\beta} f(z)\|, \quad \text{whenever } \|\tilde{\nabla}_{1/\beta} f(z)\| \leq \epsilon,\ f(z) \leq \xi. \tag{3.62}$$

To argue that when Z is compact, the above error bound is independent of ϵ, we use the following two steps: (1) for all $z \in Z \cap \text{dom}(h)$ such that $\|\tilde{\nabla}_{1/\beta} f(z)\| \leq \delta$, it is clear that the error bound (3.60) holds true; (2) for all $z \in Z \cap \text{dom}(h)$ such that $\|\tilde{\nabla}_{1/\beta} f(z)\| \geq \delta$, the ratio $\frac{\text{dist}\,(z,Z^*)}{\|\tilde{\nabla}_{1/\beta} f(z)\|}$ is a continuous function and well defined over the compact set $Z \cap \text{dom}(h) \cap \left\{ z \mid \|\tilde{\nabla}_{1/\beta} f(z)\| \geq \delta \right\}$. Thus, the above ratio must be bounded from above by a constant τ' (independent of b, and no greater than $\max_{z,z' \in Z} \|z - z'\|/\delta$). Combining (1) and (2) yields the desired error bound over the set $Z \cap \text{dom}(h)$. **Q.E.D.**

Proof of Theorem 5

From Theorem 4 we know that (x^r, z^r, λ^r) converges to the set of stationary solutions of problem (3.2). Let (x^*, z^*, λ^*) be one of such stationary solution. Then by the definition of the Q function and the fact that the successive differences of the gradients goes to zero (cf. (3.47)), we have

$$Q^* = f(z^*) = \sum_{i=1}^{N} 1/N g_i(z^*) + g_0(z^*) + p(z^*). \tag{3.63}$$

Then by Lemma 7 - (2) we know that $f(z^r) = \sum_{i=1}^{N} 1/N g_i(z^r) + g_0(z^r) + p(z^r)$ will finally settle at some isocost surface of f, i.e., there exists some *finite* $\bar{r} > 0$ such that for all $r > \bar{r}$ and $\bar{v} \in \mathbb{R}$ such that

$$f(\bar{z}^r) = \bar{v}, \quad \forall\ r \geq \bar{r} \tag{3.64}$$

where $\bar{z}^r = \arg\min_{z \in Z^*} \|z^r - z\|$. Therefore, combining the fact that $\|x^{r+1} - x^r\| \to 0$, $\|z^{r+1} - z^r\| \to 0$, $\|x_i^{r+1} - z^{r+1}\| \to 0$ and $\|\lambda^{r+1} - \lambda^r\| \to 0$ (cf. (3.87), (3.88)), it is easy to see that

$$L(\bar{z}^r, \bar{x}^r, \bar{\lambda}^r) = f(\bar{z}^r) = \bar{v}, \quad \forall\ r \geq \bar{r}, \tag{3.65}$$

where $\bar{x}^r, \bar{\lambda}^r$ are defined similarly as $\bar{z}^r$.

Now we prove that the expectation of $\Delta^{r+1} := Q^{r+1} - \bar{v}$ diminishes Q-linearly. All the expectation below is w.r.t. the natural history of the algorithm. The proof consists of the following steps:

Step 1: There exists $\sigma_1 > 0$ such that

$$\mathbb{E}[Q^r - Q^{r+1}] \geq \sigma_1 \left(\mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} \mathbb{E}\|1/N \nabla g_i(z^r) - 1/N \nabla g_i(y_i^{r-1})\|^2 \right);$$

Step 2: There exists $\tau > 0$ such that

$$\mathbb{E}\|z^r - \bar{z}^r\|^2 \leq \tau \|\mathbb{E}[\nabla_{1/\beta} \tilde{f}(z^r)]\|^2;$$

Step 3: There exists $\sigma_2 > 0$ such that

$$\|\mathbb{E}[\nabla_{1/\beta} \tilde{f}(z^r)]\|^2 \leq \sigma_2 \left(\mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} \mathbb{E}\|1/N \nabla g_i(z^r) - 1/N \nabla g_i(y_i^{r-1})\|^2 \right);$$

Step 4: There exists $\sigma_3 > 0$ such that the following relation holds true for all $r \geq \bar{r}$

$$\mathbb{E}[Q^{r+1} - \bar{v}] \leq \sigma_3 \left(\mathbb{E}\|z^r - \bar{z}^r\|^2 + \mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} \mathbb{E}\|1/N \nabla g_i(z^r) - 1/N \nabla g_i(y_i^{r-1})\|^2 \right).$$

These steps will be verified one by one shortly. But let us suppose that they all hold true. Below we show that linear convergence can be obtained.

Combining step 4 and step 2 we conclude that there exists $\sigma_3 > 0$ such that for all $r \geq \bar{r}$

$$\mathbb{E}[Q^{r+1} - \bar{v}] \leq \sigma_3 \left(\tau \|\mathbb{E}[\nabla_{1/\beta} \tilde{f}(z^{r-1})]\|^2 + \mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} \mathbb{E}\|1/N \nabla g_i(z^r) - 1/N \nabla g_i(y_i^{r-1})\|^2 \right).$$

Then if we bound $\|\mathbb{E}(G^r)\|^2$ using step 3, we can simply make a $\sigma_4 > 0$ such that

$$\mathbb{E}[Q^{r+1} - \bar{v}] \leq \sigma_4 \left(\mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} \mathbb{E}\|1/N \nabla g_i(z^r) - 1/N \nabla g_i(y_i^{r-1})\|^2 \right).$$

Finally, applying step 1 we reach the following bound for $\mathbb{E}[Q^{r+1} - \bar{v}]$:

$$\mathbb{E}[Q^{r+1} - \bar{v}] \leq \frac{\sigma_4}{\sigma_1}\mathbb{E}[Q^r - Q^{r+1}], \quad \forall\ r \geq \bar{r},$$

which further implies that for $\sigma_5 = \frac{\sigma_4}{\sigma_1} > 0$, we have

$$\mathbb{E}[\Delta^{r+1}] \leq \frac{\sigma_5}{1+\sigma_5}\mathbb{E}[\Delta^r], \quad \forall\ r \geq \bar{r}.$$

Now let us verify the correctness of each step. Step 1 can be directly obtained from equation (3.12). Step 2 is exactly Lemma (7). Step 3 can be verified using a similar derivation as in (5.126)[2].

Below let us prove the step 4, which is a bit involved. From (3.7) we know that

$$z^{r+1} = \arg\min_z h(z) + g_0(z) + \sum_{i=1}^N \langle \lambda_i^r, x_i^{r+1} - z\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - z\|^2.$$

This implies that

$$\begin{aligned} &h(z^{r+1}) + g_0(z^{r+1}) + \sum_{i=1}^N \langle \lambda_i^r, x_i^{r+1} - z^{r+1}\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - z^{r+1}\|^2 \\ &\leq h(\bar{z}^r) + g_0(\bar{z}^r) + \sum_{i=1}^N \langle \lambda_i^r, x_i^{r+1} - \bar{z}^r\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2. \end{aligned} \tag{3.66}$$

Rearranging the terms, we obtain

$$h(z^{r+1}) + g_0(z^{r+1}) - h(\bar{z}^r) - g_0(\bar{z}^r) \leq \sum_{i=1}^N \langle \lambda_i^r, z^{r+1} - \bar{z}^r\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2.$$

Using this inequality we have:

$$\begin{aligned} Q^{r+1} - \bar{v} \leq &\sum_{i=1}^N 1/N\left(g_i(z^{r+1}) - g_i(\bar{z}^r)\right) + \langle \lambda_i^r, z^{r+1} - \bar{z}^r\rangle \\ &+ \sum_{i=1}^N \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2 + \|1/N(\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2. \end{aligned} \tag{3.67}$$

[2]We simply need to replace $-z^{r-1} + \text{prox}_h^{1/\beta}[u^{r-1} - \beta\nabla g_0(z^{r-1})]$ in step (a) of (5.126) by $-z^r + \text{prox}_h^{1/\beta}[u^r - \beta\nabla g_0(z^r)]$ and using the same derivation.

The first term in RHS can be bounded as follows:

$$\begin{aligned}
&\sum_{i=1}^{N} 1/N \left(g_i(z^{r+1}) - g_i(\bar{z}^r)\right) \\
&\overset{(a)}{\leq} \sum_{i=1}^{N} 1/N \langle \nabla g_i(\bar{z}^r), z^{r+1} - \bar{z}^r \rangle + L_i/2N \|z^{r+1} - \bar{z}^r\|^2 \\
&\leq \sum_{i=1}^{N} 1/N \langle \nabla g_i(\bar{z}^r) + \nabla g_i(z^{r+1}) - \nabla g_i(z^{r+1}), z^{r+1} - \bar{z}^r \rangle + L_i/2N \|z^{r+1} - \bar{z}^r\|^2 \\
&\overset{(b)}{\leq} \sum_{i=1}^{N} 1/N \langle \nabla g_i(z^{r+1}), z^{r+1} - \bar{z}^r \rangle + 3L_i/2N \|z^{r+1} - \bar{z}^r\|^2,
\end{aligned}$$

where (a) is true due to the descent lemma; and (b) comes from the Lipschitz continuity of the ∇g_i.

Plugging the above bound into (3.67), we further have:

$$\begin{aligned}
Q^{r+1} - \bar{v} &\leq \sum_{i=1}^{N} 1/N \langle \nabla g_i(z^{r+1}) - \nabla g_i(y_i^{r-1}), z^{r+1} - \bar{z}^r \rangle + 3L_i/2N \|z^{r+1} - \bar{z}^r\|^2 \\
&\quad + \frac{\eta_i}{2} \|x_i^{r+1} - \bar{z}^r\|^2 + \|1/N(\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2 \\
&= \sum_{i=1}^{N} 1/N \langle \nabla g_i(z^{r+1}) + \nabla g_i(z^r) - \nabla g_i(z^r) - \nabla g_i(y_i^{r-1}), z^{r+1} - \bar{z}^r \rangle \\
&\quad + \frac{\eta_i}{2} \|x_i^{r+1} - \bar{z}^r\|^2 + \|1/N(\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2 + 3L_i/2N \|z^{r+1} - \bar{z}^r\|^2,
\end{aligned}$$

where in the first inequality we have used the fact that $\lambda_i^r = -\frac{1}{N} \nabla g_i(y_i^{r-1})$; cf . (5.100). Applying the Cauchy-Schwartz inequality we further have:

$$\begin{aligned}
Q^{r+1} - \bar{v} &\leq \sum_{i=1}^{N} 1/2 \|1/N \left(\nabla g_i(z^{r+1}) + \nabla g_i(z^r)\right)\|^2 + 1/2 \|z^{r+1} - \bar{z}^r\|^2 \\
&\quad + \sum_{i=1}^{N} 1/2 \|1/N \left(\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\right)\|^2 + 1/2 \|z^{r+1} - \bar{z}^r\|^2 \\
&\quad + \frac{\eta_i}{2} \|x_i^{r+1} - \bar{z}^r\|^2 + \|1/N(\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2 + 3L_i/2N \|z^{r+1} - \bar{z}^r\|^2 \\
&\leq \sum_{i=1}^{N} \left[\frac{L_i^2}{2N^2} \|z^{r+1} - z^r\|^2 + \frac{3}{2N^2} \|g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2 + \frac{\eta_i}{2} \|x_i^{r+1} - \bar{z}^r\|^2 \right] \\
&\quad + (1 + 3L_i/2N) \|z^{r+1} - \bar{z}^r\|^2. \qquad (3.68)
\end{aligned}$$

Now let us bound $\sum_{i=1}^{N} \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2$ in the above inequality:

$$\begin{aligned}\sum_{i=1}^{N} \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2 &= \sum_{i=1}^{N} \frac{\eta_i}{2}\|x_i^{r+1} - z^{r+1} + z^{r+1} - \bar{z}^r\|^2 \\ &\leq \sum_{i=1}^{N} \eta_i\|x_i^{r+1} - z^{r+1}\|^2 + \eta_i\|z^{r+1} - \bar{z}^r\|^2 \\ &= \sum_{i=1}^{N} \eta_i\|x_i^{r+1} - z^r + z^r - z^{r+1}\|^2 + \eta_i\|z^{r+1} - \bar{z}^r\|^2 \\ &\leq \sum_{i=1}^{N} 2\eta_i\|x_i^{r+1} - z^r\|^2 + 2\eta_i\|z^r - z^{r+1}\|^2 + \eta_i\|z^{r+1} - \bar{z}^r\|^2.\end{aligned}$$

Using the fact that $x_i^{r+1} = z^r$ when $i \neq i_r$ we further have:

$$\begin{aligned}\sum_{i=1}^{N} \frac{\eta_i}{2}\|x_i^{r+1} - \bar{z}^r\|^2 &\leq 2\eta_{i_r}\|x_{i_r}^{r+1} - z^r\|^2 + \sum_{i=1}^{N} 2\eta_i\|z^r - z^{r+1}\|^2 + \eta_i\|z^{r+1} - \bar{z}^r\|^2 \\ &= \frac{2}{\alpha_{i_r}^2\eta_{i_r}}\|\lambda_{i_r} + 1/N\nabla g_{i_r}(z^r)\|^2 + \sum_{i=1}^{N} 2\eta_i\|z^r - z^{r+1}\|^2 + \eta_i\|z^{r+1} - \bar{z}^r\|^2 \\ &= \frac{2}{\alpha_{i_r}^2\eta_{i_r}N^2}\|\nabla g_{i_r}(z^r) - \nabla g_{i_r}(y_{i_r}^{r-1})\|^2 \\ &+ \sum_{i=1}^{N} 2\eta_i\|z^r - z^{r+1}\|^2 + \eta_i\|z^{r+1} - z^r + z^r - \bar{z}^r\|^2 \\ &\leq \frac{2}{\alpha_{i_r}^2\eta_{i_r}N^2}\|\nabla g_{i_r}(z^r) - \nabla g_{i_r}(y_{i_r}^{r-1})\|^2 \\ &+ \sum_{i=1}^{N} 4\eta_i\|z^r - z^{r+1}\|^2 + 2\eta_i\|z^r - \bar{z}^r\|^2. \qquad (3.69)\end{aligned}$$

Take expectation on both sides of the above equation and set $p_i = \alpha_i$, we obtain:

$$\begin{aligned}\sum_{i=1}^{N} \frac{\eta_i}{2}\mathbb{E}\|x_i^{r+1} - \bar{z}^r\|^2 \leq \sum_{i=1}^{N} \frac{2}{\alpha_i\eta_i}\mathbb{E}\|\nabla g_i(z^r) - \nabla g_i(y_i^{r-1})\|^2 \\ + \sum_{i=1}^{N} 4\eta_i\mathbb{E}\|z^r - z^{r+1}\|^2 + 2\eta_i\mathbb{E}\|z^r - \bar{z}^r\|^2.\end{aligned}$$

Combining equations (3.68) and (3.69), eventually one can find $\sigma_3 > 0$ such that

$$\mathbb{E}[Q^{r+1} - \bar{v}] \leq \sigma_3\left(\mathbb{E}\|z^r - \bar{z}\|^2 + \mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N}\mathbb{E}\|1/N\nabla g_i(z^r) - 1/N\nabla g_i(y_i^{r-1})\|^2\right),$$

which completes the proof of Step 4.

In summary, we have shown that Step 1 - 4 all hold true. Therefore we have shown that the NESTT-G converges Q-linearly. **Q.E.D.**

Some Key Properties of NESTT-E

To facilitate the following derivation, in this section we collect some key properties of NESTT-E.

First, for $i = i_r$, using the optimality condition for x_i update step (3.16) we have the following identity:

$$\frac{1}{N}\nabla g_{i_r}(x_{i_r}^{r+1}) + \lambda_{i_r}^r + \alpha_{i_r}\eta_{i_r}(x_{i_r}^{r+1} - z^{r+1}) = 0. \tag{3.70}$$

Combined with the dual variable update step (3.17) we obtain

$$\frac{1}{N}\nabla g_{i_r}(x_{i_r}^{r+1}) = -\lambda_{i_r}^{r+1}. \tag{3.71}$$

Second, the optimality condition for the z-update is given by:

$$z^{r+1} = \text{prox}_h\left[z^{r+1} - \nabla_z(L(x^r, z, \lambda^r) - h(z))\right] \tag{3.72}$$

$$= \text{prox}_h\left[z^{r+1} - \sum_{i=1}^N \eta_i\left(z^{r+1} - x_i^r - \frac{\lambda_i^r}{\eta_i}\right) - \nabla g_0(z^{r+1})\right]. \tag{3.73}$$

3.6.4 Proof of Theorem 6

To prove this result, we need a few lemmas.

For notational simplicity, define new variables $\{\hat{x}_i^{r+1}\}$, $\{\hat{\lambda}_i^{r+1}\}$ by

$$\hat{x}_i^{r+1} := \arg\min_{x_i}\ U_i(x_i, z^{r+1}, \lambda_i^r), \quad \hat{\lambda}_i^{r+1} := \lambda_i^r + \alpha_i\eta_i\left(\hat{x}_i^{r+1} - z^{r+1}\right), \quad \forall i. \tag{3.74}$$

These variables are the *virtual variables* generated by updating all variables at iteration $r+1$. Also define:

$$L^r := L(x^r, z^r; \lambda^r), \quad w := (x, z, \lambda), \quad \beta := \frac{1}{\sum_{i=1}^N \eta_i}, \quad c_i := \frac{L_i^2}{\alpha_i\eta_i N^2} - \frac{\gamma_i}{2} + \frac{1-\alpha_i}{\alpha_i}\frac{L_i}{N}$$

First, we need the following lemma to show that the size of the successive difference of the dual variables can be upper bounded by that of the primal variables. This is a simple consequence of (3.71); also see [R2, Lemma 2.1]. We include the proof for completeness.

Lemma 8 *Suppose assumption A holds. Then for NESTT-E algorithm, the following are true:*

$$\|\lambda_i^{r+1} - \lambda_i^r\|^2 \leq \frac{L_i^2}{N^2}\|x_i^{r+1} - x_i^r\|^2, \quad \|\hat{\lambda}_i^{r+1} - \lambda_i^r\|^2 \leq \frac{L_i^2}{N^2}\|\hat{x}_i^{r+1} - x_i^r\|^2, \ \forall\ i. \tag{3.75a}$$

Proof. We only show the first inequality. The second one follows an analogous argument.

To prove (3.75a), first note that the case for $i \neq i_r$ is trivial, as both sides of (3.75a) are zero. For the index i_r, we have a closed-form expression for $\lambda_{i_r}^r$ following (3.71). Notice that for any given i, the primal-dual pair (x_i, λ_i) is always updated at the same iteration. Therefore, if for each i we choose the initial solutions in a way such that $\lambda_i^0 = -\nabla g_i(x_i^0)$, then we have

$$\frac{1}{N}\nabla g_i(x_i^{r+1}) = -\lambda_i^{r+1} \quad \forall\ i = 1, 2, \cdots N. \tag{3.76}$$

Combining (3.76) with Assumption A-(a) yields the following:

$$\|\lambda_i^{r+1} - \lambda_i^r\| = \frac{1}{N}\|\nabla g_i(x_i^{r+1}) - \nabla g_i(x_i^r)\| \leq \frac{L_i}{N}\|x_i^{r+1} - x_i^r\|.$$

The proof is complete. **Q.E.D.**

Second, we bound the successive difference of the potential function.

Lemma 9 *Suppose Assumption A holds true. Then the following holds for NESTT-E*

$$\mathbb{E}[L^{r+1} - L^r | x^r, z^r] \leq -\frac{\gamma_z}{2}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^N p_i c_i \|x_i^r - \hat{x}_i^{r+1}\|^2. \tag{3.77}$$

Proof. First let us split $L^{r+1} - L^r$ in the following way:

$$L^{r+1} - L^r = L^{r+1} - L(x^{r+1}, z^{r+1}; \lambda^r) + L(x^{r+1}, z^{r+1}; \lambda^r) - L^r. \tag{3.78}$$

The first two terms in (3.78) can be bounded by

$$\begin{aligned} L^{r+1} - L(x^{r+1}, z^{r+1}; \lambda^r) &= \sum_{i=1}^N \langle \lambda_i^{r+1} - \lambda_i^r, x_i^{r+1} - z^{r+1} \rangle \\ &\overset{\text{(a)}}{=} \frac{1}{\alpha_{i_r}\eta_{i_r}}\|\lambda_{i_r}^{r+1} - \lambda_{i_r}^r\|^2 \overset{\text{(b)}}{\leq} \frac{L_{i_r}^2}{N^2\alpha_{i_r}\eta_{i_r}}\|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \end{aligned} \tag{3.79}$$

where in (a) we have used (3.17), and the fact that $\lambda_i^{r+1} - \lambda_i^r = 0$ for all variable blocks except i_rth block; (b) is true because of Lemma 8.

The last two terms in (3.78) can be written in the following way:

$$L(\{x_i^{r+1}\}, z^{r+1}; \lambda^r) - L^r = L(x^{r+1}, z^{r+1}; \lambda^r) - L(x^r, z^{r+1}; \lambda^r) + L(x^r, z^{r+1}; \lambda^r) - L^r. \tag{3.80}$$

The first two terms in (3.80) characterizes the change of the Augmented Lagrangian before and after the update of x. Note that x updates do not directly optimize the augmented Lagrangian. Therefore the characterization of this step is a bit involved. We have the following:

$$\begin{aligned}
&L(x^{r+1}, z^{r+1}; \lambda^r) - L(x^r, z^{r+1}; \lambda^r) \\
&\overset{(a)}{\leq} \sum_{i=1}^{N} \left(\left\langle \nabla_i L(x^{r+1}, z^{r+1}; \lambda^r), x_i^{r+1} - x_i^r \right\rangle - \frac{\gamma_i}{2} \|x_i^{r+1} - x_i^r\|^2 \right) \\
&\overset{(b)}{=} \left\langle \nabla_{i_r} L(x^{r+1}, z^{r+1}; \lambda^r), x_{i_r}^{r+1} - x_{i_r}^r \right\rangle - \frac{\gamma_{i_r}}{2} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \\
&\overset{(c)}{=} \left\langle \eta_{i_r}(1 - \alpha_{i_r})(x_{i_r}^{r+1} - z^{r+1}), x_{i_r}^{r+1} - x_{i_r}^r \right\rangle - \frac{\gamma_{i_r}}{2} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \\
&\overset{(d)}{=} \left\langle \frac{1 - \alpha_{i_r}}{\alpha_{i_r}} (\lambda_{i_r}^{r+1} - \lambda_{i_r}^r), x_{i_r}^{r+1} - x_{i_r}^r \right\rangle - \frac{\gamma_{i_r}}{2} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \\
&\leq \frac{1 - \alpha_{i_r}}{\alpha_{i_r}} \left(\frac{1}{2L_{i_r}/N} \|\lambda_{i_r}^{r+1} - \lambda_{i_r}^r\|^2 + \frac{L_{i_r}}{2N} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \right) - \frac{\gamma_{i_r}}{2} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 \\
&\overset{(e)}{\leq} \frac{1 - \alpha_{i_r}}{\alpha_{i_r}} \frac{L_{i_r}}{N} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2 - \frac{\gamma_{i_r}}{2} \|x_{i_r}^{r+1} - x_{i_r}^r\|^2
\end{aligned} \tag{3.81}$$

where

- (a) is true because $L(x, z, \lambda)$ is strongly convex with respect to x_i.
- (b) is true because when $i \neq i_r$, we have $x_i^{r+1} = x_i^r$.
- (c) is true because $x_{i_r}^{r+1}$ is optimal solution for the problem $\min U_{i_r}(x_{i_r}, z^{r+1}, \lambda_{i_r}^r)$ (satisfying (3.70)), and we have used the optimality of such $x_{i_r}^{r+1}$.
- (d) and (e) are due to Lemma 8.

Similarly, the last two terms in (3.80) can be bounded using equation (3.70) and the strong convexity of function L with respect to the variable z. Therefor We have:

$$L(x^r, z^{r+1}, \lambda^r) - L^r \leq -\frac{\gamma_z}{2} \|z^{r+1} - z^r\|^2. \tag{3.82}$$

Combining equations (3.79), (3.81) and (3.82), eventually we have:

$$L^{r+1} - L(x^r, z^{r+1}, \lambda^r) \leq c_{i_r}\|x^r_{i_r} - x^{r+1}_{i_r}\|^2 \tag{3.83}$$

$$L^{r+1} - L^r \leq -\frac{\gamma_z}{2}\|z^{r+1} - z^r\|^2 + c_{i_r}\|x^r_{i_r} - x^{r+1}_{i_r}\|^2 \tag{3.84}$$

Taking expectation on both side of this inequality with respect to i_r, we can conclude that:

$$\mathbb{E}[L^{r+1} - L^r \mid z^r, x^r] \leq -\frac{\gamma_z}{2}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^{N} p_i c_i \|x^r_i - \hat{x}^{r+1}_i\|^2 \tag{3.85}$$

where p_i is the probability of picking ith block. The lemma is proved. **Q.E.D.**

Lemma 10 *Suppose that Assumption A is satisfied, then* $L^r \geq \underline{f}$.

Proof. Using the definition of the augmented Lagrangian function we have:

$$\begin{aligned}
L^{r+1} &= \sum_{i=1}^{N}\left(\frac{1}{N}g_i(x_i^{r+1}) + \langle\lambda_i^{r+1}, x_i^{r+1} - z^{r+1}\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - z^{r+1}\|^2\right) + g_0(z^{r+1}) + p(z^{r+1}) \\
&\overset{(a)}{=} \sum_{i=1}^{N}\left(\frac{1}{N}g_i(x_i^{r+1}) + \frac{1}{N}\langle\nabla g_i(x_i^{r+1}), z^{r+1} - x_i^{r+1}\rangle + \frac{\eta_i}{2}\|x_i^{r+1} - z^{r+1}\|^2\right) + g_0(z^{r+1}) + p(z^{r+1}) \\
&\overset{(b)}{\geq} \sum_{i=1}^{N}\frac{1}{N}g_i(z^{r+1}) + \left(\frac{\eta_i}{2} - \frac{L_i}{2N}\right)\|z^{r+1} - x_i^{r+1}\|^2 + g_0(z^{r+1}) + p(z^{r+1}) \\
&\overset{(c)}{\geq} \sum_{i=1}^{N}\frac{1}{N}g_i(z^{r+1}) + g_0(z^{r+1}) + p(z^{r+1}) \geq \underline{f}
\end{aligned} \tag{3.86}$$

where (a) is true because of equation (3.71); (b) follows Assumption A-(b); (c) follows Assumption A-(d). The desired result is proven. **Q.E.D.**

Proof of Theorem 6. We first show that the algorithm converges to the set of stationary solutions, and then establish the convergence rate.

Step 1. Convergence to Stationary Solutions. Combining the descent estimate in Lemma 9 as well as the lower bounded condition in Lemma 10, we can again apply the Supermartigale Convergence Theorem [15, Proposition 4.2] and conclude that

$$\|x_i^{r+1} - x_i^r\| \to 0, \quad \|z^{r+1} - z^r\| \to 0, \text{with probability 1.} \tag{3.87}$$

From Lemma 8 we have that the constraint violation is satisfied

$$\|\lambda^{r+1} - \lambda^r\| \to 0, \quad \|x_i^{r+1} - z^r\| \to 0. \tag{3.88}$$

The rest of the proof follows similar lines as in [R2, Theorem 2.4]. Due to space limitations we omit the proof.

Step 2. Convergence Rate. We first show that there exists a $\sigma_1(\alpha) > 0$ such that

$$\|\tilde{\nabla} L(w^r)\|^2 + \sum_{i=1}^N \frac{L_i^2}{N^2}\|x_i^r - z^r\|^2 \leq \sigma_1(\alpha)\left(\|z^r - z^{r+1}\|^2 + \sum_{i=1}^N \|x_i^r - \hat{x}_i^{r+1}\|^2\right). \tag{3.89}$$

Using the definition of $\|\tilde{\nabla} L^r(w^r)\|$ we have:

$$\|\tilde{\nabla} L^r(w^r)\|^2 = \|z^r - \text{prox}_h\left[z^r - \nabla_z(L^r - h(z^r))\right]\|^2 + \sum_{i=1}^N \left\|\frac{1}{N}\nabla g_i(x_i^r) + \lambda_i^r + \eta_i(x_i^r - z^r)\right\|^2. \tag{3.90}$$

From the optimality condition of the z update (3.73) we have:

$$z^{r+1} = \text{prox}_h\left[z^{r+1} - \sum_{i=1}^N \eta_i\left(z^{r+1} - x_i^r - \frac{\lambda_i^r}{\eta_i}\right) - \nabla g_0(z^{r+1})\right].$$

Using this, the first term in equation (3.90) can be bounded as:

$$\begin{aligned}
&\|z^r - \text{prox}_h\left[z^r - \nabla_z(L^r - h(z^r))\right]\| \\
&= \left\|z^r - z^{r+1} + z^{r+1} - \text{prox}_h\left[z^r - \sum_{i=1}^N \eta_i(z^r - x_i^r - \frac{\lambda_i^r}{\eta_i}) - \nabla g_0(z^r)\right]\right\| \\
&\leq \|z^r - z^{r+1}\| + \left\|\text{prox}_h\left[z^{r+1} - \sum_{i=1}^N \eta_i\left(z^{r+1} - x_i^r - \frac{\lambda_i^r}{\eta_i}\right) - \nabla g_0(z^{r+1})\right]\right. \\
&\qquad \left. - \text{prox}_h\left[z^r - \sum_{i=1}^N \eta_i(z^r - x_i^r - \frac{\lambda_i^r}{\eta_i}) - \nabla g_0(z^r)\right]\right\| \\
&\leq 2\|z^{r+1} - z^r\| + \left(\sum_{i=1}^N \eta_i + L_0\right)\|z^r - z^{r+1}\|,
\end{aligned} \tag{3.91}$$

where in the last inequality we have used the nonexpansiveness of the proximity operator.

Similarly, the optimality condition of the x_i subproblem is given by

$$\frac{1}{N}\nabla g_i(\hat{x}_i^{r+1}) + \lambda_i^r + \alpha_i\eta_i(\hat{x}_i^{r+1} - z^{r+1}) = 0. \tag{3.92}$$

Applying this identity, the second term in equation (3.90) can be written as follows:

$$\begin{aligned}
&\sum_{i=1}^{N}\left\|\frac{1}{N}\nabla g_i(x_i^r)+\lambda_i^r+\eta_i(x_i^r-z^r)\right\|^2\\
&\overset{(a)}{=}\sum_{i=1}^{N}\left\|\frac{1}{N}\nabla g_i(x_i^r)-\frac{1}{N}\nabla g_i(\hat{x}_i^{r+1})+\eta_i(x_i^r-z^r)-\alpha_i\eta_i(\hat{x}_i^{r+1}-z^{r+1})\right\|^2\\
&=\sum_{i=1}^{N}\left\|\frac{1}{N}\nabla g_i(x_i^r)-\frac{1}{N}\nabla g_i(\hat{x}_i^{r+1})+\eta_i(x_i^r-\hat{x}_i^{r+1}+\hat{x}_i^{r+1}-z^{r+1}+z^{r+1}-z^r)-\alpha_i\eta_i(\hat{x}_i^{r+1}-z^{r+1})\right\|^2\\
&\overset{(b)}{\leq}4\sum_{i=1}^{N}\left[\left(\frac{L_i^2}{N^2}+\eta_i^2+\frac{(1-\alpha_i)^2L_i^2}{N^2\alpha_i^2}\right)\|\hat{x}_i^{r+1}-x_i^r\|^2+\eta_i^2\|z^{r+1}-z^r\|^2\right],
\end{aligned}\tag{3.93}$$

where (a) holds because of equation (3.92); (b) holds because of Lemma 8.

Finally, combining (3.91) and (3.93) leads to the following bound for proximal gradient

$$\begin{aligned}
\|\tilde{\nabla}L^r\|^2\leq&\left(4\sum_{i=1}^{N}\eta_i^2+\left(2+L_0+\sum_{i=1}^{N}\eta_i\right)^2\right)\|z^r-z^{r+1}\|^2\\
&+\sum_{i=1}^{N}4\left(\frac{L_i^2}{N^2}+\eta_i^2+\frac{(1-\alpha_i)^2L_i}{N^2\alpha_i^2}\right)\|x_i^r-\hat{x}_i^{r+1}\|^2.
\end{aligned}\tag{3.94}$$

Also note that:

$$\begin{aligned}
\sum_{i=1}^{N}\frac{L_i^2}{N^2}\|x_i^r-z^r\|^2&\leq\sum_{i=1}^{N}3\frac{L_i^2}{N^2}\left[\|x_i^r-\hat{x}_i^{r+1}\|^2+\|\hat{x}_i^{r+1}-z^{r+1}\|^2+\|z^{r+1}-z^r\|^2\right]\\
&=\sum_{i=1}^{N}3\frac{L_i^2}{N^2}\left[\|x_i^r-\hat{x}_i^{r+1}\|^2+\frac{1}{\alpha_i^2\eta_i^2}\|\hat{\lambda}_i^{r+1}-\lambda_i^r\|^2+\|z^{r+1}-z^r\|^2\right]\\
&\leq\sum_{i=1}^{N}3\frac{L_i^2}{N^2}\left[\|x_i^r-\hat{x}_i^{r+1}\|^2+\frac{L_i^2}{\alpha_i^2\eta_i^2N^2}\|\hat{x}_i^{r+1}-x_i^r\|^2+\|z^{r+1}-z^r\|^2\right].
\end{aligned}\tag{3.95}$$

The two inequalities (3.94) – (3.95) imply that:

$$\begin{aligned}
&\|\tilde{\nabla}L^r\|^2+\sum_{i=1}^{N}\frac{L_i^2}{N^2}\|x_i^r-z^r\|^2\\
&\leq\left(\sum_{i=1}^{N}4\eta_i^2+(2+\sum_{i=1}^{N}\eta_i+L_0)^2+3\sum_{i=1}^{N}\frac{L_i^2}{N^2}\right)\|z^r-z^{r+1}\|^2\\
&+\sum_{i=1}^{N}\left(4\left(\frac{L_i^2}{N^2}+\eta_i^2+(\frac{1}{\alpha_i}-1)^2\frac{L_i^2}{N^2}\right)+3\left(\frac{L_i^4}{\alpha_iN^4\eta_i^2}+\frac{L_i^2}{N^2}\right)\right)\|x_i^r-\hat{x}_i^{r+1}\|^2.
\end{aligned}\tag{3.96}$$

Define the following quantities:

$$\hat{\sigma}_1(\alpha) = \max_i \left\{ 4\left(\frac{L_i^2}{N^2} + \eta_i^2 + \left(\frac{1}{\alpha_i} - 1\right)^2 \frac{L_i^2}{N^2}\right) + 3\left(\frac{L_i^4}{\alpha_i \eta_i^2 N^4} + \frac{L_i^2}{N^2}\right)\right\}$$

$$\tilde{\sigma}_1 = \sum_{i=1}^N 4\eta_i^2 + (2 + \sum_{i=1}^N \eta_i + L_0)^2 + 3\sum_{i=1}^N \frac{L_i^2}{N^2}.$$

Setting $\sigma_1(\alpha) = \max(\hat{\sigma}_1(\alpha), \tilde{\sigma}_1) > 0$, we have

$$\|\tilde{\nabla} L^r\|^2 + \sum_{i=1}^N \frac{L_i^2}{N^2}\|x_i^r - z^r\|^2 \leq \sigma_1(\alpha)\left(\|z^r - z^{r+1}\|^2 + \sum_{i=1}^N \|x_i^r - \hat{x}_i^{r+1}\|^2\right). \tag{3.97}$$

From Lemma 9 we know that

$$\mathbb{E}[L^{r+1} - L^r | z^r, x^r] \leq -\frac{\gamma_z}{2}\|z^{r+1} - z^r\|^2 + \sum_{i=1}^N p_i c_i \|x_i^r - \hat{x}_i^{r+1}\|^2 \tag{3.98}$$

Note that $\gamma_z = \sum_{i=1}^N \eta_i - L_0$, then define $\hat{\sigma}_2$ and $\tilde{\sigma}_2$ as

$$\hat{\sigma}_2(\alpha) = \max_i \left\{ p_i \left(\frac{\gamma_i}{2} - \frac{L_i^2}{\alpha_i \eta_i N^2} - \frac{1 - \alpha_i}{\alpha_i}\frac{L_i}{N}\right)\right\}$$

$$\tilde{\sigma}_2 = \frac{\sum_{i=1}^N \eta_i - L_0}{2}.$$

We can set $\sigma_2(\alpha) = \max(\hat{\sigma}_2(\alpha), \tilde{\sigma}_2)$ to obtain

$$E[L^r - L^{r+1} | x^r, z^r] \geq \sigma_2(\alpha)\left(\sum_{i=1}^N \|\hat{x}_i^{r+1} - x_i^r\|^2 + \|z^{r+1} - z^r\|^2\right). \tag{3.99}$$

Combining (3.89) and (3.99) we have

$$H(w^r) = \|\tilde{\nabla} L^r\|^2 + \sum_{i=1}^N L_i^2/N\|x_i^r - z^r\|^2 \leq \frac{\sigma_1(\alpha)}{\sigma_2(\alpha)} E[L^r - L^{r+1} | F^r].$$

Let us set $C(\alpha) = \frac{\sigma_1(\alpha)}{\sigma_2(\alpha)}$ and take expectation on both side of the above equation to obtain:

$$\mathbb{E}[H(w^r)] \leq C(\alpha) E[L^r - L^{r+1}]. \tag{3.100}$$

Summing both sides of the above inequality over $r = 1, \cdots, R$, we obtain:

$$\sum_{r=1}^R \mathbb{E}[H(w^r)] \leq C(\alpha) E[L^1 - L^{R+1}]. \tag{3.101}$$

Using the definition of $w^m = (x^m, z^m, \lambda^m)$, and following the same line of argument as Theorem (4) we eventually conclude that

$$\mathbb{E}[H(w^m)] \leq \frac{C(\alpha)\mathbb{E}[L^1 - L^{R+1}]}{R}. \tag{3.102}$$

The proof is complete. **Q.E.D.**

3.6.5 Proof of Proposition 1

Applying the optimality condition on z subproblem in (3.32) we have:

$$z^{r+1} = \underset{z}{\operatorname{argmin}}\, h(z) + g_0(z) + \frac{\beta}{2}\|z - u^{r+1}\|^2 \tag{3.103}$$

where the variable u^{r+1} is given by (cf. (5.108))

$$u^{r+1} = \beta \sum_{i=1}^{N} (\lambda_i^r + \eta_i x_i^{r+1}). \tag{3.104}$$

Now from one of the key properties of NESTT-G [cf. Section 3.6, equation (3.28)], we have that

$$u^{r+1} = z^r - \beta \left(\frac{1}{N\alpha_{i_r}} \left(\nabla g_{i_r}(z^r) - \nabla g_{i_r}(y_{i_r}^{r-1}) \right) + \frac{1}{N} \sum_{i=1}^{N} \nabla g_i(y_i^{r-1}) N \right). \tag{3.105}$$

This verifies the claim. **Q.E.D.**

Table 3.1: Comparison of # of gradient evaluations for NESTT-G and GD in the worst case

	NESTT-G	**GD**
# of Gradient Evaluations	$\mathcal{O}\left((\sum_{i=1}^{N}\sqrt{L_i/N})^2/\epsilon\right)$	$\mathcal{O}\left(\sum_{i=1}^{N} L_i/\epsilon\right)$
Case I: $L_i = 1,\ \forall i$	$\mathcal{O}(N/\epsilon)$	$\mathcal{O}(N/\epsilon)$
Case II : $\mathcal{O}(\sqrt{N})$ terms with $L_i = N$ the rest with $L_i = 1$	$\mathcal{O}(N/\epsilon)$	$\mathcal{O}(N^{3/2}/\epsilon)$
Case III : $\mathcal{O}(1)$ terms with $L_i = N^2$ the rest with $L_i = 1$	$\mathcal{O}(N/\epsilon)$	$\mathcal{O}(N^2/\epsilon)$

Table 3.2: Optimality gap $\|\tilde{\nabla}_{1/\beta} f(z^r)\|^2$ for different algorithms, with 100 passes of the datasets.

	SGD		**NESTT-E** ($\alpha = 10$)		**NESTT-G**		**SAGA**	
N	Uniform	Non-Uni	Uniform	Non-Uni	Uniform	Non-Uni	Uniform	Non-Uni
10	3.4054	0.2265	2.6E-16	6.16E-19	2.3E-21	6.1E-24	2.7E-17	2.8022
20	0.6370	6.9087	2.4E-9	5.9E-9	1.2E-10	2.9E-11	7.7E-7	11.3435
30	0.2260	0.1639	3.2E-6	2.7E-6	4.5E-7	1.4E-7	2.5E-5	0.1253
40	0.0574	0.3193	5.8E-4	8.1E-5	1.8E-5	3.1E-5	4.1E-5	0.7385
50	0.0154	0.0409	8.3E.-4	7.1E-4	1.2E-4	2.7E-4	2.5E-4	3.3187

CHAPTER 4. PERTURBED PROXIMAL PRIMAL DUAL ALGORITHM FOR NONCONVEX NONSMOOTH OPTIMIZATION

Abstract

In this paper we propose a perturbed proximal primal dual algorithm (PProx-PDA) for optimization problems whose objective is the sum of smooth (possibly nonconvex) and convex (possibly nonsmooth) functions subject to a linear coupling constraint. This family of problems has applications in a number of statistical and engineering applications, for example in high-dimensional subspace estimation, and distributed signal processing and learning over networks. The proposed method is of Uzawa type, in which a primal gradient descent step is performed followed by a (approximate) dual gradient ascent step. One distinctive feature of the proposed algorithm is that the primal and dual steps are both perturbed appropriately using past iterates so that a number of asymptotic convergence and rate of convergence results (to first-order stationary solutions) can be obtained. Finally, we conduct extensive numerical experiments to validate the effectiveness of the proposed algorithms.

4.1 Introduction

The Problem

Consider the following optimization problem

$$\min_{x\in X} \; f(x)+h(x), \quad \text{s.t.} \quad Ax=b, \tag{4.1}$$

where $f(x):\mathbb{R}^N\to\mathbb{R}$ is a continuous smooth function (possibly nonconvex); $A\in\mathbb{R}^{M\times N}$ is a rank deficient matrix; $b\in\mathbb{R}^M$ is a given vector; X is a convex compact set; $h(x):\mathbb{R}^N\to\mathbb{R}$ is a lower semi-continuous nonsmooth convex function. Problem (4.1) is an interesting class that can

be specialized to a number of statistical and engineering applications. We provide a few of these applications in the next subsection.

4.1.1 Motivating Applications

Sparse subspace estimation. Suppose that $\Sigma \in \mathbb{R}^{p\times p}$ is an unknown covariance matrix, $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ and $u_1, u_2, \cdots, u_p$ are its eigenvalues and eigenvectors, respectively, and they satisfy $\Sigma = \sum_{i=1}^{p} \lambda_i u_i u_i^\top$. Principal Component Analysis (PCA) aims to recover $u_1, u_2, \cdots, u_k$, where $k \leq p$, from a sample covariance matrix $\hat{\Sigma}$ obtained from i.i.d samples $\{x_i\}_{i=1}^n$. The subspace spanned by $\{u_i\}_{i=1}^k$ is called k-dimensional principal subspace, whose projection matrix is given by $\Pi^* = \sum_{i=1}^{p} u_i u_i^\top$. Therefore, PCA reduces to finding an estimate of Π^*, denoted by $\hat{\Pi}$, from the sample covariance matrix $\hat{\Sigma}$. In high dimensional setting where the number of data points is significantly smaller than the dimension i.e. $(n \ll p)$, it is desirable to find a *sparse* $\hat{\Pi}$, using the following formulation [49]

$$\min_{\Pi} \ \langle \hat{\Sigma}, \Pi \rangle + \mathcal{P}_\nu(\Pi), \qquad \text{s.t.} \quad \Pi \in \mathcal{F}^k. \tag{4.2}$$

In the above formulation, $\mathcal{F}^k$ denotes the Fantope set [135], given by $\mathcal{F}^k = \{X : \ 0 \preceq X \preceq I, \mathrm{trace}(X) = k\}$, which promotes low rankness in X. The function $\mathcal{P}_\nu(\Pi)$ is a nonconvex regularizer that enforces sparsity on Π. Typical forms of this regularization are smoothly clipped absolute deviation (SCAD) [35], and minimax concave penalty (MCP) [143]. For example, the definition of MCP regularization with parameters b and ν is given below

$$\mathcal{P}_\nu(\phi) = \iota_{|\phi|\leq b\nu}\left(\nu|\phi| - \frac{\phi^2}{2b}\right) + \iota_{|\phi|>b\nu}\left(\frac{b\nu^2}{2}\right), \tag{4.3}$$

where, ι_X denoted the indicator function for convex set X, which is defined as $\iota_X(y) = 0$ when $y \in X$, and $\iota_X(y) = \infty$ otherwise.

One particular characterization for these nonconvex penalties is that they can be decomposed as a sum of an ℓ_1 function and a concave function $q_\nu(x)$: $\mathcal{P}_\nu(\phi) = \nu|\phi| + q_\nu(\phi)$ for some $\nu \geq 0$. In a recent work [49], it is shown that with high probability, every first-order stationary solution of problem (4.2) (denoted as $\hat{\Pi}$) is of high-quality, in the sense that it satisfies the following error

$$\|\hat{\Pi} - \Pi^*\|_F \leq \frac{4C\lambda_1\sqrt{s_1}}{\lambda_k - \lambda_{k+1}}\sqrt{\frac{s}{n}} + \frac{12C\lambda_1\sqrt{m_1 m_2}}{\lambda_k - \lambda_{k+1}}\sqrt{\frac{\log p}{n}}, \tag{4.4}$$

where $s = |\text{supp}(\text{diag}(\Pi^*))|$, C, s_1, m_1, and m_2 are some constants. See [49, Theorem 3] for detailed description. In order to deal with the Fantope and the nonconvex regularizer separately, one can introduce a new variable Φ and reformulate problem (4.2) in the following manner [135]

$$\min_{\Pi,\Phi} \ \left\langle \hat{\Sigma}, \Pi \right\rangle + \mathcal{P}_\nu(\Phi) \quad \text{s.t.} \quad \Pi \in \mathcal{F}^k, \ \Pi - \Phi = 0. \tag{4.5}$$

Clearly this is special case of problem (4.1), with $x = [\Pi, \Phi]$, $f(x) = \left\langle \hat{\Sigma}, \Pi \right\rangle + q_\nu(\Phi)$, $h(x) = \nu\|\Phi\|_1$, $X = \mathcal{F}^k$, $A = [I, -I]$, $b = 0$.

The exact consensus problem over networks. Consider a network which consists of N agents who collectively optimize the following problem

$$\min_{y\in\mathbb{R}} \ f(y) + h(y) := \sum_{i=1}^{N} \left(f_i(y) + h_i(y)\right), \tag{4.6}$$

where $f_i(y) : \mathbb{R} \to \mathbb{R}$ is a smooth function, and $h_i(y) : \mathbb{R} \to \mathbb{R}$ is a convex, possibly nonsmooth regularizer (here y is assumed to be scalar for ease of presentation). Note that both f_i and h_i are only accessible by agent i. In particular, each local loss function f_i can represent: 1) a mini-batch of (possibly nonconvex) loss functions modeling data fidelity [7]; 2) nonconvex activation functions of neural networks [3]; 3) nonconvex utility functions used in applications such as resource allocation [18]. The regularization function h_i usually take the following forms: 1) convex regularizers such as nonsmooth ℓ_1 or smooth ℓ_2 functions; 2) the indicator function for closed convex set X, i.e. ι_X. This problem has found applications in various domains such as distributed statistical learning [95], distributed consensus [133], distributed communication networking [147, 82], distributed and parallel machine learning [61, 39] and distributed signal processing [117, 147]; for more applications we refer the readers to a recent survey [46].

To integrate the structure of the network into problem (4.6), we assume that the agents are connected through a network defined by an undirected, connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with $|\mathcal{V}| = N$ vertices and $|\mathcal{E}| = E$ edges. Each agent can only communicate with its immediate neighbors, and it is responsible for optimizing one component function f_i regularized by h_i. Define the incidence matrix $A \in \mathbb{R}^{E\times N}$ as following: if $e \in \mathcal{E}$ and it connects vertex i and j with $i > j$, then $A_{ev} = 1$ if $v = i$, $A_{ev} = -1$ if $v = j$ and $A_{ev} = 0$ otherwise. Using this definition, the *signed graph Laplacian matrix* L_- is given by $L_- := A^T A \in \mathbb{R}^{N\times N}$. Introducing N new variables x_i as the local copy of the

global variable y, and define $x := [x_1; \cdots; x_N] \in \mathbb{R}^N$, problem (4.6) can be equivalently expressed as

$$\min_{x \in \mathbb{R}^N} \; f(x) + h(x) := \sum_{i=1}^{N} \left(f_i(x_i) + h_i(x_i) \right), \text{ s.t. } Ax = 0. \tag{4.7}$$

This problem is precisely original problem (4.1) with correspondence $X = \mathbb{R}^N$, $b = 0$, $f(x) := \sum_{i=1}^N f_i(x_i)$, and $h(x) := \sum_{i=1}^N h_i(x_i)$.

The partial consensus problem. In the previous application, the agents are required to reach *exact* consensus, and such constraint is imposed through $Ax = 0$ in (4.7). In practice, however, consensus is rarely achieved exactly, for example due to potential disturbances in network communication; see detailed discussion in [75]. Further, in applications ranging from distributed estimation to rare event detection, the data obtained by the agents, such as harmful algal blooms, network activities, and local temperature, often exhibit distinctive spatial structure [28]. The distributed problem in these settings can be best formulated by using certain partial consensus model in which the local variables of an agent are only required to be close to those of its neighbors. To model such a *partial* consensus constraint, we denote ξ_e as the permissible tolerance for $e = (i, j) \in \mathcal{E}$, and replace the strict consensus constraint $x_i - x_j = 0$ with $\|x_i - x_j\|^2 \leq \xi_e$. Further, we define the link variable $z_e = x_i - x_j$, and set $z := \{z_e\}_{e \in \mathcal{E}}$, $Z := \{z \mid \|z_e\|^2 \leq \xi_e \; \forall \; e \in \mathcal{E}\}$. Using these notations, the partial consensus problem can be formulated as

$$\min_{x,z} \quad \sum_{i=1}^{N} \left(f_i(x_i) + h_i(x_i) \right) \quad \text{s.t.} \quad Ax - z = 0, \; z \in Z, \tag{4.8}$$

which is again a special case of problem (4.1).

4.1.2 Literature Review and Contribution.

4.1.2.1 Literature on Related Algorithms.

The Augmented Lagrangian (AL) method, also known as the methods of multipliers, is pioneered by Hestenes [58] and Powell [107]. It is a classical algorithm for solving nonconvex smooth constrained problems and its convergence is guaranteed under rather week assumptions [14, 106, 36]. A modified version of AL has been develped by Rockafellar in [113], in which a proximal term has been added to the objective function in order to make it strongly convex in each iteration. Later

Wright [73] specialized this algorithm to the linear programming problem. Many existing packages such as LANCELOT are implemented based on AL method. Recently, due to the need to solve very large scale nonlinear optimization problems, the AL and its variants regain their popularity. For example, in [29] a line search AL method has been proposed for solving problem (4.1) with $h \equiv 0$ and $X = \{x;\ l \leq x \leq u\}$. Also reference [21] has developed an AL based algorithm for nonconvex nonsmooth optimization, where subgradients of the augmented Lagrangian are used in the primal update. When the problem is convex, smooth and the constraints are linear, Lan and Monterio [77] have analyzed the iteration complexity for the AL method. More specifically, the authors analyzed the total number of Nesterov's optimal iterations [104] that are required to reach high quality primal-dual solutions. Subsequently, Liu et al [87] proposed an inexact AL (IAL) algorithm which only requires an $\epsilon-$approximated solution for the primal subproblem at each iteration. Hong et al [61] proposed a proximal primal-dual algorithm (Prox-PDA), an AL-based method mainly used to solve smooth and unconstrained distributed nonconvex problem [i.e. problem (4.7) with $h_i \equiv 0$ and $X \in \mathbb{R}^N$]. Overall, the AL based methods often require sophisticated stepsize selection, and an accurate oracle for solving the primal problem. Further, they cannot deal with problems that have both nonsmooth regularizer $h(x)$ and a general convex constraint. Therefore, it is not straightforward to apply these methods to problems such as distributed learning and high-dimensional sparse subspace estimation mentioned in the previous subsection.

Recently, the alternating direction method of multipliers (ADMM), a variant of the AL, has gained popularity for decomposing large-scale nonsmooth optimization problems [20]. The method originates in early 1970s [48, 42], and has since been studied extensively [16, 62, 32]. The main strength of this algorithm is that it is capable of decomposing a large problem into a series of small and simple subproblems, therefore making the overall algorithm scalable and easy to implement. However, unlike the AL method, the ADMM is designed for convex problems, despite its good numerical performance in nonconvex problems such as the nonnegative matrix factorization [130], phase retrieval [139], distributed clustering [39], tensor decomposition [84] and so on. Only very recently, researchers have begun to rigorously investigate the convergence of ADMM (to first-

order stationary solutions) for nonocnvex problems. Zhang [144] have analyzed a class of splitting algorithms (which includes the ADMM as a special case) for a very special class of nonconvex quadratic problems. Ames and Hong in [5] have developed an analysis for ADMM for certain ℓ_1 penalized problem arising in high-dimensional discriminant analysis. Other works along this line include [63, 79, 57, 97] and [136]; See Table 1 in [136] for a comparison of the conditions required for these works. Despite the recent progress, it appears that the aforementioned works still pose very restrictive assumptions on the problem types in order to achieve convergence. For example it is not clear whether the ADMM can be used for the distributed nonconvex optimization problem (4.7) over an arbitrary connected graph, despite the fact that for convex problem such application is popular, and the resulting algorithms are efficient.

4.1.2.2 Literature on Applications.

The sparse subspace estimation problem formulations (4.2) and (4.5) have been first considered in [30, 135] and subsequently considered in [49]. The work [135] proposes a semidefinite convex optimization problem to estimate principal subspace of a population matrix Σ based on a sample covariance matrix. The authors of [49] further show that by utilizing nonconvex regularizers it is possible to significantly improve the estimation accuracy for a given number of data points. However, the algorithm considered in [49] is not guaranteed to reach any stationary solutions.

The consensus problem (4.6) and (4.7) have been studied extensively in the literature when the objective functions are all convex; see for example [100, 88, 99, 122, 11]. Without assuming convexity of f_i's, the literature has been very scant; see recent developments in [17, 63, 55, 92]. However, all of these recent results require that the nonsmooth terms h_i's, if present, have to be identical. This assumption is unnecessarily strong and it defeats the purpose of *distributed* consensus since *global* information about the objective function has to be shared among the agents. Further, in the nonconvex setting we are not aware of any existing distributed algorithm with convergence guarantee that can deal with the more practical problem (4.8) with partial consensus.

4.1.2.3 Contributions of This work.

In this paper we develop an AL-based algorithm, named the perturbed proximal primal dual algorithm (PProx-PDA), for the challenging linearly constrained nonconvex nonsmooth problem (4.1). The proposed method is of Uzawa type [74] and has very simple update rule. It is a *single-loop* algorithm that alternates between a primal (scaled) proximal gradient descent step, and an (approximate) dual gradient ascent step. Further, by appropriately selecting the scaling matrix in the primal step, the variables can be easily updated in parallel. These features make the algorithm attractive for applications such as the high-dimensional subspace estimation and the distributed learning problems discussed in Section 4.1.1,

One distinctive feature of the PProx-PDA is the use of a novel perturbation scheme for both the primal and dual steps, which is designed to ensure a number of asymptotic convergence and rate of convergence properties (to approximate first-order stationary solutions). Specifically, we show that when certain perturbation parameter remains *constant* across the iterations, the algorithm converges globally sublinearly to the set of approximate first-order stationary solutions. Further, when the perturbation parameter diminishes to zero with appropriate rate, the algorithm converges to the set of exact first-order stationary solutions. To the best of our knowledge, the proposed algorithm represents one of the first first-order methods with convergence and rate of convergence guarantees for problems in the form of (4.1).

Notation. We use $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_F$ to denote the Euclidean norm, ℓ_1-norm, and Frobenius norm respectively. For given vector x, and matrix H, we denote $\|x\|_H^2 := x^T H x$. For two vectors a, b we use $\langle a, b\rangle$ to denote their inner product. We use $\sigma_{\max}(A)$ to denote the maximum eigenvalue for a matrix A. We use I_N to denote an $N \times N$ identity matrix. For a nonsmooth convex function $h(x)$, $\partial h(x)$ denotes the subdifferential set defined by

$$\partial h(x) = \{v \in \mathbb{R}^N; h(x) \geq h(y) + \langle v, x - y\rangle \ \forall y \in \mathbb{R}^N\}. \tag{4.9}$$

For a convex function $h(x)$ and a constant $\alpha > 0$ the proximity operator is defined as below

$$\text{prox}_h^{1/\alpha}(x) := \underset{z}{\text{argmin}} \frac{1}{2\alpha}\|x - z\|^2 + h(z). \tag{4.10}$$

4.2 Perturbed Proximal Primal Dual Algorithm

To begin with, we introduce the the augmented Lagrangian for problem (4.1)

$$L_\rho(x, y) = f(x) + h(x) + \langle \lambda, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2, \tag{4.11}$$

where $\lambda \in \mathbb{R}^M$ is the dual variable associated with the equality constraint $Ax = b$, and $\rho > 0$ is the penalty parameter for the augmented term $\|Ax - b\|^2$.

For notational simplicity, define $u(x; y) := \langle \nabla f(y), x - y\rangle$ to be the linear approximation of $f(x)$. Define $B \in \mathbb{R}^{M\times N}$ to be a scaling matrix, and introduce two new parameter $\gamma > 0$ and $\beta > 0$, where γ is a small positive number which is related to the size of the equality constraint violation, and β is the proximal parameter that regularizes the primal update. Let us choose $\gamma > 0$ and $\rho > 0$ such that $\rho\gamma < 1$. The steps of the proposed PProx-PDA algorithm is given below (Algorithm 5).

Algorithm 5 The perturbed proximal primal-dual algorithm (PProx-PDA)

Initialize: λ^0 and x^0
Repeat: update variables by

$$\begin{aligned} x^{r+1} = \arg\min_{x\in X}\ & u(x, x^r) + h(x) + \langle (1 - \rho\gamma)\lambda^r, Ax - b\rangle \\ & + \frac{\rho}{2}\|Ax - b\|^2 + \frac{\beta}{2}\|x - x^r\|^2_{B^TB} \end{aligned} \tag{4.12a}$$

$$\lambda^{r+1} = (1 - \rho\gamma)\lambda^r + \rho\left(Ax^{r+1} - b\right) \tag{4.12b}$$

Until Convergence.

In contrast to the AL method, in which the primal variable is updated by minimizing the augmented Lagrangian given in (4.11), in PProx-PDA the primal step minimizes an approximated augmented Lagrangian, where the approximation comes from: 1) replacing function $f(x)$ with the surrogate function $u(x, x^r)$; 2) perturbing λ by a positive factor $1 - \rho\gamma > 0$; 3) adding proximal term $\frac{\beta}{2}\|x - x^r\|^2_{B^TB}$. We make a few remarks about these algorithmic choices. First, the use of the linear surrogate function $u(x, y) := \langle \nabla f(y), x - y\rangle$ ensures that only first-order information is used for the primal update. Also it is worth mentioning that one can replace the function $u(x; y)$ with a wider class of "surrogate" functions satisfying certain gradient consistent conditions [109, 119],

and our subsequent analysis will still hold true. However, in order to stay focused, we choose not to present those variations. Second, the primal and dual perturbation is added to facilitate convergence analysis. Note that in the convex case, similar perturbation in the dual step has been considered for example in [101], but the purpose is to make sure that the dual variable lies in a convex and compact set under the Slater constraint qualification. Third, the appropriate choice of scaling matrix B ensures the following key properties:

a) Problem (4.12a) is strongly convex;

b) Problem (4.12a) is decomposable over different variables (or variable blocks).

Point (a) is relatively easy to see since ρ and β can be chosen to be large enough, and B^TB can be chosen to satisfy $A^TA + B^TB \succeq I$, so that the strongly convex regularization dominates the nonconvex function $f(x)$. We illustrate Point (b) through the distributed optimization problem (4.7). Let us define the *signless incidence matrix* $B := |A|$, where A is the signed incidence matrix defined in Section 4.1.1, and the absolute value is taken for each component of A. Using this choice of B, we have $B^TB = L_+ \in \mathbb{R}^{N\times N}$, which is the signless graph Laplacian whose (i,i)th diagonal entry is the degree of node i, and its (i,j)th entry is 1 if $e = (i,j) \in \mathcal{E}$, and 0 otherwise. Further, let us set $\rho = \beta$. Then x-update step (4.12a) becomes

$$x^{r+1} = \arg\min_x \sum_{i=1}^N \langle \nabla f_i(x_i^r), x_i - x_i^r\rangle + \langle (1-\rho\gamma)\lambda^r, Ax - b\rangle + \rho x^T D x - \rho x^T L_+ x^r$$

where $D := \mathrm{diag}[d_1, \cdots, d_N] \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix, with d_i denoting the degree of node i. Clearly this problem is separable over the variable x_i for all $i = 1, 2, \cdots, N$.

4.2.1 Convergence Analysis

In this subsection we provide the convergence analysis for PProx-PDA presented in Algorithm 5. We will frequently use the following identity

$$\langle b, b-a\rangle = \frac{1}{2}\left(\|b-a\|^2 + \|b\|^2 - \|a\|^2\right). \tag{4.13}$$

Also, for the notation simplicity we define

$$w^r := (x^{r+1} - x^r) - (x^r - x^{r-1}). \tag{4.14}$$

To proceed, let us make the following blanket assumptions on problem (4.1).

Assumption A.

A1. The function $f(x)$ is L-smooth i.e., there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall\ x, y \in X. \tag{4.15}$$

Further, without loss of generality, assume that $f(x) \geq 0$ for all $x \in X$.

A2. The function $h(x)$ is nonsmooth lower semi-continuous convex function, and it is assumed to be lower bounded: $h(x) \geq 0,\ \forall\ x \in X$.

A3. The constraint $Ax = b$ is feasible over $x \in X$.

A4. The feasible set X is a convex and compact set.

A5. The scaling matrix B is chosen such that $A^TA + B^TB \succeq I$.

Our first lemma characterizes the behavior of the dual variable.

Lemma 11 *Under Assumption A, the following holds true for PProx-PDA*

$$\begin{aligned}
&\frac{1-\rho\gamma}{2\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\beta}{2}\|x^{r+1} - x^r\|^2_{B^TB} \\
&\leq \frac{1-\rho\gamma}{2\rho}\|\lambda^{r} - \lambda^{r-1}\|^2 + \frac{\beta}{2}\|x^{r} - x^{r-1}\|^2_{B^TB} \\
&\quad + \frac{L}{2}\|x^{r+1} - x^r\|^2 + \frac{L}{2}\|x^r - x^{r-1}\|^2 - \gamma\|\lambda^{r+1} - \lambda^r\|^2, \quad \forall\ r \geq 1.
\end{aligned} \tag{4.16}$$

Proof 1 *From the optimality condition of the x-update in (4.12a) we obtain*

$$\begin{aligned}
&\langle \nabla f(x^r) + A^T\lambda^r(1-\rho\gamma) + \rho A^T(Ax^{r+1} - b) \\
&\quad + \beta B^TB(x^{r+1} - x^r) + \xi^{r+1}, x^{r+1} - x\rangle \leq 0,\ \forall\ x \in X,
\end{aligned} \tag{4.17}$$

where $\xi^{r+1} \in \partial h(x^{r+1})$ is a subgradint for nonsmooth function $h(x)$ at $x = x^{r+1}$. Performing this equation for $r-1$, we get

$$\begin{aligned}
&\langle \nabla f(x^{r-1}) + A^T\lambda^{r-1}(1-\rho\gamma) + \rho A^T(Ax^{r} - b) \\
&\quad + \beta B^TB(x^{r} - x^{r-1}) + \xi^{r}, x^{r} - x\rangle \leq 0,\ \forall\ x \in X,
\end{aligned} \tag{4.18}$$

where $\xi^r \in \partial h(x^r)$ *is defined similarly. Let* $x = x^r$ *in the first inequality and* $x = x^{r+1}$ *in the second, we can then add the resulting inequalities to obtain*

$$\begin{aligned}&\langle \nabla f(x^r) - \nabla f(x^{r-1}), x^{r+1} - x^r \rangle + \langle A^T(\lambda^{r+1} - \lambda^r), x^{r+1} - x^r \rangle \\ &+ \beta \langle B^T B w^r, x^{r+1} - x^r \rangle \leq \langle \xi^r - \xi^{r+1}, x^{r+1} - x^r \rangle \leq 0 \end{aligned} \tag{4.19}$$

where in the last inequality we have utilized the convexity of h*. Now let us analyze each terms in* (4.19)*. First, by the application of Young's inequality, and the assumption that* f *is* L*-smooth we have*

$$\langle \nabla f(x^{r-1}) - \nabla f(x^r), x^{r+1} - x^r \rangle \leq \frac{L}{2}\|x^{r+1} - x^r\|^2 + \frac{L}{2}\|x^r - x^{r-1}\|^2. \tag{4.20}$$

For the second term, we have the following series of equalities

$$\begin{aligned}&\langle A^T(\lambda^{r+1} - \lambda^r), x^{r+1} - x^r \rangle = \langle A(x^{r+1} - x^r), \lambda^{r+1} - \lambda^r \rangle \\ &= \langle (Ax^{r+1} - b - \gamma\lambda^r) - (Ax^r - b - \gamma\lambda^{r-1}), \lambda^{r+1} - \lambda^r \rangle + \gamma \langle \lambda^r - \lambda^{r-1}, \lambda^{r+1} - \lambda^r \rangle \\ &\overset{(4.12b),(4.13)}{=} \frac{1}{2}\left(\frac{1}{\rho} - \gamma\right)\Big(\|\lambda^{r+1} - \lambda^r\|^2 - \|\lambda^r - \lambda^{r-1}\|^2 \\ &\quad + \|(\lambda^{r+1} - \lambda^r) - (\lambda^r - \lambda^{r-1})\|^2\Big) + \gamma\|\lambda^{r+1} - \lambda^r\|^2. \end{aligned} \tag{4.21}$$

For the third term, we have

$$\begin{aligned}&\beta \langle B^T B w^r, x^{r+1} - x^r \rangle \overset{(4.13)}{=} \frac{\beta}{2}\left(\|x^{r+1} - x^r\|^2_{B^TB} - \|x^r - x^{r-1}\|^2_{B^TB} + \|w^r\|^2_{B^TB}\right) \\ &\geq \frac{\beta}{2}\left(\|x^{r+1} - x^r\|^2_{B^TB} - \|x^r - x^{r-1}\|^2_{B^TB}\right). \end{aligned} \tag{4.22}$$

Therefore, combining (4.20) – (4.22)*, we obtain the desired result in* (4.16)*.* **Q.E.D.**

Next we analyze the behavior of the primal iterations. Towards this end, let us define the following new quantity

$$T(x, \lambda) := f(x) + h(x) + \langle (1 - \rho\gamma)\lambda, Ax - b - \gamma\lambda \rangle + \frac{\rho}{2}\|Ax - b\|^2. \tag{4.23}$$

Note that this quantity is identical to the augmented Lagrangian when $\gamma = 0$. It is constructed to track the behavior of the algorithm. The next lemma analyzes the change of T in two successive iterations of the algorithm.

Lemma 12 *Suppose that $\beta > 3L$ and $\rho \geq \beta$. Then we have the following*

$$\begin{aligned} & T(x^{r+1}, \lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2 \\ & \leq T(x^r, \lambda^r) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^r\|^2 + \left(\frac{(1-\rho\gamma)(2-\rho\gamma)}{2\rho}\right)\|\lambda^{r+1} - \lambda^r\|^2 \\ & - \left(\frac{\beta - 3L}{2}\right)\|x^{r+1} - x^r\|^2, \quad \forall\ r \geq 0. \end{aligned} \tag{4.24}$$

Proof 2 *For simplicity let us define $g(x, \lambda; x^r) := T(x, \lambda) + \frac{\beta}{2}\|x - x^r\|_{B^TB}$. Then it is easy to see that if $\beta \geq 3L$, then the change of x results in the reduction of T:*

$$\begin{aligned} & T(x^{r+1}, \lambda^r) - T(x^r, \lambda^r) = g(x^{r+1}, \lambda^r; x^r) - g(x^r, \lambda^r; x^r) - \frac{\beta}{2}\|x^{r+1} - x^r\|_{B^TB} \\ & \overset{\text{(i)}}{\leq} \langle \nabla f(x^{r+1}) + \xi^{r+1} + (1-\rho\gamma)A^T\lambda^r + \rho A^T(Ax^{r+1} - b) + \beta B^TB(x^{r+1} - x^r), \\ & \qquad x^{r+1} - x^r\rangle - \frac{\beta - L}{2}\|x^{r+1} - x^r\|^2 \\ & \overset{\text{(ii)}}{\leq} -\left(\frac{\beta - 3L}{2}\right)\|x^{r+1} - x^r\|^2, \end{aligned} \tag{4.25}$$

where (i) *is true because when $\beta \geq 3L$, $\rho \geq \beta$ and $A^TA + B^TB \succeq I$, function $g(x, \lambda; x^r)$ is strongly convex with modulus $\beta - L$ [here $\xi^{r+1} \in \partial h(x^{r+1})$];* (ii) *is true due to the optimality condition* (4.17) *for x-subproblem, and the assumption that $f(x)$ is L-smooth. Second, let us analyze $T(x^{r+1}, \lambda^{r+1}) - T(x^{r+1}, \lambda^r)$ as the following*

$$\begin{aligned} & T(x^{r+1}, \lambda^{r+1}) - T(x^{r+1}, \lambda^r) \\ & = (1-\rho\gamma)\left(\langle \lambda^{r+1} - \lambda^r, Ax^{r+1} - b - \gamma\lambda^r\rangle\right) - (1-\rho\gamma)\langle \gamma\lambda^{r+1} - \gamma\lambda^r, \lambda^{r+1}\rangle \\ & \overset{(4.12b),(4.13)}{=} (1-\rho\gamma)\left(\frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\gamma}{2}(\|\lambda^r\|^2 - \|\lambda^{r+1}\|^2 - \|\lambda^{r+1} - \lambda^r\|^2)\right). \end{aligned} \tag{4.26}$$

Combining the previous two steps, we obtain the desired inequality in (4.24). **Q.E.D.**

Comparing the results of two previous lemmas, from (4.16) we can observe that term $\frac{1}{2}(\frac{1}{\rho} - \gamma)\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\beta}{2}\|x^{r+1} - x^r\|^2_{B^TB}$ is descending in $\|\lambda^{r+1} - \lambda^r\|^2$ and ascending in $\|x^{r+1} - x^r\|^2$, while from (4.24) we can see that $T(x^{r+1}, \lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2$ is behaving in an opposite manner. Therefore, let us define the following potential function P_c as a conic combination of these two terms such that it is descending in each iteration. For some $c > 0$

$$P_c(x^{r+1},\lambda^{r+1};x^r,\lambda^r) := T(x^{r+1},\lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2 + \frac{c}{2}\left(\frac{1-\rho\gamma}{\rho}\|\lambda^{r+1}-\lambda^r\|^2 + \beta\|x^{r+1}-x^r\|^2_{B^TB} + L\|x^{r+1}-x^r\|^2\right). \tag{4.27}$$

Then according to the previous two lemmas, one can conclude there are constants a_1, a_2, such that

$$\begin{aligned} &P_c(x^{r+1},\lambda^{r+1};x^r,\lambda^r) - P_c(x^r,\lambda^r;x^{r-1},\lambda^{r-1}) \\ &\leq -a_1\|\lambda^{r+1}-\lambda^r\|^2 - a_2\|x^{r+1}-x^r\|^2, \end{aligned} \tag{4.28}$$

where $a_1 = \left((1-\rho\gamma)\frac{\gamma}{2} + c\gamma - \frac{1-\rho\gamma}{\rho}\right)$, and $a_2 = (\frac{\beta-3L}{2} - cL)$. Therefore, it is easy to observe that in order to make the P_c function descent, it is sufficient to have

$$(1-\rho\gamma)\frac{\gamma}{2} + c\gamma - \frac{1-\rho\gamma}{\rho} > 0, \text{ and } \beta > (3+2c)L. \tag{4.29}$$

Therefore a sufficient condition is that

$$\tau := \rho\gamma \in (0,1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+2c)L, \quad \rho \geq \beta. \tag{4.30}$$

Next, let us show that the potential function P_c is lower bounded, when choosing particular parameters given in the previous lemma.

Lemma 13 *Suppose Assumption A is satisfied, and the algorithm parameters are chosen according to (4.30). Then the following Statement holds true*

$$\exists\ \underline{P} \quad \text{s.t. } P_c(x^{r+1},\lambda^{r+1};x^r,\lambda^r) \geq \underline{P} > -\infty, \quad \forall\ r \geq 0. \tag{4.31}$$

Proof 3 *First, we analyze terms related to $T(x^{r+1},\lambda^{r+1})$. The inner product term in (4.23) can be bounded as*

$$\begin{aligned} &\langle \lambda^{r+1} - \rho\gamma\lambda^{r+1}, Ax^{r+1} - b - \gamma\lambda^{r+1}\rangle \\ &\overset{(4.13)}{=} \frac{1}{2}\left(\frac{1-\rho\gamma}{\rho} - (1-\rho\gamma)\gamma\right)(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1}-\lambda^r\|^2) \\ &= \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1}-\lambda^r\|^2). \end{aligned} \tag{4.32}$$

Clearly, the constant in front of the above equality is positive. Taking a sum over R iterations of $T(x^{r+1}, \lambda^{r+1})$, we obtain

$$\begin{aligned}
&\sum_{r=1}^{R} T(x^{r+1}, \lambda^{r+1}) = \sum_{r=1}^{R} \Big(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2 \Big) \\
&+ \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2 + \sum_{r=1}^{R} \|\lambda^{r+1} - \lambda^r\|^2) \\
&\geq \sum_{r=1}^{R} \Big(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2 \Big) + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2) \\
&\geq -\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2, \qquad (4.33)
\end{aligned}$$

where the last inequality comes from the fact that f and h are both assumed to be lower bounded by 0. Therefore, the sum of the $T(\cdot,\cdot)$ function is lower bounded. From (4.33) we conclude that $\sum_{r=1}^{R} P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r)$ is also lower bounded by $-\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2$ for any R, because besides the term $\sum_{r=1}^{R} T(x^{r+1}, \lambda^{r+1})$, the rest of the terms are all positive. Combined with the fact that P_c is nonincreasing we conclude that the potential function is lower bounded, that is we have

$$\underline{P} \geq -\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2. \qquad (4.34)$$

This proves the claim. **Q.E.D.**

Now we are ready to present the main result on the convergence of the PProx-PDA. To proceed, let us define some approximate stationary solutions for problem (4.1).

Definition 2 Stationary solution. *Consider problem (4.1). Given $\epsilon > 0$, the tuple (x^*, λ^*) is an ϵ-stationary solution if the following holds*

$$\|Ax^* - b\|^2 \leq \epsilon, \quad \langle \nabla f(x^*) + A^T\lambda^* + \xi^*, x^* - x\rangle \leq 0, \ \ \forall\ x \in X, \qquad (4.35)$$

where ξ^* is some vector that satisfies $\xi^* \in \partial h(x^*)$.

We note that the ϵ-stationary solution slightly violates the constraint violation. This definition is closely related to the approximate KKT (AKKT) condition in the existing literature [6, 53]. It can be verified that when $X = \mathbb{R}^N$, and $h \equiv 0$, then the condition in (4.35) satisfies the stopping

criteria for reaching AKKT condition Eq. (9)-(11) in [6]. We refer the readers to [6, Section 3.1] for detailed discussion of the relationship between AKKT and KKT conditions.

We show below that by appropriately choosing the algorithm parameters, the PProx-PDA in fact converges to the set of approximate stationary solutions. The precise relationship to convergence to ϵ-stationary solutions involves bounding the size of the dual solutions, as well as the choice of various algorithm parameters. These issues will be discussed in the next subsection.

Theorem 7 *Suppose Assumption A holds. Further assume that the parameters γ, ρ, β, c satisfy (4.30). For any given $\epsilon > 0$, the following is true for the sequence (x^r, λ^r) generated by the PProx-PDA*

- *In the limit we have*

$$\lambda^{r+1} - \lambda^r \to 0, \quad x^{r+1} - x^r \to 0.$$

- *Let (x^*, λ^*) denote any limit point of the sequence (x^r, λ^r). Then (x^*, λ^*) is a $(\gamma^2\|\lambda^*\|^2)$-stationary solution of problem (4.1).*

Proof 4 *Combining the bound given in (4.28) with the fact that the potential function P_c is decreasing and lower bounded, we immediately conclude that*

$$\lambda^{r+1} - \lambda^r \to 0, \quad x^{r+1} - x^r \to 0. \tag{4.36}$$

which proves the first part.

In order to prove the second part let (x^*, λ^*) be any limit point of the sequence (x^r, λ^r). From (4.12b) we have $\lambda^{r+1} - \lambda^r = \rho(Ax^{r+1} - b - \gamma\lambda^r)$. Then combining this with (4.36) we obtain

$$Ax^* - b - \gamma\lambda^* = 0. \tag{4.37}$$

Thus, we have $\|Ax^* - b\|^2 \leq \gamma^2\|\lambda^*\|^2$; which proves the first inequality in (4.35). Further, from the optimality condition of x-subproblem we have

$$\begin{aligned}\langle \nabla f(x^r) + \xi^r + (1 - \rho\gamma)A^T\lambda^r + \rho A^T(Ax^{r+1} - b) \\ + \beta B^T B(x^{r+1} - x^r), x^{r+1} - x\rangle \leq 0, \ \forall\ x \in X,\end{aligned} \tag{4.38}$$

where $\xi^r \in \partial h(x^r)$. *From the dual variable update we have* $\lambda^{r+1} = (1-\rho\gamma)\lambda^r + \rho\left(Ax^{r+1} - b\right)$, *therefor combining this with the previous equation we have*

$$\langle \nabla f(x^r) + \xi^r + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r), x^{r+1} - x\rangle \leq 0, \ \forall\ x \in X. \tag{4.39}$$

This inequality together with (4.36) *implies* (4.35).

4.2.2 The Choice of Perturbation Parameter

In this section, we discuss how to obtain ϵ-stationary solution. First, note that Theorem 7 indicates that if λ^* is bounded, and the bound is independent of the choice of parameters γ, ρ, β, c, then one can choose $\gamma = \mathcal{O}(\sqrt{\epsilon})$ to reach an ϵ optimal solution. Such boundedness of λ^* can be ensured by assuming certain constraint qualification (CQ) at (x^*, λ^*); see a related discussion in the Appendix. In the rest of this section, we take an alternative approach to argue ϵ-stationary solution. Our general strategy is to let $\frac{1}{\rho}$ and γ proportional to the accuracy parameter ϵ, while keeping $\tau = \rho\gamma \in (0,1)$ and c fixed to some ϵ-independent constants.

Let us define the following constants for problem (4.1)

$$\begin{aligned} &d_1 = \max\{\|Ax - b\|^2 \mid x \in X\}, \quad d_2 = \max\{\|x - y\|^2 \mid x, y \in X\}, \\ &d_3 = \max\{\|x - y\|^2_{B^TB} \mid x, y \in X\}, \quad d_4 = \max\{f(x) + h(x) \mid x \in X\}. \end{aligned} \tag{4.40}$$

The lemma below provides a parameter independent bound for $\frac{\rho}{2}\|Ax^1 - b\|^2$.

Lemma 14 *Suppose* $\lambda^0 = 0, Ax^0 = b$, $\rho \geq \beta$, *and* $\beta - 3L > 0$. *Then we have*

$$\frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4, \quad \frac{\beta}{2}\|x^1 - x^0\|^2 \leq d_4 + \frac{3L}{2}d_2 \tag{4.41}$$

Proof 5 *From Lemma 12, and use the choice of* x^0 *and* λ^0, *we obtain*

$$\begin{aligned} &T(x^1, \lambda^1) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^1\|^2 + \frac{\beta - 3L}{2}\|x^1 - x^0\|^2 \\ &\leq T(x^0, \lambda^0) + \left(\frac{1-\rho\gamma}{\rho} - \frac{\gamma}{2}(1-\rho\gamma)\right)\|\lambda^1\|^2. \end{aligned}$$

Utilizing the definition of $T(x, \lambda)$ *and* (4.32), *we obtain*

$$T(x^1,\lambda^1) = f(x^1) + h(x^1) + \frac{(1-\rho\gamma)^2}{\rho}\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2$$
$$T(x^0,\lambda^0) = f(x^0) + h(x^0).$$

Combining the above, we obtain

$$\left((1-\rho\gamma)\gamma - \frac{1-\rho\gamma}{\rho} + \frac{(1-\rho\gamma)^2}{\rho}\right)\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2 + \frac{\beta - 3L}{2}\|x^1 - x^0\|^2$$
$$\leq T(x^0,\lambda^0) - f(x^1) - h(x^1)$$

By simple calculation we can show that $\left((1-\rho\gamma)\gamma - \frac{1-\rho\gamma}{\rho} + \frac{(1-\rho\gamma)^2}{\rho}\right) = 0$. *By using the assumption* $f(x^1) \geq 0$, $h(x^1) \geq 0$, *it follows that*

$$\frac{\beta - 3L}{2}\|x^1 - x^0\|^2 \leq d_4, \quad \frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4. \tag{4.42}$$

This leads to the desired claim. **Q.E.D.**

Combining the above lemma with dual update (4.12b), we can conclude that

$$\frac{1}{2\rho}\|\lambda^1\|^2 = \frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4. \tag{4.43}$$

Next, we derive an upper bound for the initial potential function $P_c(x^1,\lambda^1;x^0,\lambda^0)$. Assuming that $Ax^0 = b$, $\lambda^0 = 0$, we have

$$\begin{aligned}
P_c(x^1,\lambda^1;x^0,\lambda^0) &\overset{(4.27)}{=} T(x^1,\lambda^1) + \frac{(1-\rho\gamma)(\gamma + c/\rho)}{2}\|\lambda^1\|^2 \\
&\quad + \frac{c}{2}\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\
&\overset{(4.23),(4.32)}{\leq} f(x^1) + h(x^1) + \frac{(1-\rho\gamma)^2}{\rho}\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2 \\
&\quad + \frac{(1-\rho\gamma)(\gamma + c/\rho)}{2}\|\lambda^1\|^2 + \frac{c}{2}\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\
&\overset{(4.41)}{\leq} \left[2 + 2(1-\rho\gamma)^2 + (1-\rho\gamma)(c+\rho\gamma)\right]d_4 + \frac{c}{2}\left(2\sigma_{\max}(B^TB)(d_4 + \frac{3L}{2}d_2) + Ld_2\right) \\
&:= P_c^0
\end{aligned} \tag{4.44}$$

It is important to note that P_c^0 does not depend on ρ, γ, β individually, but only on $\rho\gamma$ and c, both of which can be chosen as absolute constants. The next lemma bounds the size of $\|\lambda^{r+1}\|^2$.

Lemma 15 *Suppose that (ρ, γ, β) are chosen according to (4.30), and the assumptions in Lemma 14 hold true. Then the following holds true for all $r \geq 0$*

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \leq P_c^0. \tag{4.45}$$

Proof 6 *We use induction to prove the lemma. The initial step $r = 0$ is clearly true. In the inductive step we assume that*

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r}\|^2 \leq P_c^0 \quad \text{for some } r \geq 1. \tag{4.46}$$

Using the fact that the potential function is decreasing (cf. (4.28)), we have

$$P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r) \leq P_c(x^1, \lambda^1; x^0, \lambda^0) \leq P_c^0. \tag{4.47}$$

Combining (4.47) with (4.32), and use the definition of P_c function in (4.27), we obtain

$$\frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) + \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \leq P_c^0. \tag{4.48}$$

If $\|\lambda^{r+1}\| - \|\lambda^r\| \geq 0$, then we have

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \leq \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) \overset{(4.48)}{\leq} P_c^0.$$

If $\|\lambda^{r+1}\| - \|\lambda^r\| < 0$, then we have

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 < \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r}\|^2 \leq P_c^0,$$

where the second inequality comes from the induction assumption (4.46). This concludes the proof of (4.45). **Q.E.D.**

From Lemma 15, and the fact that $\rho\gamma = \tau \in (0, 1)$, we have

$$\gamma\|\lambda^{r+1}\|^2 \leq \frac{2}{1-\tau}P_c^0, \quad \forall\ r \geq 0. \tag{4.49}$$

Therefore, we get

$$\gamma^2\|\lambda^{r+1}\|^2 \leq 2P_c^0\frac{\gamma}{1-\tau}, \quad \text{for all } r \geq 0\ . \tag{4.50}$$

Also note that ρ and β should satisfy (4.30), reStated below

$$\tau := \rho\gamma \in (0,1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+2c)L, \quad \rho \geq \beta. \tag{4.51}$$

Combining the above results, we have the following corollary about the choice of parameters to achieve ϵ-stationary solution.

Corollary 1 *Consider the following choices of algorithm parameters*

$$\gamma = \min\left\{\epsilon, \frac{1}{\beta}\right\}, \quad \rho = \frac{1}{2}\max\left\{\beta, \frac{1}{\epsilon}\right\}, \quad \beta > 7L, \quad c = 2. \tag{4.52a}$$

Further suppose Assumption A is satisfied, and that $Ax^0 = b$, $\lambda^0 = 0$. Then the sequence of dual variables $\{\lambda^r\}$ lies in a bounded set. Further, every limit point generated by the PProx-PDA algorithm is an ϵ-stationary solution.

Proof 7 *Using the parameters in (4.52a), we have*

$$\tau = \rho\gamma = \frac{1}{2}, \quad \frac{\gamma}{1-\rho\gamma} \leq 2\epsilon.$$

Then we can bound P_c^0 by the following

$$\begin{aligned} P_c^0 &= \left[2 + 2(1-\rho\gamma)^2 + (1-\rho\gamma)(c+\rho\gamma)\right] d_4 + \frac{c}{2}(2\sigma_{\max}(B^TB)(d_4 + \frac{3L}{2}d_2) + Ld_2) \\ &\leq (6 + 2\sigma_{\max}(B^TB))d_4 + (3\sigma_{\max}(B^TB)L + L)d_2. \end{aligned}$$

Therefore using (4.50) we conclude

$$\gamma^2\|\lambda^{r+1}\|^2 \leq 2P_c^0\frac{\gamma}{1-\rho\gamma} \leq 4((6 + 2\sigma_{\max}(B^TB))d_4 + (3\sigma_{\max}(B^TB)L + L)d_2)\epsilon.$$

Note that the constant in front of ϵ is not dependent on algorithm parameters. This implies that $\gamma^2\|\lambda^{r+1}\|^2 = \mathcal{O}(\epsilon)$. **Q.E.D.**

Remark 1 *First, in the above result, the ϵ-stationary solution is obtained by imposing the additional assumption that the initial solution is feasible, and that $\lambda^0 = 0$. Admittedly, obtaining a feasible initial solution could be challenging, but for problems such as distributed optimization (4.7) and subspace estimation (4.2), finding feasible x^0 is relatively easy. Second, the penalty parameter*

could be large because it is inversely proportional to the accuracy. Having a large penalty parameter at the beginning can make the algorithm progress slowly. In practice one can start with a smaller ρ and gradually increase it until reaching the predefined threshold. Following this idea, in the next section we will design an algorithm that allows ρ to increase unboundedly, such that in the limit the exact first-order stationary solution can be obtained.

Remark 2 *We comment that the convergence analysis carried out in this subsection differs from the recent analysis on nonconvex primal/dual type algorithms, which is first presented in Ames and Hong [5] and later generalized by [79, 63, 136, 97, 54]. Those analysis has been critically dependent on bounding the size of the successive dual variables with that of the successive primal variables. Unfortunately, this can only be done when the primal step immediately preceding the dual step is smooth and unconstrained. Therefore the algorithms and analysis presented in these works cannot be applied to our problem* (4.1)*, or the applications mentioned in Section 4.1.1.*

4.2.3 Convergence Rate Analysis

In this subsection we briefly discuss the convergence rate of the algorithm.

To begin with, assume that parameters are chosen according to (4.30), and $Ax^0 = b, \lambda^0 = 0$. Also we will choose $1/\rho$ and γ proportional to certain accuracy parameter, while keeping $\tau = \rho\gamma \in (0, 1)$ and c fixed to some absolute constants. To proceed, let us define the following quantities

$$H^r := f(x^r) + h(x^r) + \langle \lambda^r, Ax^r - b\rangle, \tag{4.53a}$$

$$G^{r+1} := \|\tilde{\nabla} H(x^r, \lambda^{r-1})\|^2 + \frac{1}{\rho^2}\|\lambda^{r+1} - \lambda^r\|^2, \tag{4.53b}$$

$$Q^r := \|\tilde{\nabla} H(x^r, \lambda^{r-1})\|^2 + \|Ax^r - b\|^2, \tag{4.53c}$$

where $\tilde{\nabla} H^r$ is the proximal gradient defined as

$$\tilde{\nabla} H^r = x^r - \mathrm{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla(H^r - h(x^r))\right]. \tag{4.54}$$

It can be checked that $Q(x^r, \lambda^{r-1}) \to 0$ if and only if a stationary solution for problem (4.1) is obtained. Therefore we say that an θ-stationary solution is obtained if $Q^r \le \theta$.

Note that the θ-stationary solution has been used in [63] for characterizing the rate for ADMM method. Compared with the ϵ-stationary solution defined in Definition 1, its progress is easier to quantify.

Using the definition of proximity operator, the optimality condition of the x-subproblem (4.12a) can be equivalently written as

$$x^{r+1} = \text{prox}^{\beta}_{h+\iota(X)}\left[x^{r+1} - \frac{1}{\beta}\left[\nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r)\right]\right].$$

By using the non-expansiveness of the prox operator, we obtain the following

$$\begin{aligned}
\|\tilde{\nabla} H^r\|^2 &= \left\|x^r - \text{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla\left[H^r - h(x^r)\right]\right]\right\|^2 \\
&= \left\|x^{r+1} - \text{prox}^{\beta}_{h+\iota(X)}\left[x^{r+1} - \frac{1}{\beta}\left[\nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r)\right]\right]\right. \\
&\quad \left. - x^r + \text{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla\left[H^r - h(x^r)\right]\right]\right\|^2 \\
&\le 2\|x^{r+1} - x^r\|^2 + \frac{4}{\beta^2}\|A^T(\lambda^{r+1} - \lambda^r)\|^2 + 4\|(I - B^T B)(x^{r+1} - x^r)\|^2 \\
&\le (2 + 4\sigma^2_{\max}(\hat{B}^T\hat{B}))\|x^{r+1} - x^r\|^2 + \frac{4\sigma_{\max}(A^T A)}{\beta^2}\|\lambda^{r+1} - \lambda^r\|^2,
\end{aligned}$$

where in the last inequality we define $\hat{B} := I - B^T B$. Therefore,

$$G^{r+1} \le b_1\|\lambda^{r+1} - \lambda^r\|^2 + b_2\|x^{r+1} - x^r\|^2, \tag{4.55}$$

where $b_1 = \frac{4\sigma_{\max}(A^T A)}{\beta^2} + \frac{1}{\rho^2}$, and $b_2 = 2 + 4\sigma^2_{\max}(\hat{B}^T\hat{B})$. Combining (4.55) with the descent estimate for the potential function P_c given in (4.28), we obtain

$$G^{r+1} \le V\left[P_c(x^r, \lambda^r; x^{r-1}, \lambda^{r-1}) - P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r)\right], \tag{4.56}$$

where we have defined

$$V := \frac{\max(b_1, b_2)}{\min(a_1, a_2)},$$

and one can check that V is in the order of $\mathcal{O}(1/\gamma)$ because a_1 is in the order of γ; cf. (4.28).

Let R denote the first time that G^{r+1} reaches below a given number $\theta > 0$. Summing both sides of (4.56) over R iterations, and utilizing the fact that P_c is lower bounded by $\underline{P}$, it follows that

$$\theta \le \frac{V\left(P_c^0 - \underline{P}\right)}{R} \overset{(4.34)}{\le} \frac{V\left(P_c^0 + \frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2\right)}{R} \overset{(4.43)}{\le} \frac{V\left(P_c^0 + (1-\tau)^2 d_4\right)}{R}$$

where d_4 is given in (4.40), and P_c^0 is given in (4.44). Note that $G^{r+1} \leq \theta$ implies that $1/\rho^2\|\lambda^{r+1} - \lambda^r\|^2 = \|Ax^{r+1} - b - \gamma\lambda^r\|^2 \leq \theta$. From (4.45) we have that

$$\|\gamma\lambda^{r+1}\| \leq \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}, \quad \forall\ r \geq 0.$$

It follows that

$$\|Ax^{r+1} - b\| \leq \frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\| + \|\gamma\lambda^r\| \leq \sqrt{\theta} + \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}.$$

It follows that whenever $G^{r+1} \leq \theta$ we have

$$Q^r := \|\tilde{\nabla}H(x^r, \lambda^{r-1})\|^2 + \|Ax^r - b\|^2 \leq \theta + \left(\sqrt{\theta} + \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}\right)^2. \tag{4.57}$$

Let us pick the parameters such that they satisfy (4.30) and the following

$$\frac{2P_c^0\gamma}{1-\rho\gamma} = \frac{2P_c^0\gamma}{1-\tau} = \theta.$$

Then whenever $G^{r+1} \leq \theta$, we have $Q^r \leq 5\theta$. It follows that the total number of iteration it takes for the stationarity gap Q^r to reach below 5θ is given by

$$R \leq \frac{V\left(P_c^0 + (1-\tau)^2 d_4\right)}{\theta} = \mathcal{O}\left(\frac{1}{\theta^2}\right), \tag{4.58}$$

where the last relation holds because V is in the order of $\mathcal{O}(\frac{1}{\gamma})$, γ is chosen in the order of $\mathcal{O}(\theta)$, and P_c^0, d_4 and τ are not dependent on the problem accuracy. The result below summarizes the our discussion.

Corollary 2 *Suppose that $Ax^0 = b$ and $\lambda^0 = 0$. Additionally, for a given $\theta > 0$, and $\tau \in (0,1)$, choose γ, ρ, c, β as follows*

$$\gamma = \frac{\theta(1-\tau)}{2P_c^0}, \quad \rho = \frac{\tau}{\gamma}, \qquad c > \frac{1}{\tau} - 1, \quad \rho \geq \beta, \quad \text{and } \beta > (3+2c)L.$$

Let R denote the first time that Q^r reaches below 5θ. Then we have $R = \mathcal{O}\left(\frac{1}{\theta^2}\right)$.

4.3 An Algorithm with Increasing Accuracy

So far we shown that PProx-PDA converges to the set of *approximate* stationary solutions by properly choosing the problem parameters following (4.30) and (1). The inaccuracy of the algorithm can be attributed to the use of perturbation parameter γ. Is it possible to gradually reduce the perturbation so that asymptotically the algorithm reaches the *exact* stationary solutions? Is it possible to avoid using very large penalty parameter ρ at the beginning of the algorithm? This section designs an algorithm that addresses these questions.

We consider a modified algorithm in which the parameters (ρ, β, γ) are *iteration-dependent*. In particular, we choose ρ^{r+1}, β^{r+1} and $1/\gamma^{r+1}$ to be increasing sequences. The new algorithm, named PProx-PDA with increasing accuracy (PProx-PDA-IA), is listed in Algorithm 6. Below

Algorithm 6 PProx-PDA with increasing accuracy (PProx-PDA-IA)

Initialize: λ^0 and x^0

Repeat: update variables by

$$x^{r+1} = \arg\min_{x\in X}\ u(x, x^r) + h(x) + \langle (1-\rho^{r+1}\gamma^{r+1})\lambda^r, Ax - b\rangle + \frac{\rho^{r+1}}{2}\|Ax-b\|^2 + \frac{\beta^{r+1}}{2}\|x - x^r\|^2_{B^TB}. \tag{4.59a}$$

$$\lambda^{r+1} = (1-\rho^{r+1}\gamma^{r+1})\lambda^r + \rho^{r+1}\left(Ax^{r+1} - b\right). \tag{4.59b}$$

Until Convergence.

we analyze the convergence of the new algorithm. Besides assuming that the optimization problem under consideration satisfies Assumption A, we make the following additional assumptions:

Assumption B

B1. Assume that

$$\rho^{r+1}\gamma^{r+1} = \tau \in (0,1), \quad \text{for some fixed constant } \tau.$$

B2. The sequence $\{\rho^r\}$ satisfies

$$\rho^{r+1} \to \infty, \ \sum_{r=1}^{\infty} \frac{1}{\rho^{r+1}} = \infty, \ \sum_{r=1}^{\infty} \frac{1}{(\rho^{r+1})^2} < \infty,$$
$$\rho^{r+1} - \rho^r = D > 0,$$

for some $D > 0$. A simple choice of ρ^{r+1} is $\rho^{r+1} = r + 1$. Similarly, the sequence $\{\gamma^{r+1}\}$ satisfies

$$\gamma^{r+1} - \gamma^r \leq 0, \ \gamma^{r+1} \to 0, \ \sum_{r=1}^{\infty} \gamma^{r+1} = \infty, \ \sum_{r=1}^{\infty} (\gamma^{r+1})^2 < \infty. \tag{4.60}$$

B3. Assume that

$$\beta^{r+1} \geq \beta^r, \quad \beta^{r+1} \to \infty, \text{ and } \sum_{r=1}^{\infty} \frac{1}{\beta^{r+1}} = \infty,$$
$$\exists\ c_0 > 1 \text{ s.t. } \beta^{r+1} = c_0 \rho^{r+1}, \ \text{ for } r \text{ large enough.} \tag{4.61}$$

B4. There exists $\Lambda > 0$ such that for every $r > 0$ we have $\|\lambda^r\| \leq \Lambda$.

We note that Assumption [B4] is somewhat restrictive because it is dependent on the iterates. In the Appendix we will show that such an assumption can be satisfied under some additional regularity conditions about problem (4.1). We choose to State [B4] here to avoid lengthy discussion on those regularity conditions before the main convergence analysis.

Similar to Lemma 11, our first step utilizes the optimality condition of two consecutive iterates to analyze the change of the primal and dual differences.

Lemma 16 *Suppose that the Assumptions A and [B1]-[B3] hold true, and that τ, D, are constants defined in assumption B. Then for large enough r, there exists constant C_1 such that*

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\leq\frac{(1-\tau)}{2}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^{r}}{\rho^{r+1}}-1)\|\lambda^{r}\|^2+\frac{\beta^{r}\rho^{r}}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&+\frac{\rho^{r}L}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}-\frac{\tau}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2\\
&+\frac{L\rho^r+D(L+\beta^{r+1}\|B^TB\|)}{2}\|x^{r+1}-x^r\|^2_{B^TB}-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}. \qquad (4.62)
\end{aligned}$$

Proof 8 *Suppose that $\xi^{r+1}\in\partial h(x^{r+1})$. From the optimality condition for x-subproblem (4.59a) we have for all $x\in dom(h)$*

$$\langle\nabla f(x^r)+A^T\lambda^{r+1}+\beta^{r+1}B^TB(x^{r+1}-x^r)+\xi^{r+1},x^{r+1}-x\rangle\leq 0.$$

Performing the above inequality for the $(r-1)th$ iteration, we have

$$\langle\nabla f(x^{r-1})+A^T\lambda^{r}+\beta^{r}B^TB(x^{r}-x^{r-1})+\xi^{r},x^{r}-x\rangle\leq 0,\ \forall\ x\in dom(h).$$

Plugging in $x=x^r$ in the first inequality, $x=x^{r+1}$ in the second inequality and add them together, and use the fact that h is convex, we obtain

$$\begin{aligned}
&\langle\nabla f(x^r)-\nabla f(x^{r-1})+A^T(\lambda^{r+1}-\lambda^r)\\
&+\beta^{r+1}B^TB(x^{r+1}-x^r)-\beta^rB^TB(x^r-x^{r-1}),x^{r+1}-x^r\rangle\leq 0. \qquad (4.63)
\end{aligned}$$

Let us analyze the above inequality term by term. First, using Young's inequality and the assumption that f is L-smooth i.e. (4.15) we have

$$\langle\nabla f(x^{r-1})-\nabla f(x^r),x^{r+1}-x^r\rangle\leq\frac{L}{2}\|x^{r+1}-x^r\|^2+\frac{L}{2}\|x^r-x^{r-1}\|^2.$$

Second, note that we have

$$\begin{aligned}
&\langle A^T(\lambda^{r+1}-\lambda^r), x^{r+1}-x^r\rangle = \langle \lambda^{r+1}-\lambda^r, A(x^{r+1}-x^r)\rangle \\
&= \langle \lambda^{r+1}-\lambda^r, Ax^{r+1}-b-\gamma^{r+1}\lambda^r+\gamma^{r+1}\lambda^r+\gamma^r\lambda^{r-1}-\gamma^r\lambda^{r-1}-Ax^r+b\rangle \\
&\overset{(4.59b)}{=} \langle \lambda^{r+1}-\lambda^r, \frac{\lambda^{r+1}-\lambda^r}{\rho^{r+1}}+\gamma^{r+1}\lambda^r-\gamma^r\lambda^{r-1}-\frac{\lambda^r-\lambda^{r-1}}{\rho^r}\rangle \\
&= \frac{1}{\rho^r}\langle \lambda^{r+1}-\lambda^r, \lambda^{r+1}-\lambda^r-(\lambda^r-\lambda^{r-1})\rangle + \left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}\right)\|\lambda^{r+1}-\lambda^r\|^2 \\
&\quad + \langle \lambda^{r+1}-\lambda^r, \lambda^r-\lambda^{r-1}\rangle\gamma^r + \langle \lambda^{r+1}-\lambda^r, \lambda^r\rangle(\gamma^{r+1}-\gamma^r) \\
&\overset{(4.13)}{=} \frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2+\|(\lambda^{r+1}-\lambda^r)-(\lambda^r-\lambda^{r-1})\|^2\right) \\
&\quad + \gamma^r\|\lambda^{r+1}-\lambda^r\|^2 + \left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}\right)\|\lambda^{r+1}-\lambda^r\|^2 \\
&\quad + \frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2-\|\lambda^{r+1}-\lambda^r\|^2).
\end{aligned}$$

Summarizing, we have

$$\begin{aligned}
\langle A^T(\lambda^{r+1}-\lambda^r), x^{r+1}-x^r\rangle &\geq \frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2\right) \\
&\quad + \frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2) \\
&\quad + \left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}+\gamma^r-\frac{1}{2}(\gamma^{r+1}-\gamma^r)\right)\|\lambda^{r+1}-\lambda^r\|^2 \\
&\overset{(B1)}{=} \frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2\right) \\
&\quad + \frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2) \\
&\quad + \left(\gamma^r-\left(\frac{1}{\tau}-\frac{1}{2}\right)(\gamma^r-\gamma^{r+1})\right)\|\lambda^{r+1}-\lambda^r\|^2.
\end{aligned}$$

Third, notice that

$$\begin{aligned}
&\langle \beta^{r+1}B^TB(x^{r+1}-x^r)-\beta^rB^TB(x^r-x^{r-1}), x^{r+1}-x^r\rangle \\
&\overset{(4.13)}{=} (\beta^{r+1}-\beta^r)\|x^{r+1}-x^r\|^2_{B^TB}+\frac{\beta^r}{2}\left(\|x^{r+1}-x^r\|^2_{B^TB}-\|x^r-x^{r-1}\|^2_{B^TB}+\|w^r\|^2_{B^TB}\right) \\
&= \frac{\beta^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}-\frac{\beta^r}{2}\|x^r-x^{r-1}\|^2_{B^TB}+\frac{\beta^{r+1}-\beta^r}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{\beta^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

Therefore, from the above three steps, we can bound (4.63) by

$$\begin{aligned}
&\frac{(1-\tau)}{2\rho^r}\|\lambda^{r+1}-\lambda^r\|^2+\frac{1}{2}(\gamma^{r+2}-\gamma^{r+1})\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{L}{2}\|x^{r+1}-x^r\|^2\\
&\leq \frac{(1-\tau)}{2\rho^r}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{1}{2}(\gamma^{r+1}-\gamma^{r})\|\lambda^{r}\|^2+\frac{\beta^{r}}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&+\frac{L}{2}\|x^r-x^{r-1}\|^2-\left(\gamma^r-(\frac{1}{\tau}-\frac{1}{2})(\gamma^r-\gamma^{r+1})\right)\|\lambda^{r+1}-\lambda^r\|^2+L\|x^{r+1}-x^r\|^2\\
&+\frac{1}{2}(\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r))\|\lambda^{r+1}\|^2-\frac{\beta^r}{2}\|w^r\|^2_{B^TB}. \qquad (4.64)
\end{aligned}$$

Multiplying ρ^r on both sides, we obtain

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\quad+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2\\
&\leq \frac{(1-\tau)}{2}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^{r}}{\rho^{r+1}}-1)\|\lambda^{r}\|^2+\frac{\beta^{r}\rho^r}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&\quad+\frac{\rho^rL}{2}\|x^r-x^{r-1}\|^2-\rho^r\left(\gamma^r-(\frac{1}{\tau}-\frac{1}{2})(\gamma^r-\gamma^{r+1})\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&\quad+L\rho^r\|x^{r+1}-x^r\|^2+\frac{\rho^r}{2}(\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r))\|\lambda^{r+1}\|^2\\
&\quad+\frac{(\beta^{r+1})(\rho^{r+1}-\rho^r)}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{L(\rho^{r+1}-\rho^r)}{2}\|x^{r+1}-x^r\|^2-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}$$

where we have used the following fact

$$0\geq\left(\frac{\rho^r}{\rho^{r+2}}-\frac{\rho^r}{\rho^{r+1}}\right)=\frac{\rho^r}{\rho^{r+1}}\left(\frac{\rho^{r+1}}{\rho^{r+2}}-1\right)\geq\left(\frac{\rho^{r+1}}{\rho^{r+2}}-1\right).$$

Further, by Assumption B we have $\rho^{r+1}-\rho^r=D$, also we have $\|x^{r+1}-x^r\|^2_{B^TB}\leq\|B^TB\|\|x^{r+1}-x^r\|^2$. Therefore, we reach

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\quad+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2\\
&\leq \frac{(1-\tau)}{2}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^{r}}{\rho^{r+1}}-1)\|\lambda^{r}\|^2+\frac{\beta^{r}\rho^r}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&\quad+\frac{\rho^rL}{2}\|x^r-x^{r-1}\|^2-\frac{\tau}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2\\
&\quad+\frac{L\rho^r+D(L+\beta^{r+1}\|B^TB\|)}{2}\|x^{r+1}-x^r\|^2-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}, \qquad (4.65)
\end{aligned}$$

where the last inequality is true using the following relations:

- *To bound the term* $\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r)$ *we have*

$$\begin{aligned}\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r) &= \left(\frac{\tau}{\rho^{r+2}}-\frac{\tau}{\rho^{r+1}}-\frac{\tau}{\rho^{r+1}}+\frac{\tau}{\rho^r}\right)\\ &= \tau D\frac{\rho^{r+2}-\rho^r}{\rho^r\rho^{r+1}\rho^{r+2}} = \frac{2\tau D^2}{\rho^r\rho^{r+1}\rho^{r+2}}.\end{aligned}$$

  Thus there exists a constant C_1 *such that*

$$\frac{\rho^r}{2}\left(\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r)\right)\leq\frac{C_1(\gamma^{r+1})^2}{2}.$$

- *For large enough* r

$$\tau\geq\frac{D(2-\tau)}{\rho^{r+1}}.$$

The proof of the lemma is complete. **Q.E.D.**

Now let us analyze the behavior of $T(x,\lambda)$ which is originally defined in (4.23) in order to bound the descent of the primal variable. In this case, because T is also a function of ρ and γ (which is also time varying), we denote it as $T(x,\lambda;\rho,\gamma)$.

Lemma 17 *Suppose that the Assumptions Assumptions A and [B1]-[B3] hold true,* τ *and* D *are constants defined in Assumption B. Then we have*

$$\begin{aligned}&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+D(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\ &\leq T(x^r,\lambda^r;\rho^{r+1},\gamma^{r+1})+\left((1-\tau)\frac{\gamma^{r+1}}{2}-\frac{D(\gamma^{r+1})^2}{2\tau}+D(\gamma^{r+1})^2\right)\|\lambda^r\|^2\\ &\quad-\left(\frac{\beta^{r+1}-3L}{2}\right)\|x^{r+1}-x^r\|^2\\ &\quad+(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{2\tau^2(1-\tau)}\right)\|\lambda^{r+1}-\lambda^r\|^2\\ &\quad+\frac{(1-\tau)(\gamma^{r+1}-\gamma^{r+2})}{2}\|\lambda^{r+1}\|^2+\frac{D(\gamma^{r+2})^2}{2}\|\lambda^{r+1}\|^2\\ &\quad+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2.\end{aligned}\tag{4.66}$$

Proof 9 *Following the same analysis as in](4.25), we have that the T function has the following descent when only changing the primal variable*

$$\begin{aligned} &T(x^{r+1},\lambda^r;\rho^{r+1},\gamma^{r+1}) - T(x^r,\lambda^r;\rho^{r+1},\gamma^{r+1}) \\ &\leq -\left(\frac{\beta^{r+1}-3L}{2}\right)\|x^{r+1}-x^r\|^2. \end{aligned} \tag{4.67}$$

Second, following (4.26), it is easy to verify that

$$\begin{aligned} &T(x^{r+1},\lambda^{r+1};\rho^{r+1},\gamma^{r+1}) - T(x^{r+1},\lambda^r;\rho^{r+1},\gamma^{r+1}) \\ &\leq (1-\tau)\left(\frac{\|\lambda^{r+1}-\lambda^r\|^2}{\rho^{r+1}} + \frac{\gamma^{r+1}}{2}\left(\|\lambda^r\|^2 - \|\lambda^{r+1}\|^2 - \|\lambda^{r+1}-\lambda^r\|^2\right)\right) \\ &\leq (1-\tau)\Bigg(\frac{\|\lambda^{r+1}-\lambda^r\|^2}{\rho^{r+1}} + \frac{\gamma^{r+1}}{2}\|\lambda^r\|^2 - \frac{\gamma^{r+2}}{2}\|\lambda^{r+1}\|^2 \\ &\quad - (\frac{\gamma^{r+1}}{2} - \frac{\gamma^{r+2}}{2})\|\lambda^{r+1}\|^2 - \frac{\gamma^{r+1}}{2}\|\lambda^{r+1}-\lambda^r\|^2\Bigg). \end{aligned} \tag{4.68}$$

The most involving step is the analysis of the change of T when the parameters ρ and γ are changed. We first have the following bound

$$\begin{aligned} &T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2}) - T(x^{r+1},\lambda^{r+1};\rho^{r+1},\gamma^{r+1}) \\ &= (1-\tau)(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2 + \frac{\rho^{r+1}-\rho^r}{2}\|Ax^{r+1}-b\|^2 \\ &= (1-\tau)(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2 \\ &+ \underbrace{\frac{D}{2}\|(Ax^{r+1}-b)-\gamma^{r+1}\lambda^r\|^2}_{(a)} - \underbrace{\frac{D}{2}\|\gamma^{r+1}\lambda^r\|^2}_{(b)} + \underbrace{D\langle\gamma^{r+1}\lambda^r, Ax^{r+1}-b\rangle}_{(c)}. \end{aligned} \tag{4.69}$$

The term (a) in (4.69) is given by

$$\frac{D}{2}\|(Ax^{r+1}-b)-\gamma^{r+1}\lambda^r\|^2 = \frac{D}{(\rho^{r+1})^2}\|\lambda^{r+1}-\lambda^r\|^2. \tag{4.70}$$

The term (b) in (4.69) is given by

$$-\frac{D}{2}\|\gamma^{r+1}\lambda^r\|^2 = -(\gamma^{r+1})^2\frac{D}{2}\|\lambda^r\|^2. \tag{4.71}$$

The term (c) in (4.69) is given by

$$\begin{aligned} &D\langle\gamma^{r+1}\lambda^r, Ax^{r+1}-b\rangle = D\langle\gamma^{r+1}\lambda^r, \frac{\lambda^{r+1}-\lambda^r}{\rho^{r+1}} + \gamma^{r+1}\lambda^r\rangle \\ &= D(\gamma^{r+1})^2\|\lambda^r\|^2 + D\frac{(\gamma^{r+1})^2}{2\tau}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 - \|\lambda^{r+1}-\lambda^r\|^2). \end{aligned} \tag{4.72}$$

So collecting terms, we have

$$\begin{aligned}
&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+\frac{D}{2}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
\leq& T(x^{r},\lambda^{r};\rho^{r+1},\gamma^{r+1})+\left((1-\tau)\frac{\gamma^{r+1}}{2}-\frac{D(\gamma^{r+1})^2}{2\tau}+\frac{D}{2}(\gamma^{r+1})^2\right)\|\lambda^{r}\|^2\\
&-(\frac{\beta^{r+1}-3L}{2})\|x^{r+1}-x^r\|^2\\
&+(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&+\frac{1-\tau}{2}(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2+(\gamma^{r+2})^2\frac{D}{2}\|\lambda^{r+1}\|^2\\
&+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2. \qquad (4.73)
\end{aligned}$$

The lemma is proved. **Q.E.D.**

In the next step we construct and estimate the descent of the potential function. For some given $c>0$, we construct the following potential function

$$\begin{aligned}
P_c^{r+1}:=&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2}) \qquad (4.74)\\
&+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+\frac{D}{2}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
&+c\Big(\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2\\
&+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\Big).
\end{aligned}$$

Lemma 18 *Suppose that the Assumptions A and [B1]-[B3] hold true, and let τ and D be the constants defined in Assumption B. Then for large enough r we have the following for the potential function P_c*

$$\begin{aligned}
P_c^{r+1}-P_c^r\leq&-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\
&-c\frac{\tau}{4}\|\lambda^{r+1}-\lambda^r\|^2+D_0(\gamma^{r+1})^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}, \qquad (4.75)
\end{aligned}$$

where D_0 is a positive constant.

Proof 10 *According to Lemma 16 and Lemma 17, for large enough r we have*

$$P_c^{r+1} - P_c^r \leq -\left(\frac{\beta^{r+1} - 3L}{2} - cL\rho^r - cDL - c\beta^{r+1}\|B^TB\|\right)\|x^{r+1} - x^r\|^2$$
$$-\left(c\frac{\tau}{2} - (1-\tau)\left(\frac{1}{\rho^{r+1}} - \frac{\gamma^{r+1}}{2} + \frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\right)\|\lambda^{r+1} - \lambda^r\|^2$$
$$+\frac{(1-\tau)(\gamma^{r+1} - \gamma^{r+2})}{2}\|\lambda^{r+1}\|^2 + \frac{D(\gamma^{r+2})^2}{2}\|\lambda^{r+1}\|^2 - c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}$$
$$+D\frac{(\gamma^{r+1})^2 - (\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2 + c\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2. \tag{4.76}$$

From the properties of perturbation parameter γ^r given in (4.60) we can observe that

$$\gamma^{r+1} - \gamma^{r+2} \leq \frac{D}{\tau}\gamma^{r+1}\gamma^{r+2} \leq \frac{D}{\tau}(\gamma^{r+1})^2.$$

Utilizing this result together with the Assumption [B6] related to dual variable λ, we obtain the following relations for large enough r

$$\frac{(1-\tau)(\gamma^{r+1} - \gamma^{r+2})}{2}\|\lambda^{r+1}\|^2 \leq D\frac{(1-\tau)(\gamma^{r+1})^2\Lambda}{2\tau}. \tag{4.77}$$

Similarly we also have

$$c\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2 \leq \frac{cC_1\Lambda(\gamma^{r+1})^2}{2}.$$

Moreover, since $(\gamma^{r+1})^2 - (\gamma^{r+2})^2 \leq (\gamma^{r+1})^2$, and $\gamma^{r+2} \leq \gamma^{r+1}$, we have

$$D\frac{(\gamma^{r+1})^2 - (\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2 \leq \frac{D\Lambda}{2\tau}(\gamma^{r+1})^2,$$
$$(\gamma^{r+2})^2\frac{D}{2}\|\lambda^{r+1}\|^2 \leq \frac{D\Lambda}{2}(\gamma^{r+1})^2. \tag{4.78}$$

Let us set

$$D_0 := \frac{D(1-\tau)\Lambda}{2\tau} + \frac{cC_1\Lambda}{2} + \frac{D\Lambda}{2\tau} + \frac{D\Lambda}{2},$$

which adds up the constants in front of $(\gamma^{r+1})^2$ in the above terms. We can therefore bound the difference of the potential function by

$$P_c^{r+1} - P_c^r \leq -\left(\frac{\beta^{r+1} - 3L}{2} - cL\rho^r - cDL - c\beta^{r+1}\|B^TB\|\right)\|x^{r+1} - x^r\|^2$$
$$-\left(c\frac{\tau}{2} - (1-\tau)\left(\frac{1}{\rho^{r+1}} - \frac{\gamma^{r+1}}{2} + \frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\right)\|\lambda^{r+1} - \lambda^r\|^2$$
$$+D_0(\gamma^{r+1})^2 - c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}. \tag{4.79}$$

Since $(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\to 0$, *we can find* r_0 *large enough such that for* $r\geq r_0$

$$(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{2\tau^2(1-\tau)}\right)\leq\frac{c\tau}{4}. \tag{4.80}$$

Thus, for $r\geq r_0$ *we have*

$$\begin{aligned}P_c^{r+1}-P_c^r\leq&-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\&-c\frac{\tau}{4}\|\lambda^{r+1}-\lambda^r\|^2+D_0(\gamma^{r+1})^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}.\end{aligned} \tag{4.81}$$

The claim is proved. **Q.E.D.**

Note that by Assumption B we have that

$$\sum_{r=1}^{\infty}(\gamma^{r+1})^2<\infty. \tag{4.82}$$

Therefore to ensure the potential function decrease eventually, we need to pick the constants in the following way [note that by (4.61), $c_0\rho^{r+1}=\beta^{r+1}$]

$$\frac{c_0\rho^{r+1}-3L}{2}-cL\rho^r-cDL-cc_0\rho^{r+1}\|B^TB\|\geq 0. \tag{4.83}$$

It is clear that if constant c is picked such that

$$0<c\leq\frac{c_0}{2(L+c_0\|B^TB\|)}. \tag{4.84}$$

Then the above inequality is satisfied for large enough r.

In this step we show that the potential function is lower bounded.

Lemma 19 *Suppose that the Assumptions A and [B1]-[B3] hold true, and that the constant* c *is chosen such that*

$$0<c\leq\frac{1-\tau}{D}. \tag{4.85}$$

Then the potential function P_c^r *defined in* (4.74) *is lower bounded.*

Proof 11 *Let us rearrange the terms of the potential function*

$$
\begin{aligned}
P_c^{r+1} &= T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2}) \\
&+ \left(\frac{(1-\tau)D(\gamma^{r+2})^2}{2\tau} + \frac{(1-\tau-cD)\gamma^{r+2}}{2} \right) \|\lambda^{r+1}\|^2 \\
&+ c\left(\frac{(1-\tau)}{2}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1} - x^r\|^2_{B^TB} + \frac{L\rho^{r+1}}{2}\|x^{r+1} - x^r\|^2 \right).
\end{aligned} \tag{4.86}
$$

First of all, we note that if we set $0 < c \leq \frac{1-\tau}{D}$ *then the coefficient in front of* $\|\lambda^{r+1}\|^2$ *is positive. Let us analyze* $T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$*. We have the following*

$$
\begin{aligned}
&\langle \lambda^{r+1} - \rho^{r+2}\gamma^{r+2}\lambda^{r+1}, Ax^{r+1} - b - \gamma^{r+2}\lambda^{r+1} \rangle \\
&= \frac{1-\tau}{\rho^{r+1}}\langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle + (1-\tau)\langle \lambda^{r+1}, \gamma^{r+1}\lambda^r - \gamma^{r+2}\lambda^{r+1} \rangle \\
&= \frac{1-\tau}{\rho^{r+1}}\langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle + (1-\tau)\gamma^{r+1}\langle \lambda^{r+1}, \lambda^r - \lambda^{r+1} \rangle \\
&+ (1-\tau)(\gamma^{r+1} - \gamma^{r+2})\|\lambda^{r+1}\|^2 \\
&\geq \left(\frac{1-\tau}{\rho^{r+1}} - (1-\tau)\gamma^{r+1} \right) \langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle \\
&= \frac{1}{2\rho^{r+1}}(1-\tau)^2(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2) \\
&\geq \frac{(1-\tau)^2}{2}\left(\frac{1}{\rho^{r+1}}\|\lambda^{r+1}\|^2 - \frac{1}{\rho^r}\|\lambda^r\|^2\right).
\end{aligned} \tag{4.87}
$$

It follows that the sum $\sum_{r=1}^{\infty} T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$ *is lower bounded. The claim can then be proved by using a similar argument as in Lemma 13.* **Q.E.D.**

Finally we put all the previous lemmas together to present the main convergence results for the PProx-PDA-IA.

Theorem 8 *Suppose that Assumptions A–B hold true, and that* τ*,* c *and* D *are picked such that (4.84) and (4.85) are satisfied. Then every limit point of the sequence generated by PProx-PDA-IA is a stationary solution of problem (4.1).*

Proof 12 *In this proof we pick a special case of* B *satisfying* $B^TB = I$*, in order to avoid unnecessarily complicated notation. The proof is a modification of the classical result in [15, Proposition 3.5].*

Combining Lemma 16 and Lemma 19, we have

$$\sum_{r=1}^{\infty} \beta^{r+1}\|x^{r+1}-x^r\|^2 < \infty, \quad \sum_{r=1}^{\infty} \|\lambda^{r+1}-\lambda^r\|^2 < \infty, \tag{4.88}$$

$$\sum_{r=1}^{\infty} (\beta^{r+1})^2\|(x^{r+1}-x^r)-(x^r-x^{r-1})\|^2 < \infty. \tag{4.89}$$

From (4.88) we have $\lambda^{r+1}-\lambda^r \to 0$, *which implies that From (4.89), we have*

$$(\rho^{r+1})(Ax^{r+1}-b)-\tau\lambda^r \to 0. \tag{4.90}$$

Combined with the fact that λ^r *is bounded, and* $\rho^{r+1} \to \infty$, *we conclude*

$$Ax^{r+1}-b \to 0. \tag{4.91}$$

Let (x^*, λ^*) *be a limit point of* (x^{r+1}, λ^{r+1}). *Comparing the optimality condition of the problem (4.1) and the optimality condition of* x*-subproblem (4.59a), in order to argue convergence to stationary solutions, we need to show*

$$\beta^{r+1}\|x^{r+1}-x^r\| \to 0. \tag{4.92}$$

Next we show such a claim. To proceed, let us define

$$v^{r+1} := \beta^{r+1}(x^{r+1}-x^r). \tag{4.93}$$

From (4.89), it is easy to show that

$$\|v^{r+1}-v^r\| = \|\beta^{r+1}(x^{r+1}-x^r)-\beta^r(x^r-x^{r-1})\| \to 0. \tag{4.94}$$

From the first inequality in (4.88), we have

$$\sum_{r=1}^{\infty} \frac{1}{\beta^{r+1}}\|v^{r+1}\|^2 \to 0. \tag{4.95}$$

This relation combined with Assumption [B3] implies: $\liminf \|v^{r+1}\| = 0$.

Let us pass a subsequence $\mathcal{K}$ *to* (x^r, λ^r) *and denote* (x^*, λ^*) *as its limit point. For notational simplicity, in the following the index set* $\{r\}$ *all belongs to the set* $\mathcal{K}$. *We already know from the*

previous argument that $\liminf_{r\to\infty} \|v^{r+1}\| = 0$. *Then it is clear that* $\lim_{r\to\infty} \|v^{r+1}\| = 0$ *if and only the following condition is true*

$$\lim_{r\to\infty} \|v^{r+1} - v^{r+t}\| = 0, \quad \forall\ t > 0. \tag{4.96}$$

Let us construct a new sequence

$$z^{r+1} = A^T\lambda^{r+1} + v^{r+1}. \tag{4.97}$$

Clearly $\liminf_{r\to\infty} z^{r+1} = A^T\lambda^*$, *because along the subsequence* λ^r *converges to* λ^*. *It is also easy to show that* (4.96) *is true if and only if the following is true*

$$\lim_{r\to\infty} \|z^{r+1} - z^{r+t}\| = 0, \quad \forall\ t > 0. \tag{4.98}$$

Suppose that (4.98) *is not true. Hence there exists an* $\epsilon > 0$ *such that* $\|z^r\| < \|A^T\lambda^*\| + \epsilon/2$ *for infinitely many* r, *and* $\|z^{r+1}\| > \|A^T\lambda^*\| + \epsilon/2$ *for infinitely many* r. *Then there exists an infinite subset of iteration indices* $\mathcal{R}$ *such that for each* $r \in \mathcal{R}$, *there exits a* $t(r)$ *such that*

$$\begin{aligned} &\|z^r\| < \|A^T\lambda^*\| + \epsilon/2, \quad \|z^{t(r)}\| > \|A^T\lambda^*\| + \epsilon, \\ &\|A^T\lambda^*\| + \epsilon/2 < \|z^t\| \le \|A^T\lambda^*\| + \epsilon, \quad \forall\ r < t < t(r). \end{aligned} \tag{4.99}$$

Also from the fact that $\|v^{r+1} - v^r\| \to 0$ *and* $\|\lambda^{r+1} - \lambda^r\| \to 0$, *we can conclude that* $\|z^{r+1} - z^r\| \to 0$. *Therefore, we must have*

$$\|z^r\| \ge \frac{3\epsilon}{8} + \|A^T\lambda^*\|. \tag{4.100}$$

Let r *be large enough such that*

$$\left|\|A^T\lambda^*\| - \|A^T\lambda^r\|\right| \le \|A^T(\lambda^* - \lambda^r)\| \le \frac{\epsilon}{4}. \tag{4.101}$$

Then we have

$$\|v^t\| \le \|A^T\lambda^t\| + \|A^T\lambda^*\| + \epsilon \le 2(\|A^T\lambda^*\| + \epsilon),\ \forall\ r < t < t(r), \tag{4.102a}$$

$$\|v^t\| \ge \|z^t\| - \|A^T\lambda^t\| \overset{(4.101)}{\ge} \|z^t\| - \|A^T\lambda^*\| - \frac{\epsilon}{4} \overset{(4.99)}{\ge} \frac{\epsilon}{4},\ \forall\ r < t < t(r). \tag{4.102b}$$

$$\|v^r\| \ge \|z^r\| - \|A^T\lambda^r\| \ge \|z^r\| - \|A^T\lambda^*\| - \frac{\epsilon}{4} \overset{(4.100)}{\ge} \frac{\epsilon}{8}. \tag{4.102c}$$

From the definition of $t(r)$ *we have that for all* $r \in \mathcal{R}$ *the following is true*

$$\frac{\epsilon}{2} \leq \|z^{t(r)}\| - \|z^r\| \leq \sum_{t=r}^{t(r)-1} \|z^{t+1} - z^t\|. \tag{4.103}$$

Next, we make the following simplification that $X \equiv \mathbb{R}$ *and* $h \equiv 0$ *to avoid lengthy discussion. The subsequent proof holds true for the general case as well, using the same techniques presented in [119, Theorem 4]. From the optimality condition* (4.63)*, and with the above simplification, we obtain*

$$z^{t+1} - z^t = \nabla f(x^t) - \nabla f(x^{t-1}), \tag{4.104}$$

which implies that

$$\|z^{t+1}\| - \|z^t\| \leq L\|x^t - x^{t-1}\| = \frac{L}{\beta^t}\|v^t\|. \tag{4.105}$$

Combining this result with (4.103)*, we obtain*

$$\frac{\epsilon}{2} < L \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}\|v^t\| \overset{(4.102a)}{\leq} 2L(\|A^T\lambda^*\| + \epsilon) \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}. \tag{4.106}$$

Which implies that

$$\frac{\epsilon}{4L(\|A^T\lambda^*\| + \epsilon)} \leq \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}. \tag{4.107}$$

Using the descent of the potential function (4.79) *we have, for* $r \in \mathcal{R}$ *and* r *large enough*

$$\begin{aligned} P_c^{t(r)} - P_c^r &\leq -\sum_{t=r}^{t(r)-1} \frac{C_5}{\beta^{t+1}}\|v^{t+1}\|^2 + \sum_{t=r}^{t(r)-1} C_3(\gamma^{t+1})^2\|\lambda^{t+1}\|^2 \\ &\leq -\frac{C_5}{L(\|A^T\lambda^*\| + \epsilon)}\frac{\epsilon^2}{64} \end{aligned} \tag{4.108}$$

where the last inequality we have used the fact that

$$\lim_{R_0 \to \infty} \sum_{r=R^0}^{\infty} C_3(\gamma^{t+1})^2\|\lambda^{t+1}\|^2 \to 0,$$

and equations (4.102b) *and* (4.107)*. This means that the potential function goes to* $-\infty$*, a contradiction. Therefore we conclude that*

$$\lim_{r \to \infty} \|z^{r+1} - z^{r+t}\| = 0, \quad \forall\ t > 0. \tag{4.109}$$

which further implies that

$$\lim_{r\to\infty} \|v^{r+1} - v^{r+t}\| = 0, \quad \forall\ t > 0. \tag{4.110}$$

Combined with the fact that $\liminf \|v^{r+1}\| = 0$*, we conclude that*

$$\lim_{r\to\infty} \|v^{r+1}\| = 0. \tag{4.111}$$

We conclude that every limit point of the sequence is a KKT point. **Q.E.D.**

4.4 Numerical Results

In this section, we customize the proposed algorithms to a number of applications in Section 4.1.1, and compare with the State-of-the-art algorithms.

4.4.1 Distributed Nonconvex Quadratic Problem

In this subsection we consider the nonconvex ℓ_1 penalized, nonnegative, sparse principal component analysis (SPCA) problem [8]. Distributed version of this problem [which is a special case of problem (4.1)] can be modeled as below

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{N} -x_i^\top \Sigma_i x_i + \alpha\|x_i\|_1 \\ \text{s.t.} \quad & \|x_i\|^2 \le 1, \quad x_i \ge 0, \quad i = 1, \cdots, N \\ & Ax = 0; \quad \texttt{Consensus Constraint} \end{aligned} \tag{4.112}$$

where $x_i \in \mathbb{R}^r$ for each i; $x := \{x_i\}_{i=1}^N$ stacks all x_i's, $\Sigma_i \in \mathbb{R}^{d\times d}$ is the covariance matrix for the mini-batch data in node i; $\alpha > 0$ is a constant that controls the sparsity. Let us define $\bar{x} := \frac{1}{N}\sum_{i=1}^N x_i$, $h(\bar{x}) := \alpha\|\bar{x}\|_1$, $f(\bar{x}) := \sum_{i=1}^N \bar{x}^T\Sigma_i\bar{x}$, and $X := \{x_i \mid \|\bar{x}\|^2 \le 1, \bar{x} \ge 0\}$. The stationary gap and the constraint violation for this problem is defined as below

$$\text{staionary-gap} = \left\| \bar{x} - \text{prox}_{h+\iota_X}\left[\bar{x} - \nabla f(\bar{x})\right] \right\|^2, \quad \text{con-vio} = \|Ax\|^2. \tag{4.113}$$

At this point, one can certainly use Algorithm 1 or Algorithm 2 to solve problem (4.112). However, the resulting x- subproblems for both algorithms are difficult to solve due to the fact that computing

the proximity operator for nonsmooth function $\alpha\|x\|_1 + \iota_{\|x\|^2\leq 1}(x) + \iota_{x\geq 0}(x)$ does not have a closed form (where $\iota_X(x)$ represents the indicator function for convex set X). On the contrary, the proximity operators for the individual component functions all have closed-form. To utilize such a problem structure, we divide the agents into three subsets, each with a distinctive regularizer. Let us denote $r = \lfloor N/3 \rfloor$. The new reformulation is given below

$$\begin{aligned} \min \quad & \sum_{i=1}^{r} \big(- x_i^\top \Sigma_i x_i + \frac{N\alpha}{r}\|x_i\|_1 \big) - \sum_{i=r+1}^{2r} x_i^\top \Sigma_i x_i - \sum_{i=2r+1}^{N} x_i^\top \Sigma_i x_i \qquad (4.114) \\ \text{s.t.} \quad & \|x_i\|^2 \leq 1, \quad i = r+1, \cdots, 2r \\ & x_i \geq 0, \qquad i = 2r+1, \cdots, N \\ & Ax = 0 \quad \texttt{Consensus Constraint}. \end{aligned}$$

To the best of our knowledge, no existing methods for nonconvex distributed optimization can effectively deal with the above problem (at least not with theoretical convergence guarantee to stationary solution). The major difficulty is to deal with the *agent-specific* nonsmooth terms. For comparison purpose, we consider the DSG algorithm [102], and the NEXT algorithm [92]. In our numerical result, the graph $\mathcal{G}$ is generated based on the scheme proposed in [141]. In this scheme a random graph with N nodes and radius R is generated with nodes uniformly distributed over a unit square, and two nodes connect to each other if their distance is less than R. The test problems are generated in the following manner. The number of agents, the network radius, the problem dimension, and the sparsity parameter to be $N = 20, R = 0.7, d = 10, \alpha = 0.01$, respectively. For PProx-PDA algorithm we set perturbation parameter $\gamma = 10^{-4}$, and ρ and β are picked such that they satisfy the theoretical bounds given in (4.51). For PProx-PDA-IA we set the increasing penalty $\rho = \beta = 40r$, and decreasing perturbation $\gamma = 10^{-3}/r$. For the DSG algorithm the stepsize is set $0.1/r$ (this choice is made so that DSG has the best performance). The parameters for NEXT are tuned according to the description in [92, Theorem 3]. Each algorithm is run for 20 independents trials, with random initialization and randomly generated data. The results are plotted in Fig. 4.1 and 4.2. In the figures, dashed lines with light colors are used to show the performance for each individual trial, while the solid dark lines are the average performance over all 20 trials. From the

plots it can be observed that the proposed algorithms, especially the increasing stepsize version, outperform both DSG and NEXT.

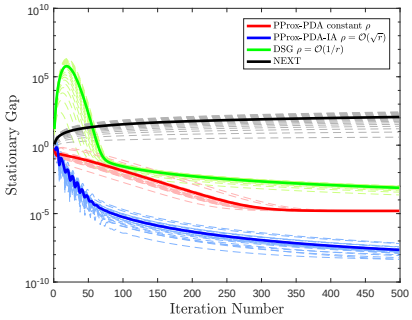


Figure 4.1: Comparison of proposed algorithms with DSG [102] and NEXT [92] in terms of stationary gap for problem 4.114 with parameters $N = 20, R = 0.7, d = 10, \alpha = 0.01$.

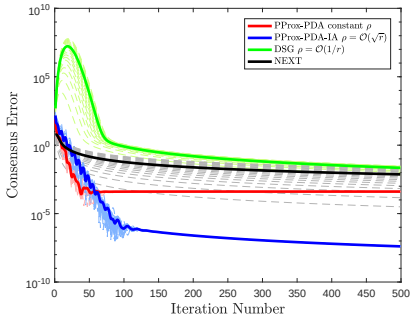


Figure 4.2: Comparison of proposed algorithms with DSG [102] and NEXT [92] in terms of constraint violation for problem 4.114 with parameters $N = 20, R = 0.7, d = 10, \alpha = 0.01$.

To see more numerical results we compare different algorithms with different problem setups. The algorithms are run for 20 independent trials with randomly generated data and random initial solutions in each individual trials. All algorithm parameters are set to be the same as in the previous experiment. The comparison results are displayed in Table 4.1. The first column describes the problem parameters including number of agents N, number of variables n, and the network radius R, while 'Alg1' and 'Alg2' stand for PProx-PDA and PProx-PDA-IA, respectively. It can be observed that in all scenarios the proposed algorithms outperform DSG.

Table 4.1: Comparison of proposed algorithms with DSG algorithm. Alg1 and Alg2 denote PProx-PDA and PProx-PDA-IA algorithms respectively.

	Stationary-Gap			Cons-Vio		
Parameters	Alg1	Alg2	DSG	Alg1	Alg2	DSG
$N = 5, n = 80, R = 0.7$	1.9E-4	6.0E-5	9.0E-4	6.0E-6	9.5E-7	4.3E-5
$N = 20, n = 15, R = 0.7$	1.3E-4	5.0E-8	9.4E-5	1.7E-3	6.8E-6	0.013
$N = 30, n = 20, R = 0.5$	6.3E-5	2.1E-8	2.6E-4	7.0E-3	6.4E-7	0.06
$N = 40, n = 30, R = 0.5$	2.0E-4	4.9E-8	1.5E-3	8.1E-3	1.5E-6	0.05

4.4.2 Nonconvex subspace estimation

In this subsection we study the problem of *sparse subspace estimation* (4.2). We compare the proposed PProx-PDA and PProx-PDA-IA with the ADMM algorithm proposed in [49, Algorithm

1]. Note that the latter is a heuristic algorithm that does not have convergence guarantee. We first consider a problem with the number of samples, problem dimension, and MCP parameters chosen as $n = 80$, $p = 128$, $\nu = 3$, $b = 3$, respectively. For PProx-PDA we set perturbation parameter $\gamma = 10^{-4}$, and ρ and β are chosen to satisfy the theoretical bounds given in (4.51). For PProx-PDA-IA we set increasing penalty $\rho = \beta = 5r$, and decreasing perturbation $\gamma = 10^{-4}/r$. The data set is generated following the same procedure as in [49]. In particular, we set $s = 5$ and $k = 1$, the leading eigenvalue of its covariance matrix Σ is set as $\nu_1 = 100$, and its corresponding eigenvector is sparse such that only the first $s = 5$ entries are nonzero, and they take the value $1/\sqrt{5}$. The rest of the eigenvalues are set to be 1, and their eigenvectors are chosen arbitrarily. For all three algorithms we measure the stationarity gap, the constraint violation.The result, which are from 20 independent trials with random initial solutions, are plotted in Fig. 4.3– 4.4. As shown in these figures, compared to the ADMM algorithm, the PProx-PDA-IA algorithm converges faster, and to better solutions.

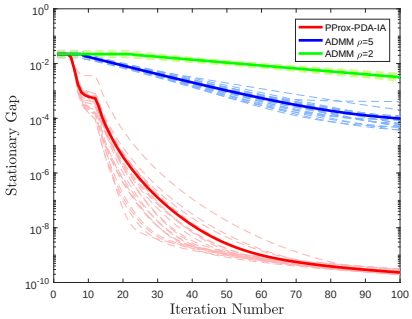


Figure 4.3: Comparison of proposed algorithms with ADMM in terms of stationary gap for nonconvex subspace estimation problem with MCP Regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively.

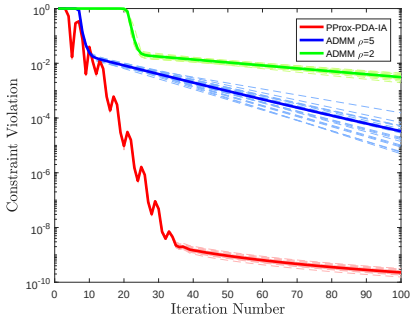


Figure 4.4: Comparison of proposed algorithms with ADMM in terms of constraint violation $\|Ax\|^2$ for nonconvex subspace estimation problem with MCP Regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively.

Our next experiment is designed to see the effect that the problem parameters (i.e. n, p, k, and s) have on the solution quality. Here, we compare the PProx-PDA-IA [with $\rho = \mathcal{O}(r), \gamma = \mathcal{O}(1/r)$] with ADMM algorithm with stepsize $\rho = 5$. Both algorithms will be run for 200 iterations. In this

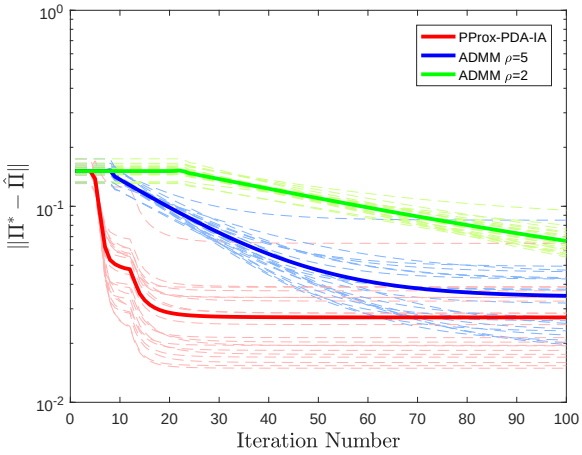


Figure 4.5: Comparison of proposed algorithms with ADMM in terms of Global Error for non-convex subspace estimation problem with MCP Regularization. The problem parameters are $n = 80$, $p = 128$, $\nu = 3$, $b = 3$. The solid lines and dotted lines represent the single performance and the average performance, respectively.

experiment we generate data sets with $s = 10$, $k = 5$, and vary other problem parameter. For this dataset the top five eigenvalues are set as $\lambda_1 = \cdots = \lambda_4 = 100$ and $\lambda_5 = 10$. To generate their corresponding eigenvectors we sample its nonzero entries from a standard Gaussian distribution, and then orthnormalize them while retaining the first $s = 10$ rows to be nonzero [49]. The rest of the eigenvalues are set as $\lambda_6 = \cdots = \lambda_p = 1$, and the associated eigenvectors are chosen arbitrarily. The results in terms of the error $\|\hat{\Pi} - \Pi^*\|$ are shown in Table 4.2. In all scenarios the proposed algorithm PProx-PDA-IA outperforms ADMM.

Further, the True Positive Rate (TPR) and False Positive Rate (FPR) [70] are measured and the results are displayed in Table 4.3 to see the recovery results. For this problem the event of being zero in vector $v = |\text{supp}(\text{diag}(\hat{\Pi}))|$ (here $\hat{\Pi}$ denotes the output of the algorithm) is considered as *positive event*. Let P denotes the number of positives, and S denotes the number of non-zeros in the ground truth vector denoted by Π^*. Further, let us use FP and TP to denote *false positive* and *true positive* respectively. In particular, FP counts the number of positive events (i.e. zeros in our case) in vector $\hat{\Pi}$ which are nonzero in ground truth vector Π^*. In contrast, TP counts the number of zeros in $\hat{\Pi}$ which are true zeros in Π^*. Given these notations, the FPR and TPR are

Table 4.2: Comparison of PProx-PDA-IA with ADMM in terms of Global Error $\|\hat{\Pi} - \Pi^*\|$ for nonconvex subspace estimation problem with MCP Regularization.

	$\|\hat{\Pi} - \Pi^*\|$	
Parameters	PProx-PDA-IA	ADMM
$n = 30, p = 128, k = 1, s = 5$	0.045 ± 0.02	0.052 ± 0.02
$n = 80, p = 128, k = 1, s = 5$	0.024 ± 0.01	0.028 ± 0.08
$n = 120, p = 128, k = 1, s = 5$	0.020 ± 0.07	0.021 ± 0.06
$n = 150, p = 200, k = 1, s = 5$	0.022 ± 0.07	0.022 ± 0.07
$n = 80, p = 128, k = 1, s = 10$	0.048 ± 0.01	0.062 ± 0.01
$n = 80, p = 128, k = 5, s = 10$	0.21 ± 0.05	0.29 ± 0.02
$n = 128, p = 128, k = 5, s = 10$	0.18 ± 0.02	0.25 ± 0.02
$n = 70, p = 128, k = 5, s = 10$	0.26 ± 0.03	0.33 ± 0.03

Table 4.3: Recovery results for PPox-PDA-IA and ADMM in terms of TPR and FPR.

	TPR		FPR	
Parameters	PProx-PDA-IA	ADMM	PProx-PDA-IA	ADMM
$n = 30, p = 128, k = 1, s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80, p = 128, k = 1, s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 120, p = 128, k = 1, s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 150, p = 200, k = 1, s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80, p = 128, k = 1, s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80, p = 128, k = 5, s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.53 ± 0.03	0.56 ± 0.04
$n = 128, p = 128, k = 5, s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.57 ± 0.01	0.59 ± 0.02
$n = 70, p = 128, k = 5, s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.53 ± 0.05	0.54 ± 0.01

defined as follows

$$FPR = \frac{FP}{S}, \quad TPR = \frac{TP}{P}. \tag{4.115}$$

In terms of TPR both algorithms work perfectly well. However, PProx-PDA-IA gets lower FPR compare to the ADMM algorithm.

4.4.3 Partial Consensus

The partial consensus optimization problem has been introduced in (4.8). As Stated in the introduction, we are not aware of any existing algorithm that is able to perform nonconvex partial consensus optimization with guaranteed performance. Let us consider *regularized logistic regression* problem [7] in a network with N nodes, in mini-bach setup i.e. each node stores b (batch size) data points, and each component function is given by

$$f_i(x_i) = \frac{1}{Nb}\left[\sum_{j=1}^{b}\log(1+\exp(-y_{ij}x_i^T v_{ij})) + \sum_{k=1}^{M}\frac{\hat{\beta}\hat{\alpha}x_{i,k}^2}{1+\hat{\alpha}x_{i,k}^2}\right],$$

where $v_{ij} \in \mathbb{R}^M$ and $y_{ij} \in \{1, -1\}$ are the feature vector and the label for the jth date point in i-th agent, $\hat{\alpha}$ and $\hat{\beta}$ are the regularization parameters [7].

We set $N = 20$, $M = 10$, $b = 100$, $\hat{\beta} = 0.01$, $\hat{\alpha} = 1$. The graph $\mathcal{G}$ is generated similar to the problem in subsection 4.4.1. The PProx-PDA and PProx-PDA-IA algorithms are implemented for the above problem. Both algorithms stop after 1000 iterations, and we measure the averaged performance over 20 trials, where in each trial the data matrix and the initial solutions are generated randomly independent. In Fig. 4.6 the stationary gap for the problem has been plotted. It can be observed that the gap is vanishing as the algorithm proceeds, and it appears that PProx-PDA-IA is faster than PProx-PDA. Fig. 4.7 displays the constraint violation for the PProx-PDA algorithm. It is also interesting to observe that when reducing the constraint violation error (represented by $\zeta > 0$), the resulting solution indeed achieves higher degrees of consensus.

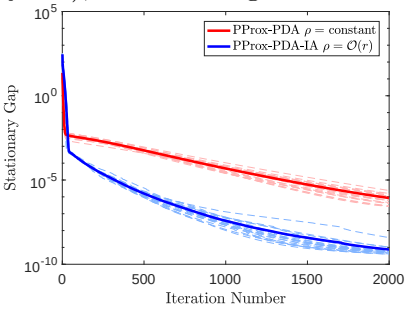


Figure 4.6: The stationary gap achieved by the proposed methods for the partial consensus problem. The solid lines and dotted lines represent the single performance and the average performance, respectively.

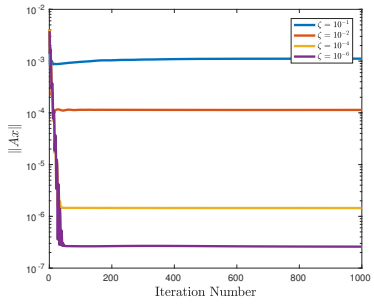


Figure 4.7: Constraint Violation $\|Ax\|$ achieved by the proposed methods for the partial consensus problem with different permissible tolerance ζ.

4.5 Conclusion

In this paper, we have proposed perturbed primal-dual based algorithms for optimizing nonconvex and linearly constrained problems. The proposed methods are of Uzawa type, in which a primal gradient descent step is performed followed by a (approximate) dual gradient ascent step. We have performed theoretical analysis for the convergence of the algorithm, and tested their performance

on a number of statistical and engineering applications. In the future, we plan to investigate, both in theory and in practice, whether the perturbation is necessary for primal-dual type algorithms to reach stationary solutions. Further, we plan to extend the proposed algorithms to problems with stochastic objective functions.

Acknowledgment. The authors would like to thanks Dr. Quanquan Gu who provided us with the codes to perform the numerical results in [49].

4.6 Appendix. Constraint qualification

In this section, we justify Assumption [B4], which imposes the boundedness of the dual variable. In particular, we discuss two situations in which the dual variables are guaranteed to be bounded. Throughout this section, we will assume that Assumption A and [B1]-[B3] hold true.

Case 1). In this case, we make use of some constraint qualification to argue the boundedness of the dual variables.

Assume that the so-called Robinson's condition is satisfied for problem (4.1) at $\hat{x}$ [114, Chap. 3]. This means the following holds

$$\left\{ \; Ad_x \;\middle|\; d_x \in \mathcal{T}_X(\hat{x}) \; \right\} = \mathbb{R}^M, \tag{4.116}$$

where d_x is the tangent direction for convex set X, and $\mathcal{T}_X(\hat{x})$ is the tangent cone to the feasible set X at the point $\hat{x}$. Utilizing this assumption we prove that the dual variable is bounded.

Lemma 20 *Suppose that the Robinson's condition holds true for problem (4.1). Then the sequence of dual variable λ^r generated by (4.59b) is bounded.*

Proof 13 *Let us argue by contradiction. Suppose that the dual variable is not bounded, i.e.,*

$$\|\lambda^r\| \to \infty. \tag{4.117}$$

Using Assumption [B3] we have the following identity (for large enough r)

$$\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2 = \frac{(\beta^{r+1})^2}{2c_0}\|x^{r+1}-x^r\|^2. \tag{4.118}$$

From Lemma 18, we have that [also cf. (4.88)]

$$\sum_{r=1}^{\infty} \|\lambda^{r+1} - \lambda^r\|^2 < \infty, \tag{4.119}$$

which implies that

$$\frac{1}{\rho^{r+1}}\|\lambda^{r+1}\|^2 - \frac{1}{\rho^r}\|\lambda^r\|^2 \to 0. \tag{4.120}$$

Plugging this result into (4.87), we conclude that the following inner product is lower bounded

$$\langle \lambda^{r+1} - \rho^{r+2}\gamma^{r+2}\lambda^{r+1}, Ax^{r+1} - b - \gamma^{r+2}\lambda^{r+1}\rangle,$$

and this further implies that $T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$ is lower bounded [by using the definition of T function in (4.23)]. By Lemma 19 (resp. Lemma 18), we conclude that the potential function is lower (resp. upper) bounded. Examine the definition of the potential function in (4.86) and use the choice of c in (4.85) we conclude that except $T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$, all the rest of the terms are all nonnegative. Using the lower boundedness of T, we conclude that the term $\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1} - x^r\|^2$ in the potential function is bounded. Therefore, there exists D_1 such that

$$\beta^{r+1}\|x^{r+1} - x^r\| \le D_1. \tag{4.121}$$

Note that all the results mentioned above do not assume the boundedness of the dual variable.

From the optimality condition of x^{r+1} we have for all $x \in X$

$$\langle \nabla f(x^r) + \xi^{r+1} + A^T\lambda^{r+1} + \beta^{r+1}B^TB(x^{r+1} - x^r), x - x^{r+1})\rangle \ge 0.$$

Note that $\beta^{r+1}\|x^{r+1} - x^r\|$ is a bounded sequence, so does $\beta^{r+1}B^TB(x^{r+1} - x^r)$. Suppose that $\{\lambda^r\}$ is not bounded, let us define a new bounded sequence as $\mu^r = \lambda^r/\|\lambda^r\|$. Let (x^*, μ^*) be a limit point of $\{x^{r+1}, \mu^{r+1}\}$. Assume that the Robinson's condition holds at x^*. Dividing both sides of the above inequality by $\|\lambda^{r+1}\|$ we obtain for all $x \in X$

$$\begin{aligned}\langle \nabla f(x^r)/\|\lambda^{r+1}\| + \xi^{r+1}/\|\lambda^{r+1}\| + A^T\mu^{r+1} \\ + \beta^{r+1}B^TB(x^{r+1} - x^r)/\|\lambda^{r+1}\|, x - x^{r+1}\rangle \ge 0.\end{aligned}$$

Passing limit, and utilizing the assumption that $\|\lambda^{r+1}\| \to \infty$, *and that* X *is a compact set, we obtain*

$$\langle A^T \mu^*, x - x^* \rangle \geq 0, \ \forall \ x \in X.$$

Utilizing the Robinson's condition, we know that there exists a scaling constant $c > 0$ *that such* $c\langle A, x - x^* \rangle = -\mu^*$. *Therefore we must have* $\mu^* = 0$. *However, this contradicts to the fact that* $\|\mu^*\| = 1$. *Therefore, we conclude that* $\{\lambda^r\}$ *is a bounded sequence.* **Q.E.D.**

Case 2). In this section, we verify Assumption [B4] by further imposing conditions on the constraint set and the nonsmooth terms.

Specifically we consider the following problem

$$\min_{\{x_k\}} \ f(x) + h(x) := f(x) + \sum_{k=1}^{K} h_k(x_k) \quad \text{s.t.} \quad \sum_{k=1}^{K} A_k x_k = b, \tag{4.122}$$

where h_k is a convex nonsmooth term that can include both regularizer and indicator functions for convex set X. Setting $K = 1$, the above problem is equivalent to the original problem (4.1).

Assumption C. Assume that for one of the block, say K satisfies the following:

$$X_K = \mathbb{R}^{n_K}, \quad \partial h_K \text{ has bounded domain.} \tag{4.123}$$

Note that the second of the above condition is possible for example when $h_K(x_K) = \|x_K\|_q$ for some constant $q \geq 1$. Further, we assume that the partial gradient of f with respect to x_K, denoted by $\nabla_K f(x)$, is bounded for all $x_K \in \text{dom}\,(h_K)$, and that A_K has full row rank.

Given the above assumption, the following lemma characterizes the bound for the dual variable.

Lemma 21 *Suppose that the Assumption C holds true. Then there exists constant* Λ *such that*

$$\|\lambda^{r+1}\|^2 \leq \Lambda, \quad \forall \ r \geq 0, \tag{4.124}$$

Proof 14 *First, from the optimality condition of* x*-update* (4.59a) *we have that for all* k, *and for all* $x_k \in dom(h_k)$

$$\langle \nabla_k f(x^r) + A_k^T \lambda^{r+1} + \beta^{r+1} B^T B(x_k^{r+1} - x_k^r) + \xi_k^{r+1}, x_k^{r+1} - x_k \rangle \leq 0, \tag{4.125}$$

where $\nabla_k f(x^r)$ *denotes the partial derivative of* $f(x)$ *with respect to the block variable* x_k *at* $x = x^r$*; and* $\xi_k^{r+1} \in \partial h_k(x^{r+1})$*. In particular for the block* K *because it is unconstrained, we have*

$$0 = \nabla_K f(x^r) + A_K^T \lambda^{r+1} + \xi_K^{r+1} + \beta^{r+1} B^T B(x_K^{r+1} - x_K^r). \tag{4.126}$$

Rearranging terms, we obtain

$$-A_K^T \lambda^{r+1} = \nabla_K f(x^r) + \xi_K^{r+1} + \beta^{r+1} B^T B(x_K^{r+1} - x_K^r). \tag{4.127}$$

From Assumption C we know that theres exists M_0 *such that* $\|\nabla_K f(x^r) + \xi_K^{r+1}\| \leq M_0$*. Together with the previous identity, we get*

$$\|A_K^T \lambda^{r+1}\|^2 \leq 2M_0^2 + 2(\beta^{r+1})^2 \|B^T B(x_K^{r+1} - x_K^r)\|^2 \quad \forall\ r. \tag{4.128}$$

Utilizing the fact that $\sigma_K^2 \|\lambda^{r+1}\|^2 \leq \|A_K^T \lambda^{r+1}\|^2$*, where* σ_K^2 *denoted the smallest nonzero eigenvalue of* $A_K^T A_K$*, we further have*

$$\|\lambda^{r+1}\|^2 \leq \frac{2}{\sigma_K^2} \left[(\beta^{r+1})^2 \|B^T B(x_K^{r+1} - x_K^r)\|^2 + D_0^2\right] \quad \forall\ r. \tag{4.129}$$

Here $\sigma_K > 0$ *because we have assumed that* A_K^T *is full column rank in Assumption C. Combining this with equation* (4.121) *one can find constant* Λ *such that* $\|\lambda^{r+1}\|^2 \leq \Lambda$*. The proof is complete.* **Q.E.D.**

Appendix B

In this section we show how the sufficient conditions developed in Appendix A can be applied to problems discussed in Section 4.1.1. Specifically, we will focus on the sparse subspace estimation problem (4.5) and the inexact consensus problem (4.8).

We first show that Assumption C is satisfied for sparse subspace estimation problem (4.5). Recall that for this problem we have two block variables (Π, Φ), and $h(\Phi) := \|\Phi\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\phi_{ij}|$.

The subdifferential the ℓ_1 function can be expressed below, and it is obviously is bounded

$$\partial|\phi_{ij}| = \begin{cases} 1 & \text{if } \phi_{ij} > 0; \\ [-1,1] & \text{if } \phi_{ij} = 0; \\ -1 & \text{if } \phi_{ij} < 0. \end{cases}$$

Then we show that $\nabla_\Phi f(\Pi,\Phi)$ is bounded where $f(\Pi,\Phi) = \left\langle \hat{\Sigma}, \Pi \right\rangle + q_\nu(\Phi)$, and $\nabla_\Phi f(\Pi,\Phi) = \nabla_\Phi q_\nu(\Phi)$. For the MCP regularization with parameter b, we have $q_\nu(\Phi) = \sum_{i=1}^p \sum_{j=1}^p q_\nu(\phi_{ij})$, where

$$q_\nu(\phi_{ij}) = \begin{cases} \frac{-\phi_{ij}^2}{2b} & \text{if } |\phi_{ij}| \le b\nu; \\ -\nu|\phi_{ij}| + \frac{b\nu^2}{2} & \text{if } |\phi_{ij}| > b\nu. \end{cases}$$

Also, for $q_\nu(\phi_{ij})$ one can simply check that

$$\frac{\partial q_\nu(\phi_{ij})}{\partial \phi_{ij}} = \begin{cases} \frac{-\phi_{ij}}{b} & \text{if } |\phi_{ij}| < b\nu; \\ -\nu \text{ sign}(\phi_{ij}) & \text{if } |\phi_{ij}| > b\nu. \end{cases}$$

This is obviously a bounded function. Finally the matrix $A_\Phi = -I$ is full row rank matrix. In summary, we have validated all the conditions in Assumption C.

Next we consider the partial consensus problem given in (4.8). To proceed, we note that the Robinson's condition reduces to the well-known Mangasarian-Fromovitz constraint qualification (MFCQ) if we set $X = \mathbb{R}^N$, and write out explicitly the inequality constraints as $g(x) \le 0$. [114, Lemma 3.17]. To State the MFCQ, consider the following optimization problem

$$\begin{aligned} \min_{y \in \mathbb{R}^N} \quad & f(y) \\ \text{s.t.} \quad & p(y) = 0; \\ & g(y) \le 0, \end{aligned} \tag{4.130}$$

where $f : \mathbb{R}^N \to \mathbb{R}$, $p : \mathbb{R}^N \to \mathbb{R}^M$ and $g : \mathbb{R}^N \to \mathbb{R}^P$ are all continuously differentiable functions. For given feasible solution $\hat{y}$ let us use $\mathcal{A}(\hat{y})$ to denote the indices for active inequality constraints,

that is

$$\mathcal{A}(\hat{y}) := \{1 \leq j \leq P \text{ s.t. } g_j(\hat{y}) = 0\}. \tag{4.131}$$

Then MFCQ holds for optimization problem (4.130) at point $\hat{y}$ if we have: 1) The rows of Jacobian matrix of $p(y)$ denoted by $\nabla p(\hat{y})$ are linearly independent. 2) There exists a vector $d_y \in \mathbb{R}^N$ such that

$$\nabla p(\hat{y}) d_y = 0, \nabla g_j(\hat{y})^T d_y < 0, \ \forall \ j \in \mathcal{A}(\hat{y}). \tag{4.132}$$

Below we show that MFCQ holds true for problem (4.8) at any point (x, z) that satisfies $z \in Z$. Comparing this problem with (4.130) we have the following specifications. The optimization variable $y = [x; z]$, where $x \in \mathbb{R}^N$ stacks all $x_i \in \mathbb{R}$ from N nodes (here we assume $x_i \in \mathbb{R}$ only for the ease of presentation). Also, $z \in \mathbb{R}^E$ stacks all $z_e \in \mathbb{R}$ for $e \in \mathcal{E}$. The equality constraint is written as $p(y) = [A, -I]y = 0$, where $A \in \mathbb{R}^{E \times N}$ and I is an $E \times E$ identity matrix. Finally, for the inequality constraint we have $g_e(y) = z_e^2 - \xi_e \leq 0$, and the active set is given by $\mathcal{A}(y) := \{e \mid z_e^2 - \xi_e = 0\}$.

To show that MFCQ holds, consider a solution $y^* := (x^*, z^*)$ such that $z^* \in Z$. First observe that the Jacobian of equality constraint is $\nabla p(y^*) = [A, -I]$ which has full row rank. In order to verify the second condition we need to find a vector $d_y := [d_x; d_z] \in \mathbb{R}^{N+E}$ such that

$$Ad_x = d_z, \tag{4.133a}$$

$$z_e [d_z]_e < 0 \quad \text{for } e \in \mathcal{A}(y^*), \tag{4.133b}$$

where $[d_z]_e$ denotes the eth component of vector d_z. To proceed, let us take $(d_x, d_z) = -(x^*, z^*)$. Clearly $Ad_x = d_z$ still holds true. Further, suppose the eth constraint is active, i.e., $|z_e^*| = \sqrt{\xi_e}$, then clearly $\left|[d_z]_e\right| = \sqrt{\xi_e}$. It follows that $[d_z]_e \times z_e^* = -\xi_e < 0$ for all $e \in \mathcal{A}(y^*)$. Therefore condition (4.133b) is also satisfied.

CHAPTER 5. ZEROTH ORDER NONCONVEX MULTI-AGENT OPTIMIZATION

Abstract

In this paper we consider distributed optimization problems over a multi-agent network, where each agent can only partially evaluate the objective function, and it is allowed to exchange messages with its immediate neighbors. Differently from all existing works on distributed optimization, our focus is given to optimizing a class of difficult *non-convex* problems, and under the challenging setting where each agent can only access the *zeroth-order information* (i.e., the functional values) of its local functions. For different types of network topologies such as undirected connected networks or star networks, we develop efficient distributed algorithms and rigorously analyze their convergence and rate of convergence (to the set of stationary solutions). Numerical results are provided to demonstrate the efficiency of the proposed algorithms.

5.1 INTRODUCTION

Distributed optimization and control has found wide range of applications in emerging research areas such as data-intensive optimization [65, 146], signal and information processing [47, 117], multi-agent network resource allocation [134], communication networks [83], just to name a few. Typically this type of problems is expressed as minimizing the sum of additively separable cost functions, given below

$$\min_{x\in\mathbb{R}^M} \; g(x) := \sum_{i=1}^{N} f_i(x), \tag{5.1}$$

where N denotes the number of agents in the network; $f_i : \mathbb{R}^M \to \mathbb{R}$ represents some (possibly nonsmooth and nonconvex) cost function related to the agent i. It is usually assumed that each agent i has complete information on f_i, and they can only communicate with their neighbors.

Therefore the key objectives of the individual agents are: 1) to achieve consensus with its neighbors about the optimization variable; 2) to optimize the global objective function $g(x)$.

Extensive research has been done on consensus based distributed optimization, but these works are mostly restricted to the family of *convex* problems where $f_i(x)$'s are all convex functions. In [100] a first-order method based on the average consensus termed decentralized subgradient (DSG) has been proposed. Following this work, many other first-order algorithms have been proposed to solve distributed convex optimization problems under different assumptions on the underlying problem. For example in [100] DSG is extended to the case where quantized information are used. In [129] a local constraint set is added to each local optimization problem. A dual averaging subgradient method is developed and analyzed in [34]. In [99] an algorithm termed subgradient-push has been developed for a time-varying directed network. Other related algorithms can be found in [88, 89, 123, 69, 10]. The first-order methods presented so far only converge to a neighborhood of solution set unless using diminishing stepsizes, however using diminishing stepsizes often makes the convergence slow. In order to overcome such a difficulty, recently the authors of [51] and [122] have proposed two methods, named incremental aggregated gradient (IAG) and exact first-order algorithm (EXTRA), both of which are capable of achieving fast convergence using constant stepsizes. Another class of algorithms for solving problem (5.1) in the convex cases are designed based on primal-dual methods, such as the Alternating Direction Method of Multipliers (ADMM) [124, 138, 64, 68, 12], many of its variants [60, 25, 98], and distributed dual decomposition method [1].

Despite the fact that distributed optimization in convex setting has a broad applicability, many important applications are inherently nonconvex. For example, the resource allocation in ad-hoc network [134], flow control in communication networks [130], and distributed matrix factorization [59, 54], just to name a few. Unfortunately, without the key assumption of the convexity of f_i's, the existing algorithms and analysis for convex problems are no longer applicable. Recently a few works have started to consider algorithms for nonconvex distributed optimization problems. For example, in [148] an algorithm based on dual subgradient method has been proposed, but it

relaxes the exact consensus constraint. In [17] a distributed stochastic projection algorithm has been proposed, and the algorithm converges to KKT solutions when certain diminishing stepsizes are used. The authors of [63] proposed an ADMM based algorithm, and they provided one of the first global convergence rate analysis for distributed nonconvex optimization. More recently, a new convexification-decomposition based approach named NEXT has been proposed in [92, 130], which utilizes the technique of *gradient tracking* to effectively propagate the information about the local functions over the network. In [59, 55, 57, 61], a number of primal-dual based algorithms with global convergence rate guarantee have been designed for different network structures.

A key drawback for all the above mentioned algorithms, convex or nonconvex, is that they require at least first-order gradient information, and sometime even the second or higher order information, in order to guarantee global convergence. Unfortunately, in many real-world problems, obtaining such information can be very expensive, if not impossible. For example, in simulation-based optimization [126], the objective function of the problem under consideration can only be evaluated using repeated simulation. In certain scenarios of training deep neural network [76], the relationship between the decision variables and the objective function is too complicated to derive explicit form of the gradient. Further, in bandit optimization [2, 37], a player tries to minimize a sequence of loss functions generated by an adversary, and such loss function can only be observed at those points in which it is realized. In these scenarios, one has to utilize techniques from derivative-free optimization, or optimization using zeroth-order information [127, 27]. Accurately estimating a gradient often requires extensive simulation (see [41]). In certain application domains, the complexity of each simulation may require significant computational time (e.g. hours). Even when such simulations are parallelized approaches based upon a centralized gradient estimation are impractical due to the need to synchronize. In contrast, a zeroth-order distributed approach has limited simulation requirements for each node and does not require synchronization.

Recently, Nesterov [105] has proposed a general framework of zeroth-order gradient based algorithms, for both convex and nonconvex problems. It has been shown that for convex (resp. nonconvex) smooth problems the proposed algorithms require $\mathcal{O}(\frac{M}{\epsilon^2})$ iterations (M denotes the

Figure 5.1: Left: Mesh Network (MNet); Right: Star Network (SNet)

dimension of the problem) to achieve an ϵ-optimal (resp. ϵ-stationary i.e. $\|\nabla f(x)\|^2 \leq \epsilon$) solution. Further, for both convex and nonconvex problems, the convergence rate for zeroth-order gradient based-algorithms is at most $\mathcal{O}(M)$ times worse than that of the first-order gradient-based algorithms. Ghadimi and Lan [45] developed a stochastic zeroth-order gradient method which works for convex and nonconvex optimization problems. Duchi et al. [33] proposed a stochastic zeroth-order Mirror Descent based algorithm for solving stochastic convex optimization problems. In [43] a zeroth-order ADMM algorithm has been proposed for solving convex optimization problems. The complexity of $\mathcal{O}(\frac{1}{\sqrt{T}})$ has been proved for the proposed algorithm, where T denotes the total number of iterations. Recently an asynchronous stochastic zeroth-order gradient descent (ASZD) algorithm is proposed in [81] for solving stochastic nonconvex optimization problem. Following this work, a variance reduced version of ASZD denoted by AsyDSZOVR is proposed in [67] for solving the same problem to improve the convergence rate from $\mathcal{O}(\frac{1}{\sqrt{T}})$ in ASZD to $\mathcal{O}(\frac{1}{T})$ in AsyDSZOVR. However, the zeroth-order based methods reviewed above are all centralized algorithms, hence they cannot be implemented in a distributed setting.

In this work we are interested in developing algorithms for the challenging problem of nonconvex distributed optimization, under the setting where each agent i can only access the *zeroth-order information* of its local functions f_i. For two different types of network topologies, namely, the undirected mesh network (MNet) (cf. Fig. 5.1) and the star networks (SNet) (cf. Fig. 5.1), we develop efficient distributed algorithms and rigorously analyze their convergence and rate of convergence (to the set of stationary solutions).

In particular, the MNet refers to a network whose nodes are connected to a subset of nodes through an undirected link, and such a network is very popular in applications such as distributed machine learning [40, 96], and distributed signal processing [46, 117]. On the other hand, the SNet has a central controller, which is connected to all the rest of the nodes. Such a network is popular in parallel computing; see for example [145, 80, 55]. The main contributions of our work is given below.

- For MNet, we design an algorithm capable of dealing with nonconvexity and zeroth-order information in the distributed setting. The proposed algorithm is based upon a primal-dual based zeroth-order scheme, which is shown to converge to the set of stationary solutions of problem (5.1) (with nonconvex but smooth f_i's), in a globally sublinear manner.

- For SNet we propose a stochastic primal-dual based method, which is able to further utilize the special structure of the network (i.e., the presence of the central controller) and deal with problem (5.1) with nonsmooth objective. Theoretically, we show that the proposed algorithm also converges to the set of stationary solutions in a globally sublinearly manner.

To the best of our knowledge, these algorithms are the first ones for distributed nonconvex optimization that are capable of utilizing zeroth-order information, while possessing global convergence rate guarantees.

Notation. We use $\|\cdot\|$ to denote the Euclidean norm, and use $\|\cdot\|_F$ to denote the Frobenius norm. If A is a matrix, $A^\top$ represent its transpose. For a given vector a and matrix H, we define $\|a\|_H^2 := a^T H a$. The notation $\langle a, b\rangle$ is used to denote the inner product of two vectors a, b. To denote an $M \times M$ identity matrix we use I_M. $\mathbb{E}[\cdot]$ denotes taking expectation with respect to all random variables, and $\mathbb{E}_v[\cdot]$ denote taking expectation with respect to the random variable v.

Preliminaries. We present some basic concepts and key properties related to derivative-free optimization [105]. Suppose $\mu > 0$ is the so-called smoothing parameter, then for a standard Gaussian random vector $\phi \in \mathbb{R}^Q$ the smoothed version of function f is defined as follows

$$f_\mu(z) = \mathbb{E}_\phi[f(z + \mu\phi)] = \frac{1}{(2\pi)^{\frac{Q}{2}}} \int f(z + \mu\phi) e^{-\frac{1}{2}\|\phi\|^2} \, d\phi.$$

Let us assume that $f : \mathbb{R}^Q \to \mathbb{R}$ is $\hat{L}$-smooth (denoted as $f \in \mathcal{C}^1_{\hat{L}}$), i.e. there exists a constant $\hat{L} > 0$ such that

$$\|\nabla f(z_1) - \nabla f(z_2)\| \leq \hat{L}\|z_1 - z_2\|, \ \forall \ z_1, z_2 \in \mathrm{dom}(f). \tag{5.2}$$

Then it can be shown that the function $f_\mu \in \mathcal{C}^1_{L_\mu}$ for some $L_\mu \leq \hat{L}$, and its gradient is given by

$$\nabla f_\mu(z) = \frac{1}{(2\pi)^{\frac{Q}{2}}} \int \frac{f(z + \mu\phi) - f(z)}{\mu} \phi e^{-\frac{1}{2}\|\phi\|^2} \, d\phi. \tag{5.3}$$

Further, for any $z \in \mathbb{R}^Q$, it is proved in [105, Theorem 1, Lemma 3] that

$$|f_\mu(z) - f(z)| \leq \frac{\mu^2}{2}\hat{L}Q, \tag{5.4}$$

$$\|\nabla f_\mu(z) - \nabla f(z)\| \leq \frac{\mu}{2}\hat{L}(Q+3)^{\frac{3}{2}}, \ \forall z \in \mathrm{dom}\,(f). \tag{5.5}$$

A stochastic zeroth-order oracle ($\mathcal{SZO}$) takes $z \in \mathrm{dom}\,(f)$ and returns a noisy functional value of $f(z)$, denoted by $\mathcal{H}(z;\xi)$, where $\xi \in \mathbb{R}$ is a random variable characterizing the stochasticity of $\mathcal{H}$. We make the following assumption regarding $\mathcal{H}(z;\xi)$ and $\nabla f(z)$.

Assumption A. We assume the following

A1. $\mathrm{Dom}(f)$ is an open set, and there exists $K \geq 0$ such that $\forall\ z \in \mathrm{dom}\,(f)$, we have: $\|\nabla f(z)\| \leq K$;

A2. For all $z \in \mathrm{dom}\,(f)$, $\mathbb{E}_\xi[\mathcal{H}(z,\xi)] = f(z)$;

A3. For some $\sigma \geq 0$, $\mathbb{E}[\|\nabla\mathcal{H}(z,\xi) - \nabla f(z)\|^2] \leq \sigma^2$, where $\nabla\mathcal{H}(z,\xi)$ denotes any stochastic estimator for $\nabla f(z)$.

These assumptions are standard in zeroth-order optimization. See for example [43, Def 1.3 and Lemma 4.2], [105, Eq. 4], and [45, A3]. Utilizing the $\mathcal{SZO}$ to obtain the functional values, one can show that the following quantity is an unbiased estimator for $\nabla f_\mu(z)$

$$G_\mu(z,\phi,\xi) = \frac{\mathcal{H}(z+\mu\phi,\xi) - \mathcal{H}(z,\xi)}{\mu}\phi, \tag{5.6}$$

where the constant $\mu > 0$ is smoothing parameter; $\phi \in \mathbb{R}^Q$ is a standard Gaussian random vector. In particular, we have

$$\mathbb{E}_{\xi,\phi}[G_\mu(z,\phi,\xi)] = \mathbb{E}_\phi\Big[\mathbb{E}_\xi[G_\mu(z,\phi,\xi) \mid \phi]\Big] = \nabla f_\mu(z). \tag{5.7}$$

Furthermore, for given J independent samples of $\{(\phi_j,\xi_j)\}_{j=1}^J$, we define $\bar{G}_\mu(z,\xi,\phi)$ as the sample average:

$$\bar{G}_\mu(z,\xi,\phi) := \frac{1}{J}\sum_{j=1}^{J} G_\mu(z,\phi_j,\xi_j), \tag{5.8}$$

where $\phi := \{\phi_j\}_{j=1}^J$, $\xi := \{\xi_j\}_{j=1}^J$. It is easy to see that for any $J \geq 1$, $\bar{G}_\mu(z,\xi,\phi)$ is an unbiased estimator of $\nabla f_\mu(z)$. Utilizing the above notations and definitions we have the following lemma regarding the $\bar{G}_\mu(z,\xi,\phi)$.

Lemma 22 *[43, Lemma 4.2] Suppose that Assumption A holds true. Then we have the following*

$$\mathbb{E}_{\xi,\phi}[\|\bar{G}_\mu(z,\xi,\phi) - \nabla f_\mu(z)\|^2] \leq \frac{\tilde{\sigma}^2}{J}, \tag{5.9}$$

where $\tilde{\sigma} := 2Q[K^2 + \sigma^2 + \mu^2\hat{L}^2Q]$.

5.2 Zeroth-Order Algorithm over MNet

5.2.1 System Model

Consider a network of agents represented by a graph $\mathcal{G} := \{\mathcal{V},\mathcal{E}\}$, with $|\mathcal{V}| = N$ (N nodes) and $|\mathcal{E}| = E$ (E edges). Each node $v \in \mathcal{V}$ represents an agent in the network, and each edge $e_{ij} = (i,j) \in \mathcal{E}$ indicates that node i and j are neighbors. Let $\mathcal{N}_i := \{j \mid (i,j) \in \mathcal{E}\}$ denotes the set of neighbors of agent i, and assume that $|\mathcal{N}_i| = d_i$. We assumed that each node can only communicate with its d_i single-hop neighbors in $\mathcal{N}_i$.

We consider the following reformulation of problem (5.1)

$$\min_{z_i \in \mathbb{R}^M} \sum_{i=1}^{N} f_i(z_i), \quad \text{s.t. } z_i = z_j,\ \forall\ e_{ij} \in \mathcal{E}, \tag{5.10}$$

where for each agent $i = 1, \cdots N$ we introduce a local variable $z_i \in \mathbb{R}^M$. If the graph $\mathcal{G}$ is a connected graph, then problem (5.10) is equivalent to problem (5.1). For simplicity of presentation let us set $Q := NM$, and define a new variable $z := \{z_i\}_{i=1}^N \in \mathbb{R}^{Q\times 1}$. Throughout this section, we will assume that each function $f_i : \mathbb{R}^M \to \mathbb{R}$ is a nonconvex and smooth function. Below we present a few network related quantities to be used shortly.

- *The incidence matrix:* For a given graph $\mathcal{G}$, the *incidence matrix* $\tilde{A} \in \mathbb{R}^{E\times N}$ is a matrix where for each edge $k = (i, j) \in \mathcal{E}$ and when $j > i$, we set $\tilde{A}(k, i) = 1$ and $\tilde{A}(k, j) = -1$. The rest of the entries of $\tilde{A}$ are all zero. For example, for the network in Fig. 5.1 the edge set is $\mathcal{E} = \{e_{12}, e_{14}, e_{34}\}$, therefore the incidence matrix is given by

$$\tilde{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

  Define the *extended incidence matrix* as

$$A := \tilde{A} \otimes I_M \in \mathbb{R}^{EM\times Q}. \tag{5.11}$$

- *The degree matrix*: For a given graph $\mathcal{G}$, the *degree matrix* $\tilde{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with $\tilde{D}(i, i) = d_i$ where d_i is the degree of node i; let $D := \tilde{D} \otimes I_M \in \mathbb{R}^{Q\times Q}$.

- *The signed/signless Laplacian matrix*: For a given graph $\mathcal{G}$ with its extended incidence matrix given by (5.11), its signed and signless Laplacian matrices are expressed as

$$L^- := A^\top A \in \mathbb{R}^{Q\times Q}. \tag{5.12a}$$

$$L^+ := 2D - A^\top A \in \mathbb{R}^{Q\times Q}. \tag{5.12b}$$

Using the above notations, one can easily check that problem (5.10) can be written compactly as below

$$\min_{z\in\mathbb{R}^Q} \; g(z) = \sum_{i=1}^N f_i(z_i), \quad \text{s.t. } Az = 0, \tag{5.13}$$

where we have defined $z := \{z_i\}_{i=1}^N \in \mathbb{R}^{Q\times 1}$. The Lagrangian function for this problem is defined by

$$L(z,\lambda) := g(z) + \langle \lambda, Az\rangle, \tag{5.14}$$

where $\lambda \in \mathbb{R}^{EM\times 1}$ is the dual variable associated with the constraint $Az = 0$. The stationary solution set for the problem (5.13) is given by

$$S = \{(z^*,\lambda^*) \mid \nabla_z L(z^*,\lambda^*) = 0 \text{ and } Az^* = 0\}, \tag{5.15}$$

where $\nabla_z L(z^*,\lambda^*)$ denotes the gradient of Lagrangian function with respect to the variable z evaluated at (z^*,λ^*).

5.2.2 The Proposed Algorithm

In this subsection we present a Zeroth-Order NonconvEx, over MNet (ZONE-M) algorithm which is capable of solving distributed nonconvex optimization problem in an efficient manner [to the set of stationary solutions as defined in (5.15)]. To proceed, let us first construct the augmented Lagrangian (AL) function for problem (5.13)

$$L_\rho(z,\lambda) := g(z) + \langle \lambda, Az\rangle + \frac{\rho}{2}\|Az\|^2, \tag{5.16}$$

where $\lambda \in \mathbb{R}^{EM\times 1}$ is the dual variable associated with the constraint $Az = 0$, and $\rho > 0$ denotes the penalty parameter. To update the primal variable z, the AL is first approximated using a quadratic function with a degree-matrix weighted proximal term $\|z - z^r\|_D^2$, followed by one step of zeroth-order gradient update to optimize such a quadratic approximation. After the primal update, an approximated dual ascent step is performed to update λ. The algorithm steps are detailed in Algorithm 7. Note that the ZONE-M is a variant of the popular method called *Method of Multipliers* (MM), whose steps are expressed below [58]

$$z^{r+1} = \underset{z\in\mathbb{R}^Q}{\operatorname{argmin}}\, L_\rho(z,\lambda^r), \tag{5.17}$$

$$\lambda^{r+1} = \lambda^r + \rho A z^{r+1}. \tag{5.18}$$

Algorithm 7 The ZONE-M Algorithm

1: **Input:** $z^0 \in \mathbb{R}^Q$, $\lambda^0 \in \mathbb{R}^{EM}$, $D \in \mathbb{R}^{Q\times Q}$, $A \in \mathbb{R}^{EM\times Q}$, $T \geq 1$, $J \geq 1$, $\mu > 0$

2: **for** $r = 0$ **to** $T-1$ **do**

For each $i = 1, \cdots, N$, generate $\phi_{i,j}^r \in \mathbb{R}^M$, $j = 1, 2, \cdots, J$ from an i.i.d standard Gaussian distribution and calculate $\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r) \in \mathbb{R}^M$ by

$$\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r) = \frac{1}{J}\sum_{j=1}^{J} \frac{\mathcal{H}_i(z_i^r + \mu\phi_{i,j}^r, \xi_{i,j}^r) - \mathcal{H}_i(z_i^r, \xi_{i,j}^r)}{\mu}\phi_{i,j}^r, \tag{5.19}$$

where we have defined $\phi_i^r := \{\phi_{i,j}^r\}_{j=1}^J$, $\xi_i^r := \{\xi_{i,j}^r\}_{j=1}^J$; Set

$$G_{\mu,r}^J := \{\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r)\}_{i=1}^N \in \mathbb{R}^Q$$

Update z and λ by

$$z^{r+1} = \underset{z}{\operatorname{argmin}}\ \langle G_\mu^J(z^r, \phi^r, \xi^r) + A^\top\lambda^r + \rho A^\top A z^r, z - z^r\rangle + \rho\|z - z^r\|_D^2, \tag{5.20}$$

$$\lambda^{r+1} = \lambda^r + \rho A z^{r+1}. \tag{5.21}$$

3: **end for**

4: **Output:** Choose (z^u, λ^u) uniformly and randomly from $\{(z^{r+1}, \lambda^{r+1})\}_{r=0}^{T-1}$.

However, for the problem that is of interest in this paper, the MM method is not applicable because of the following reasons: 1) The optimization problem (5.17) is not easily solvable to global optima because it is nonconvex, and we only have access to zeroth-order information; 2) It is not clear how to implement the algorithm in a distributed manner over the MNet. In contrast, the primal step of the ZONE-M algorithm (5.20) utilizes zeroth-order information and can be performed in closed-form. Further, as we elaborate below, combining the primal and the dual steps of ZONE-M yields a fully distributed algorithm.

To illustrate the distributed implementation of the proposed method, let us transform the ZONE-M algorithm to a *primal only* form. To this end, let us write down the optimality condition for (5.20) as

$$G_\mu^{J,r} + A^\top\lambda^r + \rho A^\top A z^r + 2\rho D(z^{r+1} - z^r) = 0. \tag{5.22}$$

Utilizing the definitions in (5.12a), and (5.12b), we have the following identity from (5.22)

$$G_\mu^{J,r} + A^\top \lambda^r + 2\rho D z^{r+1} - \rho L^+ z^r = 0. \tag{5.23}$$

Let us replace r in equation (5.23) with $r-1$ to get

$$G_\mu^{J,r-1} + A^\top \lambda^{r-1} + 2\rho D z^r - \rho L^+ z^{r-1} = 0. \tag{5.24}$$

Now rearranging the terms in (5.21) and using the definition in (5.12a) we have

$$A^\top(\lambda^r - \lambda^{r-1}) = \rho A^\top A z^r = \rho L^- z^r. \tag{5.25}$$

Subtracting equation (5.24) from (5.23) and utilizing (5.25) yield

$$G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^- z^r + 2\rho D(z^{r+1} - z^r) - \rho L^+(z^r - z^{r-1}) = 0.$$

Rearranging terms in the above identity, we obtain

$$z^{r+1} = z^r - \frac{1}{2\rho} D^{-1}\left[G_\mu^{J,r} - G_\mu^{J,r-1}\right] + \frac{1}{2} D^{-1}(L^+ - L^-) z^r - \frac{1}{2} D^{-1} L^+ z^{r-1}. \tag{5.26}$$

To implement such iteration, it is easy to check (by utilizing the definition of L^+ and L^-) that each agent i performs the following local computation

$$z_i^{r+1} = z_i^r - \frac{1}{2\rho d_i}\left[\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r) - \bar{G}_{\mu,i}(z_i^{r-1}, \phi_i^{r-1}, \xi_i^{r-1})\right] + \sum_{j\in\mathcal{N}_i} \frac{1}{d_i} z_j^r - \frac{1}{2}\left(\sum_{j\in\mathcal{N}_i} \frac{1}{d_i} z_j^{r-1} + z_i^{r-1}\right), \tag{5.27}$$

where $\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r)$ is defined in (5.19). Clearly, this is a fully decentralized algorithm, because to carry out such an iteration, each agent i only requires the knowledge about its local function [i.e., $\bar{G}_{\mu,i}(z_i^r, \phi_i^r, \xi_i^r)$, $\bar{G}_{\mu,i}(z_i^{r-1}, \phi^{r-1}, \xi^{r-1})$, z_i^r and z_i^{r-1}], as well as information from the agents in its neighborhood $\mathcal{N}_i$.

Remark 3 *The single variable iteration derived in* (5.26) *takes a similar form as the EXTRA algorithm proposed in [**?**], which uses the first-order gradient information. In EXTRA, the iteration is given by (for* $r \geq 2$*)*

$$z^{r+1} - z^r = W z^r - \frac{I_Q + W}{2} z^{r-1} - \alpha\left[\nabla g(z^r) - \nabla g(z^{r-1})\right].$$

where W is a double stochastic matrix.

In ZONE-M algorithm let us define $W := \frac{1}{2}D^{-1}(L^+ - L^-)$, which is a row stochastic matrix. Then iteration (5.26) becomes

$$z^{r+1} - z^r = Wz^r - \frac{I_Q + W}{2}z^{r-1} - \frac{1}{2\rho}D^{-1}\left[G_\mu^{J,r} - G_\mu^{J,r-1}\right],$$

which is similar to EXTRA algorithm. The key difference is that our algorithm is capable of utilizing zeroth-order information, to deal with nonconvex problems, while the EXTRA algorithm requires first-order (gradient) information, and it only deals with convex problems.

5.2.3 The Convergence Analysis of ZONE-M

In this subsection we provide the convergence analysis for ZONE-M algorithm. Besides Assumption A, we will further make the following assumptions.

Assumption B.

B1. Function $g(z)$ is $\hat{L}$-smooth, which satisfies (5.2).

B2. There exists a constant $\delta > 0$ such that

$$\exists\ \underline{g} > -\infty, \quad \text{s.t.} \quad g(z) + \frac{\delta}{2}\|Az\|^2 \geq \underline{g},\ \forall\ z \in \mathbb{R}^Q. \tag{5.28}$$

The above assumptions on the objective g is rather mild. The first one is standard for analyzing many first-order algorithms for nonconvex centralized optimization (see, e.g., [9]), while the second assumption only postulates that the objective function is bounded from below. Without loss of generality we can set $\underline{g} = 0$. A few examples of nonconvex functions that satisfy the Assumption A1, and B are provided below:

- The sigmoid function $\text{sig}(z) = \frac{1}{1+e^{-z}}$
- The function $\tanh(z) = \frac{1-e^{-2z}}{1+e^{-2z}}$
- The function $2\text{logit}(z) = \frac{2e^z}{e^z+1} = 1 + \tanh(\frac{z}{2})$

Note that these nonconvex functions are popular activation functions used in learning neural networks.

Before formally presenting the analysis, let us further define some additional notation to simplify the presentation. Let $\mathcal{F}^{r+1} := \{(\xi^t, \phi^t)\}_{t=1}^r$ be the σ-field generated by the entire history of algorithm up to iteration r. Let $\sigma_{\min}$ be the smallest nonzero eigenvalue of matrix $A^\top A$. Additionally, we define $w^r := (z^{r+1} - z^r) - (z^r - z^{r-1})$. Further to facilitate the proofs let us list a few relationships below.

- For any given vectors a and b we have

$$\langle b - a, b\rangle = \frac{1}{2}(\|b\|^2 + \|a - b\|^2 - \|a\|^2), \tag{5.29}$$

$$\langle a, b\rangle \leq \frac{1}{2\epsilon}\|a\|^2 + \frac{\epsilon}{2}\|b\|^2; \quad \forall\ \epsilon > 0. \tag{5.30}$$

- For n given vectors a_i we have the following

$$\Big\|\sum_{i=1}^n a_i\Big\|^2 \leq n\sum_{i=1}^n \big\|a_i\big\|^2. \tag{5.31}$$

Our convergence analysis consists of the following main steps: First we show that the successive difference of the dual variable, which represents the constraint violation, is bounded by a quantity related to the primal variable. Second we construct a special potential function whose behavior is tractable under a specific parameter selection. Third, we combine the previous results to obtain the main convergence rate analysis. Below we provide a sequence of lemmas and the main theorem. The proofs are provided in Appendix A. Unless otherwise stated, throughout this section the expectations are taken with respect to filtration $\mathcal{F}^{r+1}$ defined above.

Our first lemma bounds the change of the dual variables (in expectation) by that of the primal variables. This lemma will be used later to control the progress of the dual step of the algorithm.

Lemma 23 *Suppose Assumptions A, B hold true, and L^+ is the signless Laplacian matrix defined in (5.12b). Then for $r \geq 1$ we have the following inequity*

$$\mathbb{E}\|\lambda^{r+1} - \lambda^r\|^2 \leq \frac{9\tilde{\sigma}^2}{J\sigma_{\min}} + \frac{6L_\mu^2}{\sigma_{\min}}\mathbb{E}\|z^r - z^{r-1}\|^2 + \frac{3\rho^2\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2. \tag{5.32}$$

To proceed, we need to construct a potential function so that the behavior of the algorithm can be made tractable. For notational simplicity let us define $L_\rho^{r+1} := L_\rho(z^{r+1}, \lambda^{r+1})$. Also let $c > 0$ to be some positive constant (to be specified shortly), and set $k := 2\left(\frac{6\hat{L}^2}{\rho\sigma_{\min}} + \frac{3c\hat{L}}{2}\right)$. Let $B := L^+ + \frac{k}{c\rho} I_Q$, and define V^{r+1} as

$$V^{r+1} := \frac{\rho}{2}\left(\|Az^{r+1}\|^2 + \|z^{r+1} - z^r\|_B^2\right).$$

Using these notations, we define a potential function in the following form

$$P^{r+1} := L_\rho^{r+1} + cV^{r+1}. \tag{5.33}$$

The following lemma analyzes the behavior of the potential function as the ZONE-M algorithm proceeds.

Lemma 24 *Suppose Assumptions A, B are satisfied, and parameters c and ρ satisfy the following conditions*

$$c > \frac{6\|L^+\|}{\sigma_{\min}}, \quad \rho > \max\left(\frac{-b + \sqrt{b^2 - 8d}}{4}, \delta, \hat{L}/2\right), \tag{5.34}$$

where

$$b = -\hat{L}\left(\hat{L} + \frac{24\|L^+\|}{\sigma_{min}} + 1\right) - 3, d = -\frac{12\hat{L}^2}{\sigma_{\min}}.$$

Then for some constants $c_1, c_2, c_3 > 0$, the following inequality holds true for $r \geq 1$

$$\mathbb{E}\left[P^{r+1} - P^r\right] \leq \frac{k - c_1}{2}\mathbb{E}\|z^{r+1} - z^r\|^2 - c_2\mathbb{E}\|w^r\|_{L^+}^2 + c_3\frac{\tilde{\sigma}^2}{J} + \frac{3\mu^2(Q+3)^3}{8}, \tag{5.35}$$

where we have defined

$$c_1 := 2\rho - \hat{L}^2 - (c+1)\hat{L} - 3 > 0, \tag{5.36}$$
$$c_2 := \left(\frac{c\rho}{2} - \frac{3\rho\|L^+\|}{\sigma_{\min}}\right) > 0, \; c_3 := \frac{9}{\rho\sigma_{\min}} + \frac{3 + 6c\hat{L}}{2\hat{L}^2} > 0.$$

The key insight obtained from this step is that, a conic combination of augmented Lagrangian function, as well as the constraint violation can serve as the potential function that guides the

progress of the algorithm. We expect that such construction is of independent interest. It will be instrumental in analyzing other (probably more general) nonconvex primal-dual type algorithms.

The next lemma shows that P^{r+1} is lower bounded.

Lemma 25 *Suppose that Assumptions A, B are satisfied, and constant c is picked large enough such that*

$$c \geq \frac{-b_1 + \sqrt{b_1^2 - 4a_1 d_1}}{2a_1}, \tag{5.37}$$

where

$$a_1 = \frac{3\rho\hat{L}}{8}, b_1 = \frac{3\hat{L}^2}{\sigma_{\min}}, d_1 = -\frac{2\|L^+\|^2}{\sigma_{\min}}. \tag{5.38}$$

Then the statement below holds true

$$\exists\ \underline{P} \quad \text{s.t.} \quad \mathbb{E}[P^{r+1}] \geq \underline{P} > -\infty, \quad \forall\ r \geq 1. \tag{5.39}$$

where $\underline{P}$ is a constant that is independent of total number of iterations T.

To present our main convergence theorem, we need to have a way to measure the gap between the current iterate to the set of stationary solutions. To this end, consider the following gap function

$$\Phi(z^r, \lambda^{r-1}) := \mathbb{E}\Big[\|\nabla_z L_\rho(z^r, \lambda^{r-1})\|^2 + \|Az^r\|^2\Big]. \tag{5.40}$$

It can be easily checked that $\|\nabla_z L_\rho(z^*, \lambda^*)\|^2 + \|Az^*\|^2 = 0$ if and only if (z^*, λ^*) is a stationary solution of the problem (5.13). For notational simplicity let us write $\Phi^r := \Phi(z^r, \lambda^{r-1})$. The result below quantifies the convergence rate of ZONE-M.

Theorem 9 *Consider the ZONE-M algorithm with fixed total number of iterations T, and u is a number uniformly randomly sampled from the index set $\{1, 2, \cdots, T\}$. Suppose Assumptions A, B are satisfied, penalty parameter ρ satisfies in the condition given in Lemma 24, a_1, b_1, d_1 are those constants in (5.38), and constant c satisfies in*

$$c \geq \max\left(\frac{-b_1 + \sqrt{b_1^2 - 4a_1 d_1}}{2a_1}, \frac{6\|L^+\|}{\sigma_{\min}}\right). \tag{5.41}$$

Then there exists constants $\gamma_1, \gamma_2, \gamma_3 > 0$ such that we have the following bound

$$\mathbb{E}_u[\Phi^u] \leq \frac{\gamma_1}{T} + \frac{\gamma_2 \tilde{\sigma}^2}{J} + \gamma_3 \mu^2. \tag{5.42}$$

The explicit value for constants γ_1, γ_2, and γ_3 can be expressed as the following: Let

$$\alpha_1 = 8\hat{L} + 2\rho^2 \|L^+\|^2,\ \alpha_2 = \frac{6\hat{L}}{\rho^2 \sigma_{\min}},\ \alpha_3 = \frac{3\|L^+\|}{\sigma_{\min}},$$

and c_1, c_2 and c_3 are constants given in equation (5.36). Let us set $\zeta = \frac{\max(\alpha_1+\alpha_2, \alpha_3)}{\min(\frac{c_1-k}{2}, c_2)}$. Then we have the following expression

$$\gamma_1 = \zeta \mathbb{E}[P^1 - \underline{P}] + \alpha_2 \mathbb{E}\|z^1 - z^0\|^2$$
$$\gamma_2 = \zeta c_3 + \frac{9 + 4\rho^2 \sigma_{\min}}{\rho^2 \sigma_{\min}},\ \gamma_3 = \frac{3}{8}\zeta + 2\hat{L}^2.$$

Remark 4 *From the main result in Theorem 9 we can observe that the complexity bound of the ZONE-M depends on $\tilde{\sigma}$, and the smoothing parameter μ . Therefore, no matter how many iterations we run the algorithm, it always converges to a neighborhood of a KKT point, which is expected when only zeroth-order information is available; see [43, Theorem 4.4], and [45, Theorem 3.2]. Nevertheless, if we choose $J \in \mathcal{O}(T)$, and $\mu \in \mathcal{O}(\frac{1}{\sqrt{T}})$, we can achieve the following bound*

$$\mathbb{E}_u[\Phi^u] \leq \frac{\gamma_1}{T} + \frac{\gamma_2 \tilde{\sigma}^2}{T} + \frac{\gamma_3}{T}. \tag{5.43}$$

This indicates that ZONE-M converges in a sublinear rate.

Remark 5 *Our bound on ρ derived in (5.34) can be loose because it is obtained based on the the worst case analysis. In practice one may start with a small ρ and gradually increase it until reaching the theoretical bound. In the numerical experiments, we will see that such strategy often leads to faster empirical convergence.*

5.3 Zeroth-Order Algorithm over SNet

In this section we focus on multi-agent optimization problem over SNet (cf. Fig. 5.1). We propose the Zeroth-Order NonconvEx, over SNet (ZONE-S) algorithm for the multi-agent optimization problem.

5.3.1 System Model

Let us consider the following problem

$$\min_{x\in X} \; g(x) := \sum_{i=1}^{N} f_i(x) + r(x), \tag{5.44}$$

where $X \subseteq \mathbb{R}^M$ is a closed and convex set, $f_i : \mathbb{R}^M \to \mathbb{R}$ is smooth possibly nonconvex function, and $r : \mathbb{R}^M \to \mathbb{R}$ is a convex possibly nonsmooth function, which is usually used to impose some regularity to the solution. Let us set $f(x) := \sum_{i=1}^{N} f_i(x)$ for notational simplicity. Note that this problem is slightly more general than the one solved in the previous section [i.e., problem (5.1) with smooth objective function], because here we have included constraint set X and the nonsmooth function $r(x)$ as well.

We note that many first-order algorithms have been developed for solving problem (5.44), including SGD [112], SAG [31], SAGA [118], SVRG [71], and NESTT [57], but it is not clear how to adapt these methods and their analysis to the case with non-convex objective and zeroth-order information.

Similar to the problem over MNet, here we split the variable $x \in \mathbb{R}^M$ into local copies $z_i \in \mathbb{R}^M$, and reformulate problem (5.44) as

$$\min_{x,z} \; \sum_{i=1}^{N} f_i(z_i) + h(x) \text{ s.t. } x = z_i, \; \forall \; i = 1, \cdots, N, \tag{5.45}$$

where $h(x) := r(x) + \iota_X(x)$, [$\iota_X(x) = 0$ if $x \in X$, otherwise $\iota_X(x) = \infty$]. In this formulation we have assumed that for $i = 1, 2, \cdots N$, f_i is the local function for agent i, and $h(x)$ is handled by the central controller. Further, agent i has access to the functional values of f_i through the $\mathcal{SZO}$ as described in preliminaries.

5.3.2 Proposed Algorithm

The proposed algorithm is again a primal-dual based scheme. The augmented Lagrangian function for problem (5.45) is given by

$$L_\rho\Big(z, x; \lambda\Big) = \sum_{i=1}^{N} \left(f_i(z_i) + \langle \lambda_i, z_i - x\rangle + \frac{\rho_i}{2}\|z_i - x\|^2 \right) + h(x)$$

where λ_i, and ρ_i are respectively the dual variable and the penalty parameter associated with the constraint $z_i = x$. Let $\lambda := \{\lambda_i\}_{i=1}^N$, $\rho := \{\rho_i\}_{i=1}^N \in \mathbb{R}_{++}^N$. To proceed, let us introduce the following function for agent i

$$U_{\mu,i}(z_i, x; \lambda_i) = f_i(x) + \langle \bar{G}_{\mu,i}(x, \phi, \xi), z_i - x \rangle + \langle \lambda_i, z_i - x \rangle + \frac{\alpha_i \rho_i}{2} \|z_i - x\|^2. \tag{5.46}$$

In the above expression α_i is a positive constant, and $\bar{G}_{\mu,i}(x, \phi, \xi)$ is given by

$$\bar{G}_{\mu,i}(x, \phi, \xi) = \frac{1}{J} \sum_{j=1}^{J} \frac{\mathcal{H}_i(x + \mu\phi_j, \xi_j) - \mathcal{H}_i(x, \xi_j)}{\mu} \phi_j, \tag{5.47}$$

where $\mathcal{H}_i(x, \xi)$ is a noisy version of $f_i(x)$ obtained from $\mathcal{SZO}$, $\mu > 0$ is smoothing parameter, $\phi_j \in \mathbb{R}^M$ is a standard Gaussian random vector, ξ_j represents the noise related to the $\mathcal{SZO}$ output, and we set $\phi = \{\phi_j\}_{j=1}^J$, and $\xi = \{\xi_j\}_{j=1}^J$. To see more details about the characteristics of function $U_{\mu,i}(z_i, x; \lambda_i)$ the readers are refereed to [**?**].

The proposed algorithm is described below. At the beginning of iteration $r + 1$ the central controller broadcasts x^r to everyone. An agent indexed by $i_r \in \{1, 2, \cdots N\}$ is then randomly picked with some probability of p_{i_r}, and this agent optimizes $U_{\mu,i_r}(z_i, x^r, \lambda^r)$ [defined in (5.46)], and updates its dual variable λ_{i_r}. The rest of the nodes $j \neq i_r$ simply set $z_j^{r+1} = x^r$, and $\lambda_j^{r+1} = \lambda_j^r$. Finally the central controller updates the variable x by minimizing the augmented Lagrangian. The pseudo-code of the ZONE-S algorithm is presented in Algorithm 8.

5.3.3 Convergence Analysis of ZONE-S

In this part we analyze the behavior of ZONE-S algorithm. The proofs of the lemmas are provided in the appendix. We first make the following assumptions on the problem (5.44) besides Assumption A.

Assumption C

C1. For $i = 1, 2, \cdots, N$, function f_i and f are L_i-smooth, and L-smooth respectively.

C2. The function $g(x)$ is bounded from below over $X \cap \text{int}(\text{dom } (g))$.

C3. The function $r(x)$ is convex but possibly nonsmooth.

Algorithm 8 ZONE-S Algorithm

1: **Input:** $x^0 \in \mathbb{R}^M$, $\lambda^0 \in \mathbb{R}^M$, $T \geq 1$, $J \geq 1$, $\mu > 0$
2: **for** $r = 1$ **to** T, **do**

In central controller: Pick i_r from $\{1, 2, \cdots, N\}$ with probability $p_{i_r} = \frac{\sqrt{L_{\mu,i_r}}}{\sum_{i=1}^N \sqrt{L_{\mu,i}}}$. Generate $\phi_j^r \in \mathbb{R}^M$, $j = 1, 2, \cdots, J$ from an i.i.d standard Gaussian distribution
In agent i_r: Calculate $\bar{G}_{\mu,i_r}(x^r, \phi^r, \xi^r)$ using

$$\bar{G}_{\mu,i_r}(x^r, \phi^r, \xi^r) = \frac{1}{J}\sum_{j=1}^{J} \frac{\mathcal{H}_{i_r}(x^r + \mu\phi_j^r, \xi_j^r) - \mathcal{H}_{i_r}(x^r, \xi_j^r)}{\mu}\phi_j^r, \tag{5.48}$$

where we set $\phi^r = \{\phi_j^r\}_{j=1}^J$, and $\xi^r = \{\xi_j^r\}_{j=1}^J$.
In all agents: Update z, and λ by

$$z_{i_r}^{r+1} = x^r - \frac{1}{\alpha_{i_r}\rho_{i_r}}\left[\lambda_{i_r}^r + \bar{G}_{\mu,i_r}(x^r, \phi^r, \xi^r)\right]; \tag{5.49}$$

$$\lambda_{i_r}^{r+1} = \lambda_{i_r}^r + \alpha_{i_r}\rho_{i_r}\left(z_{i_r}^{r+1} - x^r\right); \tag{5.50}$$

$$\lambda_j^{r+1} = \lambda_j^r, \quad z_j^{r+1} = x^r, \quad \forall\ j \neq i_r. \tag{5.51}$$

In central controller: Update x by

$$x^{r+1} = \arg\min_{x \in X} L_\rho(z^{r+1}, x; \lambda^r). \tag{5.52}$$

3: **end for**
4: **Output:** x^u chosen randomly from $\{x^r\}_{r=1}^T$.

Let us define the auxiliary sequence $y^r := \{y_i^r\}_{i=1}^N$ as follows

$$y^0 = x^0,\ y_j^r = y_j^{r-1},\ \text{ if } j \neq i_r, \text{ else } y_{i_r}^r = x^r, \quad \forall\ r \geq 1. \tag{5.53}$$

Next let us define the potential function which measures the progress of algorithm

$$\tilde{Q}^r = \sum_{i=1}^N f_{\mu,i}(x^r) + \sum_{i=1}^N \frac{4}{\alpha_i\rho_i}\|\nabla f_{\mu,i}(y_i^{r-1}) - \nabla f_{\mu,i}(x^r)\|^2 + h(x^r),$$

where $f_{\mu,i}(x^r)$ denotes the smoothed version of function $f_i(x^r)$ defined in (5.3).

First, we study the behavior of the potential function. For this algorithm let us define the filtration $\mathcal{F}^{r+1}$ as the σ-field generated by $\{i_t, \phi^t, \xi^t\}_{t=1}^r$. Throughout this section the expectations are taken with respect to $\mathcal{F}^{r+1}$ unless otherwise noted.

Lemma 26 *Suppose Assumption C holds true, set* $\tilde{p} := \sum_{i=1}^N \frac{1}{p_i}$, $\beta := \frac{1}{\sum_{i=1}^N \rho_i}$, *and for* $i = 1, 2, \cdots, N$, *we pick*

$$\alpha_i = p_i = \frac{\rho_i}{\sum_{i=1}^N \rho_i}, \text{ and } \rho_i \geq \frac{5.5 L_{\mu,i}}{p_i}, \ i = 1, \cdots N. \tag{5.54}$$

Then we have the following result for ZONE-S algorithm

$$\mathbb{E}[\tilde{Q}^{r+1} - \tilde{Q}^r] \leq \frac{-1}{100\beta}\mathbb{E}\|x^{r+1} - x^r\|^2 - \sum_{i=1}^N \frac{1}{2\rho_i}\mathbb{E}\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1}))\|^2 + \frac{3\tilde{p}\beta\tilde{\sigma}^2}{J}. \tag{5.55}$$

Next we define the optimality gap as the following

$$\Psi^r := \frac{1}{\beta^2}\mathbb{E}\left\|x^r - \text{prox}_h^{1/\beta}[x^r - \beta\nabla f(x^r)]\right\|^2, \tag{5.56}$$

where $\text{prox}_h^{\gamma}[u] := \text{argmin} \ \ h(u) + \frac{\gamma}{2}\|x - u\|^2$ is the proximity operator for function h. Note that when the nonsmooth term $h \equiv 0$, Ψ^r reduces to the size of the gradient vector $\mathbb{E}\|\nabla f(x^r)\|^2$.

Finally we present the main convergence result about the proposed ZONE-S algorithm.

Theorem 10 *Suppose Assumptions A (for each function* f_i*), and Assumption C hold, and* u *is uniformly randomly sampled from* $\{1, 2, \cdots, T\}$. *Let us set* $\frac{1}{\beta} = 4(\sum_{i=1}^N \sqrt{L_{\mu,i}})^2$. *Then we have the following bounds for the optimality gap in expectation*

$$\begin{aligned}
&1) \ \mathbb{E}_u[\Psi^u] \leq 2200\Big(\sum_{i=1}^N \sqrt{L_{\mu,i}}\Big)^2 \frac{\mathbb{E}[\tilde{Q}^1 - \tilde{Q}^{T+1}]}{T} + \frac{\mu^2 L^2 (M+3)^3}{2} + \frac{1024\tilde{p}\tilde{\sigma}}{J}; \\
&2) \ \mathbb{E}_u[\Psi^u] + \mathbb{E}_u\Big[\sum_{i=1}^N 3\rho_i^2 \Big\|z_i^u - x^{u-1}\Big\|^2\Big] \\
&\quad \leq 2200\Big(\sum_{i=1}^N \sqrt{L_{\mu,i}}\Big)^2 \frac{\mathbb{E}[\tilde{Q}^1 - \tilde{Q}^{T+1}]}{T} + \frac{\mu^2 L^2 (M+3)^3}{2} + \frac{1024\tilde{p}\tilde{\sigma}}{J}.
\end{aligned}$$

Note that part (1) only measures the primal optimality gap, while part (2) also shows that the expected constraint violation shrinks in the same order.

Remark 6 *Similar to the ZONE-M, the bound for the optimality gap of ZONE-S is dependent on two* T*-independent constants, the first one* $\frac{\mu^2 L^2 (M+3)^3}{2}$ *arises from using zeroth-order gradient, and*

the second term $\frac{1024\tilde{p}\tilde{\sigma}^2}{J}$ *arise from the uncertainty in the gradient estimation. Again, if we pick* $\mu \in \mathcal{O}(\frac{1}{\sqrt{T}})$*, and* $J \in \mathcal{O}(T)$*, we obtain the following sublinear convergence rate*

$$\mathbb{E}_u[\Psi^u] \leq 2200\Big(\sum_{i=1}^{N}\sqrt{L_{\mu,i}}\Big)^2\frac{\mathbb{E}[\tilde{Q}^1-\tilde{Q}^{T+1}]}{T}+\frac{1024\tilde{p}\tilde{\sigma}^2}{T}+\frac{L^2(M+3)^3}{T}. \tag{5.57}$$

Remark 7 *The reason that ZONE-S is able to incorporate non-smooth terms, in contrast to the ZONE-M algorithm, is that it has special network structure. In particular, the non-smooth term is optimized by the central controller, and the fact that the central controller can talk to every node makes sure that the non-smooth term is optimized by using the most up-to-date information from the network.*

5.4 Numerical Results

In this section we numerically evaluate the effectiveness of the ZONE-M and ZONE-S algorithms. We consider some distributed nonconvex optimization problems in zeroth-order setup (i.e., we only have access to the noisy functional values). However, in order to simply check whether the considered problems are fit into our model here we express the explicit form of the objective functions for the considered problems. We set the noise ξ to be a zero-mean Gaussian random variable with standard deviation $\sigma = 0.01$. All the simulations are performed on Matlab 2015a on a Laptop with 4 GB memory and Intel Core i7-4510U CPU (2.00 GHz), running on Linux (Ubuntu 16.04) operating system.

5.4.1 ZONE-M Algorithm

We study the following two nonconvex distributed optimization problems.

Distributed Nonconvex Consensus. Consider minimizing sum of nonconvex functions in a distributed setting

$$\min_{z\in\mathbb{R}^Q}\sum_{i=1}^{N} f_i(z_i), \quad \text{s.t. } Az = 0. \tag{5.58}$$

where each agent i can only obtain the zeroth-order information of its local function, given by

$$f_i(z_i) = \frac{a_i}{1+e^{-z_i}} + b_i\log(1+z_i^2),$$

where a_i and b_i are constants generated from an i.i.d Gaussian distribution. Clearly the function f_i is nonconvex and smooth, and we can simply check that it satisfies assumption A, B. In our experiments the graphs are generated based on the scheme proposed in [141]. In this scheme a random graph with N nodes and radius R is generated with nodes uniformly distributed over a unit square, and two nodes connect if their distance is less than R. We set problem dimension $M = 1$, and the number of nodes in the network $N = 20$ with radius $R = 0.6$. The penalty parameter ρ is selected to satisfy theoretical bounds given in Lemma 24, the smoothing parameter is set $\mu = \frac{1}{\sqrt{T}}$, and we set $J = T$, where maximum number of iterations is picked $T = 1000$. We compare the ZONE-M algorithm with Randomized Gradient Free (RGF) algorithm with diminishing stepsize $\frac{1}{\sqrt{r}}$ (r denotes the iterations counter) proposed in [142], which is only developed for convex problems. We also compare our algorithm with a variant of ZONE-M which uses increasing penalty parameter $\rho = \sqrt{r}$. When choosing $\rho = \sqrt{r}$ neither RFG not ZONE-M has convergence guarantee. We use the optimality gap (opt-gap) and constraint violation (cons-vio), displayed below, to measure the quality of the solution generated by different algorithms

$$\begin{aligned} \text{opt-gap} &:= \left\| \sum_{i=1}^{N} \nabla f_i(z_i) \right\|^2 + \|Az\|^2, \\ \text{cons-vio} &:= \|Az\|^2. \end{aligned} \tag{5.59}$$

Figure 5.5 illustrates the comparison among different algorithms. Each point in the figure is obtained by averaging over 50 independent trials. One can observe that: 1) ZONE-M converges faster compared with RGF in both the optimality gap and the consensus error; 2) ZONE-M with increasing penalty ($\rho = \sqrt{r}$) appears to be faster than its constant stepsize counterpart.

In the next set of experiments we compare different algorithms with a number of choices of network size, i.e., $N \in \{10, 20, 40, 80\}$. For this problem we set the radius $R = 0.5$ The results (average over 50 independent trials) are reported in Table 5.1. In this table ZONE-M (C) and ZONE-M (I) denote ZONE-M with constant and increasing stepsize, respectively. We observe that ZONE-M algorithm is always faster compared with the RGF.

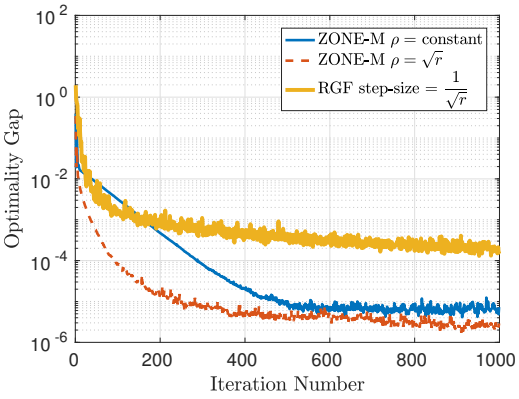


Figure 5.3 The optimality gap versus iteration counter

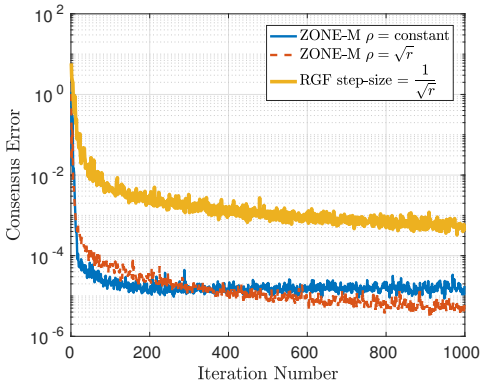


Figure 5.4 The constraint violation versus iteration counter

Figure 5.5: Comparison of different algorithms for the nonconvex consensus problem given in (5.58).

5.4.2 ZONE-S Algorithm

In this subsection we explore the effectiveness of ZONE-S algorithm. The penalty parameter ρ is selected to satisfy the conditions given in Lemma 26, or to be an increasing sequence satisfying $\rho = \sqrt{r}$. For comparison purpose we consider two additional algorithms, namely the zeroth-order gradient descent (ZO-GD) [105] (which is a centralized algorithm), and the zeroth-order stochastic gradient descent (ZO-SGD) [45]. To be notationally consistent with our algorithm we denote the stepsize for these two algorithms with $1/\rho$. For ZO-GD it has been shown that if the stepsize is set $1/\rho = \frac{1}{4L(M+4)}$, and the smoothing factor satisfies $\mu \leq \mathcal{O}(\frac{\epsilon}{ML})$, then the algorithm will converge to an ϵ-stationary solution [105, Section 7]. Also, for ZO-SGD the optimality gap decreases in the order of $\frac{1}{\sqrt{T}}$ when we pick stepsize $1/\rho < \frac{1}{2(M+4)}$, and the smoothing parameter μ satisfies $\mu \leq \frac{D_f}{(M+4)\sqrt{2N}}$, where $D_f := \left[\frac{2(f(x^1)-f^*)}{L}\right]^{1/2}$ (f^* denotes the optimal value) [45, Theorem 3.2]. Note that the theoretical results for ZO-SGD is valid only for smooth cases, however we include it here for comparison purposes.

Nonconvex Sparse Optimization Problem. Consider the following optimization problem

$$\min_{x\in\mathbb{R}^M} \sum_{i=1}^{N} f_i(x) \text{ s.t. } \|x\|_1 \leq \ell, \tag{5.60}$$

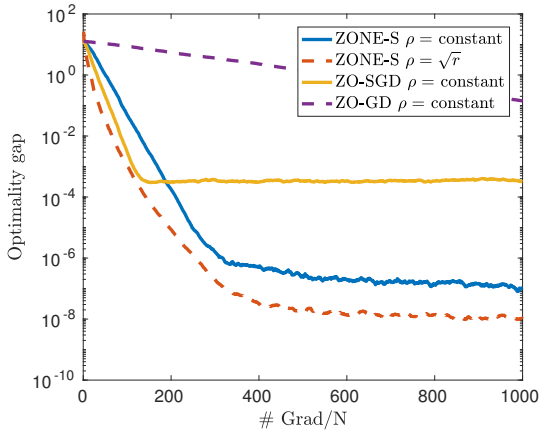


Figure 5.6: The Optimality Gap for Nonconvex Sparse Optimization problem.

where $f_i(x) = x^\top \Gamma x - \gamma^\top x$, ($\Gamma \in \mathbb{R}^{M\times M}$, and $\gamma \in \mathbb{R}^M$), and ℓ is a positive constant that controls the sparsity level of the solution. In this problem the matrix $\Gamma \in \mathbb{R}^{M\times M}$ is not necessarily a positive semidefinite matrix, thus the problem is not convex; see for example high dimensional regression problem with noisy observations in [90, problem (2.4)] . This problem is a special case of the original problem in (5.44) with $h(x)$ being the indicator function of the set $\{x \mid \|x\|_1 \leq \ell\}$.

We compare the following four algorithms: ZONE-S with constant stepsize $\rho_i = \sqrt{5.5L_{\mu,i}}\sum_{i=1}^{N}\sqrt{5.5L_{\mu,i}}$; ZONE-S with increasing penalty parameter $\rho_i = \sqrt{r}$; ZO-GD with constant stepsize ($1/\rho = \frac{1}{4L(M+4)}$), and ZO-SGD with constant step size $1/\rho = \frac{1}{2L(M+4)}$. The problem dimension is set as $N = 10$, and $M = 100$. The algorithm stops when the iteration counter reaches $T = 1000$. The results are plotted in Figure 5.6, which depicts the progress of the optimality gap [defined as in (5.56)] versus the number of iterations. Each point in this figure is obtained by averaging over 50 independent trials. We can observe that ZONE-S converges faster than the ZO-GD and ZO-SGD. Furthermore, the performance of ZONE-S improves when using the increasing stepsize, as compared to that of the constant stepsize.

5.5 Conclusion

In this work, we consider nonconvex multi-agent optimization problem under zeroth-order setup. We design algorithms to solve the problem over two popular network structures, namely MNet and

SNet. We have rigorously analyzed the convergence rate of the proposed algorithms and we have proved that both algorithms converge to the set of first-order stationary solutions under very mild conditions on the problem and by appropriately choosing the algorithm parameters.

5.6 Appendix. Proofs for ZONE-M

In this appendix we provide the proofs related to the convergence analysis of ZONE-M algorithm.

5.6.1 Proof of Lemma 23

Rearranging terms in (5.21) we get

$$\lambda^{r+1} - \lambda^r = \rho A z^{r+1}. \tag{5.61}$$

Applying this equality and the definition of L^+ in (5.12b) into the optimality condition for problem (5.20), we obtain

$$G_\mu^{J,r} + A^\top \lambda^{r+1} + \rho L^+ (z^{r+1} - z^r) = 0. \tag{5.62}$$

From (5.61) it is clear that $\lambda^{r+1} - \lambda^r$ lies in the column space of A, therefore the following is true

$$\sqrt{\sigma_{\min}} \|\lambda^{r+1} - \lambda^r\| \leq \|A^\top (\lambda^{r+1} - \lambda^r)\|, \tag{5.63}$$

where $\sigma_{\min}$ denotes the smallest non-zero eigenvalue of $A^\top A$.

Replacing r with $r-1$ in equation (5.62), and then using the definition of $w^r := (z^{r+1} - z^r) - (z^r - z^{r-1})$ we further get

$$\begin{aligned}
\left\|\lambda^{r+1} - \lambda^r\right\|^2 &\leq \frac{1}{\sigma_{\min}} \left\|G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r\right\|^2 \\
&= \frac{1}{\sigma_{\min}} \|G_\mu^{J,r} - G_\mu^{J,r-1} + \nabla g_\mu(z^r) - \nabla g_\mu(z^r) + \rho L^+ w^r\|^2 \\
&\overset{(5.31)}{\leq} \frac{3}{\sigma_{\min}} \|G_\mu^{J,r} - \nabla g_\mu(z^r)\|^2 + \frac{3}{\sigma_{\min}} \|\nabla g_\mu(z^r) - G_\mu^{J,r-1}\|^2 + \frac{3\rho^2}{\sigma_{\min}} \|L^+ w^r\|^2.
\end{aligned} \tag{5.64}$$

Let us add and subtract $\nabla g_\mu(z^{r-1})$ to the second term on the RHS of (5.64), and take the expectation on both sides

$$\begin{aligned}
&\mathbb{E}\|\lambda^{r+1}-\lambda^r\|^2 \le \frac{3}{\sigma_{\min}}\mathbb{E}\|G_\mu^{J,r}-\nabla g_\mu(z^r)\|^2 \\
&+\frac{6}{\sigma_{\min}}\mathbb{E}\|\nabla g_\mu(z^r)-\nabla g_\mu(z^{r-1})\|^2+\frac{3\rho^2}{\sigma_{\min}}\mathbb{E}\|L^+w^r\|^2+\frac{6}{\sigma_{\min}}\mathbb{E}\|\nabla g_\mu(z^{r-1})-G_\mu^{J,r-1}\|^2 \\
&\overset{\text{(i)}}{\le}\frac{9\tilde{\sigma}^2}{J\sigma_{\min}}+\frac{6L_\mu^2}{\sigma_{\min}}\mathbb{E}\|z^r-z^{r-1}\|^2+\frac{3\rho^2\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2, \qquad (5.65)
\end{aligned}$$

where (i) is true by applying Lemma 22 and utilizing the facts that $\nabla g_\mu(z)$ is L_μ-smooth and $\|L^+w^r\|^2 \le \|L^+\|\|w^r\|_{L^+}^2$. The lemma is proved. **Q.E.D.**

5.6.2 Proof of Lemma 24

Using Assumption B.1, and the fact that $D \succeq I$, it can be shown that if $2\rho \ge \hat{L}$, then function

$$L_\rho(z,\lambda)+\frac{\rho}{2}\|z-z^r\|_{L^+}^2 = g(z)+\langle\lambda,Az\rangle+\frac{\rho}{2}\|Az\|^2+\frac{\rho}{2}\|z-z^r\|_{L^+}^2,$$

is strongly convex with modulus $2\rho-\hat{L}$. See [150, Theorem 2.1]. Using this fact, let us bound $L_\rho^{r+1}-L_\rho^r$.

$$\begin{aligned}
&L_\rho^{r+1}-L_\rho^r = L_\rho^{r+1}-L_\rho(z^{r+1},\lambda^r)+L_\rho(z^{r+1},\lambda^r)-L_\rho^r \\
&\overset{\text{(i)}}{=}\langle\nabla_z L_\rho(z^{r+1},\lambda^r)+\rho L^+(z^{r+1}-z^r),z^{r+1}-z^r\rangle+\frac{1}{\rho}\|\lambda^{r+1}-\lambda^r\|^2-\frac{2\rho-\hat{L}}{2}\|z^{r+1}-z^r\|^2. \qquad (5.66)
\end{aligned}$$

where (i) is true due to the strong convexity of $L_\rho(z,\lambda)+\frac{\rho}{2}\|z-z^r\|_{L^+}^2$ with modulus $2\rho-\hat{L}$ and (5.61). Now using (5.62) we further have

$$\begin{aligned}
&L_\rho^{r+1}-L_\rho^r \le \left\langle\nabla g(z^{r+1})-G_\mu^{J,r},z^{r+1}-z^r\right\rangle+\frac{1}{\rho}\|\lambda^{r+1}-\lambda^r\|^2-\frac{2\rho-\hat{L}}{2}\|z^{r+1}-z^r\|^2 \\
&\overset{\text{(i)}}{\le}\frac{1}{\rho}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\hat{L}^2-2\rho+\hat{L}}{2}\|z^{r+1}-z^r\|^2+\frac{1}{2\hat{L}^2}\|\nabla g(z^{r+1})-G_\mu^{J,r}\|^2,
\end{aligned}$$

where (i) is application of (5.30) for $\epsilon = \hat{L}^2$. Taking expectation on both sides we get

$$\begin{aligned}\mathbb{E}\left[L_\rho^{r+1} - L_\rho^r\right] &\overset{\text{(i)}}{\leq} \frac{9\tilde{\sigma}^2}{\rho J \sigma_{\min}} + \frac{6L_\mu^2}{\rho\sigma_{\min}}\mathbb{E}\|z^r - z^{r-1}\|^2 + \frac{3\rho\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2 \\ &+ \frac{\hat{L}^2 - 2\rho + \hat{L}}{2}\mathbb{E}\|z^{r+1} - z^r\|^2 + \frac{1}{2\hat{L}^2}\mathbb{E}\|\nabla g(z^{r+1}) - G_\mu^{J,r}\|^2 \\ &\overset{\text{(ii)}}{\leq} \left(\frac{9}{\rho\sigma_{\min}} + \frac{3}{2\hat{L}^2}\right)\frac{\tilde{\sigma}^2}{J} + \frac{3\mu^2(Q+3)^3}{8} + \frac{6L_\mu^2}{\rho\sigma_{\min}}\mathbb{E}\|z^r - z^{r-1}\|^2 \\ &+ \frac{3\rho\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2 + \frac{\hat{L}^2 - 2\rho + \hat{L} + 3}{2}\mathbb{E}\|z^{r+1} - z^r\|^2, \end{aligned} \tag{5.67}$$

where in (i) we use Lemma 23 to bound $\mathbb{E}\|\lambda^{r+1} - \lambda^r\|$, in (ii) we apply (5.31), (5.5), (5.9), and the fact that $\nabla g_\mu(z)$ is L_μ-smooth with $L_\mu \leq \hat{L}$.

Next we bound $V^{r+1} - V^r$. Optimality condition for problem (5.20) together with equation (5.21) yield the following

$$\langle G_\mu^{J,r} + A^\top\lambda^{r+1} + \rho L^+(z^{r+1} - z^r), z^{r+1} - z\rangle \leq 0,\ \forall\ z \in \mathbb{R}^Q.$$

Similarly, for the $(r-1)$th iteration, we have

$$\langle G_\mu^{J,r-1} + A^\top\lambda^{r} + \rho L^+(z^{r} - z^{r-1}), z^{r} - z\rangle \leq 0,\ \forall\ z \in \mathbb{R}^Q.$$

Now let us set $z = z^r$ in first, $z = z^{r+1}$ in second equation, and add them. We obtain

$$\langle A^\top(\lambda^{r+1} - \lambda^r), z^{r+1} - z^r\rangle \leq -\langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, z^{r+1} - z^r\rangle. \tag{5.68}$$

The left hand side can be expressed in the following way

$$\begin{aligned}\langle A^\top(\lambda^{r+1} - \lambda^r), z^{r+1} - z^r\rangle &= \rho\langle Az^{r+1}, Az^{r+1} - Az^r\rangle \\ &\overset{(5.29)}{=} \frac{\rho}{2}\left(\|Az^{r+1}\|^2 - \|Az^r\|^2 + \|A(z^{r+1} - z^r)\|^2\right). \end{aligned} \tag{5.69}$$

For the right hand side we have

$$
\begin{aligned}
&- \langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, z^{r+1} - z^r \rangle \\
&= -\langle G_\mu^{J,r} - G_\mu^{J,r-1}, z^{r+1} - z^r \rangle - \langle \rho L^+ w^r, z^{r+1} - z^r \rangle \\
&\overset{(5.30)}{\leq} \frac{1}{2\hat{L}} \|G_\mu^{J,r} - G_\mu^{J,r-1}\|^2 + \frac{\hat{L}}{2}\|z^{r+1} - z^r\|^2 - \rho \langle L^+ w^r, z^{r+1} - z^r \rangle \\
&\overset{\text{(i)}}{\leq} \frac{3}{2\hat{L}} \Big(\|G_\mu^{J,r} - \nabla g_\mu(z^r)\|^2 + \|\nabla g_\mu(z^{r-1}) - G_\mu^{J,r-1}\|^2 \\
&+ \|\nabla g_\mu(z^r) - \nabla g_\mu(z^{r-1})\|^2 \Big) + \frac{\hat{L}}{2}\|z^{r+1} - z^r\|^2 - \rho \langle L^+ w^r, z^{r+1} - z^r \rangle.
\end{aligned}
$$

To get (i) we add and subtract $\nabla g_\mu(z^r) + \nabla g_\mu(z^{r-1})$ to $G_\mu^{J,r} - G_\mu^{J,r-1}$ and use (5.31). Taking expectation on both sides, we have

$$
\begin{aligned}
&- \mathbb{E}\langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, z^{r+1} - z^r \rangle \\
&\overset{(5.9)}{\leq} \frac{3}{2\hat{L}} \left(\frac{2\tilde{\sigma}^2}{J} + \hat{L}^2 \|z^r - z^{r-1}\|^2 \right) + \frac{\hat{L}}{2}\mathbb{E}\|z^{r+1} - z^r\|^2 - \rho \mathbb{E}\langle L^+ w^r, z^{r+1} - z^r \rangle \\
&\overset{\text{(i)}}{=} \frac{3\tilde{\sigma}^2}{\hat{L}J} + \frac{3\hat{L}}{2}\mathbb{E}\|z^r - z^{r-1}\|^2 + \frac{\hat{L}}{2}\mathbb{E}\|z^{r+1} - z^r\|^2 \\
&+ \frac{\rho}{2}\mathbb{E}\left(\|z^r - z^{r-1}\|_{L^+}^2 - \|z^{r+1} - z^r\|_{L^+}^2 - \|w^r\|_{L^+}^2 \right),
\end{aligned} \tag{5.70}
$$

where in (i) we apply (5.29) with $b = (L^+)^{1/2}(z^{r+1} - z^r)$ and $a = (L^+)^{1/2}(z^r - z^{r-1})$. Combining (5.68), (5.69) and (5.70), we obtain

$$
\begin{aligned}
&\frac{\rho}{2}\mathbb{E}\left(\|Az^{r+1}\|^2 - \|Az^r\|^2 + \|A(z^{r+1} - z^r)\|^2 \right) \\
&\leq \frac{3\tilde{\sigma}^2}{\hat{L}J} + \frac{3\hat{L}}{2}\mathbb{E}\|z^r - z^{r-1}\|^2 + \frac{\hat{L}}{2}\mathbb{E}\|z^{r+1} - z^r\|^2 \\
&+ \frac{\rho}{2}\mathbb{E}\left(\|z^r - z^{r-1}\|_{L^+}^2 - \|z^{r+1} - z^r\|_{L^+}^2 - \|w^r\|_{L^+}^2 \right).
\end{aligned} \tag{5.71}
$$

Recall that matrix $B := L^+ + \frac{k}{c\rho} I_Q$, and V^{r+1} is defined as

$$
\begin{aligned}
V^{r+1} &:= \frac{\rho}{2}\left(\|Az^{r+1}\|^2 + \|z^{r+1} - z^r\|_B^2 \right) \\
&= \frac{\rho}{2}\left(\|Az^{r+1}\|^2 + \|z^{r+1} - z^r\|_{L^+}^2 \right) + \frac{k}{2c}\|z^{r+1} - z^r\|^2.
\end{aligned}
$$

Rearranging terms in (5.71), we have

$$\begin{aligned}\mathbb{E}[V^{r+1}-V^r] &\leq \left(\frac{\hat{L}}{2}+\frac{k}{2c}\right)\mathbb{E}\|z^{r+1}-z^r\|^2+\frac{3\tilde{\sigma}^2}{\hat{L}J}\\ &+\left(\frac{3\hat{L}}{2}-\frac{k}{2c}\right)\mathbb{E}\|z^r-z^{r-1}\|^2-\frac{\rho}{2}\mathbb{E}\left(\|w^r\|^2_{L^+}+\|A(z^{r+1}-z^r)\|^2\right)\\ &\leq \left(\frac{\hat{L}}{2}+\frac{k}{2c}\right)\mathbb{E}\|z^{r+1}-z^r\|^2+\left(\frac{3\hat{L}}{2}-\frac{k}{2c}\right)\mathbb{E}\|z^r-z^{r-1}\|^2-\frac{\rho}{2}\mathbb{E}\|w^r\|^2_{L^+}+\frac{3\tilde{\sigma}^2}{\hat{L}J}.\end{aligned} \tag{5.72}$$

Now let us consider the definition of $P^{r+1}:=L_\rho^{r+1}+cV^{r+1}$. Utilizing (5.67), and (5.72) and definition of k as $k:=2\left(\frac{6\hat{L}^2}{\rho\sigma_{\min}}+\frac{3c\hat{L}}{2}\right)$ eventually we obtain

$$\mathbb{E}\left[P^{r+1}-P^r\right]\leq\frac{k-c_1}{2}\mathbb{E}\|z^{r+1}-z^r\|^2-c_2\mathbb{E}\|w^r\|^2_{L^+}+\frac{3\mu^2(Q+3)^3}{8}+c_3\frac{\tilde{\sigma}^2}{J}, \tag{5.73}$$

where we define,

$$c_1:=2\rho-\hat{L}^2-(c+1)\hat{L}-3,\quad c_2:=\left(\frac{c\rho}{2}-\frac{3\rho\|L^+\|}{\sigma_{\min}}\right),\quad c_3:=\frac{9}{\rho\sigma_{\min}}+\frac{3}{2\hat{L}^2}+\frac{3c}{\hat{L}}.$$

The lemma is proved. **Q.E.D.**

5.6.3 Proof of Lemma 25

Similar to the proof of Lemma 23 utilizing the optimality condition for x-subproblem we have

$$G_\mu^{J,r}+A^\top\lambda^{r+1}+\rho L^+(z^{r+1}-z^r)=0. \tag{5.74}$$

From this equation we have

$$\begin{aligned}\|A^\top\lambda^{r+1}\|^2&=\|G_\mu^{J,r}+\rho L^+(z^{r+1}-z^r)\|^2\\ &\leq 2\|G_\mu^{J,r}\|^2+2\rho^2\|L^+\|^2\|z^{r+1}-z^r\|^2.\end{aligned} \tag{5.75}$$

From here, we further have

$$\sigma_{\min}\|\lambda^{r+1}\|^2\leq 2\|G_\mu^{J,r}\|^2+2\rho^2\|L^+\|^2\|z^{r+1}-z^r\|^2. \tag{5.76}$$

Dividing both sides by $\sigma_{\min}$ yields

$$\|\lambda^{r+1}\|^2 \leq \frac{2}{\sigma_{\min}}\|G_{\mu}^{J,r}\|^2 + \frac{2\rho^2\|L^+\|^2}{\sigma_{\min}}\|z^{r+1}-z^r\|^2. \tag{5.77}$$

Now based on the definition of potential function P^{r+1} in equation (5.33) we have

$$P^{r+1} = g(z^{r+1}) + \frac{\rho}{2}\|Az^{r+1} + \frac{1}{\rho}\lambda^{r+1}\|^2 - \frac{1}{\rho^2}\|\lambda^{r+1}\|^2 + \frac{c\rho}{2}\|Az^{r+1}\|^2 + \frac{c\rho}{2}\|z^{r+1}-z^r\|_B^2, \tag{5.78}$$

where $B := L^+ + \frac{k}{2}I$ [note that $k = 2(\frac{6\hat{L}^2}{\rho\sigma_{\min}} + \frac{3c\hat{L}}{2})$]. Plugging (5.77) into (5.78), and utilizing the fact that $g(z^{r+1}) \geq 0$, $\frac{c\rho}{2}\|Az^{r+1}\|^2 \geq 0$, and $\|Az^{r+1} + \frac{1}{\rho}\lambda^{r+1}\|^2 \geq 0$ we get

$$P^{r+1} \geq -\frac{2}{\rho^2\sigma_{\min}}\|G_{\mu}^{J,r}\|^2 - \frac{2\|L^+\|^2}{\sigma_{\min}}\|z^{r+1}-z^r\|^2 + \frac{c\rho}{2}\|z^{r+1}-z^r\|_B^2. \tag{5.79}$$

Since L^+ is PSD matrix we have $B \succeq \frac{k}{2}I$. Also because $\frac{6\hat{L}^2}{\rho\sigma_{\min}} \geq 0$, we have $B \succeq \frac{3c\hat{L}}{2}I$. Utilizing this, we can simplify the above inequality as follows

$$P^{r+1} \geq -\frac{2}{\rho^2\sigma_{\min}}\|G_{\mu}^{J,r}\|^2 + (z^{r+1}-z^r)H(z^{r+1}-z^r), \tag{5.80}$$

where $H := (-\frac{2\|L^+\|^2}{\sigma_{\min}} + \frac{3\hat{L}^2}{\sigma_{\min}}c + \frac{3\rho\hat{L}^2}{8}c^2)I$. Therefore, if

$$c \geq \frac{-b_1 + \sqrt{b_1^2 - 4a_1d_1}}{2a_1}, \tag{5.81}$$

where

$$a_1 = \frac{3\rho\hat{L}}{8}, b_1 = \frac{3\hat{L}^2}{\sigma_{\min}}, d_1 = -\frac{2\|L^+\|^2}{\sigma_{\min}},$$

then we have $(z^{r+1}-z^r)H(z^{r+1}-z^r) \geq 0$. Hence, with this choice of c we get the following bound for the potential function

$$P^{r+1} \geq -\frac{2}{\rho^2\sigma_{\min}}\|G_{\mu}^{J,r}\|^2. \tag{5.82}$$

Tacking expectation on both sides we have

$$\mathbb{E}[P^{r+1}] \geq -\frac{2}{\rho^2\sigma_{\min}}\mathbb{E}\|G_{\mu}^{J,r}\|^2. \tag{5.83}$$

Now let us prove that $\mathbb{E}\|G_\mu^{J,r}\|^2$ is upper bounded as follows:

$$\begin{aligned}
\mathbb{E}\|G_\mu^{J,r}\|^2 &= \mathbb{E}\|G_\mu^{J,r} - \nabla g_\mu(z^r) + \nabla g_\mu(z^r)\|^2 \\
&\leq 2\mathbb{E}\|G_\mu^{J,r} - \nabla g_\mu(z^r)\|^2 + 2\mathbb{E}\|\nabla g_\mu(z^r)\|^2 \\
&\overset{\text{(i)}}{\leq} \frac{2\tilde{\sigma}}{J} + 2\mathbb{E}\|\nabla g_\mu(z^r)\|^2 \\
&\overset{\text{(ii)}}{\leq} 2\tilde{\sigma} + 4\mathbb{E}\|\nabla g(z^r)\|^2 + \mu^2\hat{L}^2(Q+3)^3 \\
&\overset{\text{(iii)}}{\leq} 2\tilde{\sigma} + 4K^2 + \mu^2\hat{L}^2(Q+3)^3, \qquad (5.84)
\end{aligned}$$

where (i) is true due to Lemma 1, (ii) comes from the fact that $J \geq 1$, and $\|\nabla g_\mu(z^r)\|^2 \leq 2\|\nabla g(z^r)\|^2 + \frac{\mu^2}{2}\hat{L}^2(Q+3)^3$ [45, Theorem 3.1], and in (iii) we use assumption A1. Therefore, we have proved that there exists a constant $K_2 := 2\tilde{\sigma} + 4K^2 + \mu^2\hat{L}^2(Q+3)^3$ such that $\mathbb{E}\|G_\mu^{J,r}\|^2 \leq K_2$. Finally, plugging this bound into equation (5.83), we get

$$\mathbb{E}[P^{r+1}] \geq -\frac{2}{\rho^2\sigma_{\min}}K_2. \qquad (5.85)$$

Since K_2 is not dependent on T, in order to prove the Lemma we just need to set T-independent lower bound $\underline{P} := -\frac{2}{\rho^2\sigma_{\min}}K_2$.

5.6.4 Proof of Theorem 9

Let us bound the optimality gap given in (5.40) term by term. First we bound the gradient of AL function with respect to variable z in point (z^{r+1}, λ^r) in the following way

$$\begin{aligned}
\|\nabla_z L_\rho(z^{r+1}, \lambda^r)\|^2 &= \|\nabla g(z^{r+1}) + A^\top\lambda^r + \rho A^\top A z^{r+1}\|^2 \\
&\overset{(5.21)}{=} \|\nabla g(z^{r+1}) + A^\top\lambda^{r+1}\|^2 \\
&\overset{(5.74)}{=} \|\nabla g(z^{r+1}) - G_\mu^{J,r} - \rho L^+(z^{r+1} - z^r)\|^2 \\
&\overset{(5.31)}{\leq} 2\|\nabla g(z^{r+1}) - G_\mu^{J,r}\|^2 + 2\rho^2\|L^+(z^{r+1} - z^r)\|^2 \\
&\overset{\text{(i)}}{\leq} 4(\|\nabla g(z^{r+1}) - \nabla g_\mu(z^r)\|^2 + \|\nabla g_\mu(z^r) - G_\mu^{J,r}\|^2) + 2\rho^2\|L^+(z^{r+1} - z^r)\|^2, \qquad (5.86)
\end{aligned}$$

where in (i) we add and subtract $\nabla g_\mu(z^r)$ to $\nabla g(z^{r+1}) - G_\mu^{J,r}$ and apply (5.31). Further, let us take expectation on both sides of (5.86)

$$\begin{aligned}
&\mathbb{E}\|\nabla_z L_\rho(z^{r+1},\lambda^r)\|^2 \\
&\leq 4\mathbb{E}\Big(\|\nabla g(z^{r+1}) - \nabla g_\mu(z^r)\|^2 + \|\nabla g_\mu(z^r) - G_\mu^{J,r}\|^2\Big) + 2\rho^2\mathbb{E}\|L^+(z^{r+1}-z^r)\|^2 \\
&\overset{\text{(i)}}{\leq} 8\mathbb{E}\Big(\|\nabla g(z^{r+1}) - \nabla g_\mu(z^{r+1})\|^2 + \hat{L}^2\|z^{r+1}-z^r\|^2\Big) + \frac{4\tilde{\sigma}^2}{J} + 2\rho^2\mathbb{E}\|L^+(z^{r+1}-z^r)\|^2 \\
&\overset{(5.5)}{\leq} 2\mu^2\hat{L}^2(Q+3)^3 + 8\hat{L}^2\mathbb{E}\|z^{r+1}-z^r\|^2 + \frac{4\tilde{\sigma}^2}{J} + 2\rho^2\mathbb{E}\|L^+(z^{r+1}-z^r)\|^2, \qquad (5.87)
\end{aligned}$$

where in (i) we applied (5.9), (5.31), and the fact that $\nabla g_\mu(z)$ is L_μ-smooth with $L_\mu \leq \hat{L}$. Second, let us bound the expected value of the constraint violation. Utilizing the equation (5.21) we have

$$\|Az^{r+1}\|^2 = \frac{1}{\rho^2}\|\lambda^{r+1}-\lambda^r\|^2.$$

Taking expectation on the above identity, and utilizing the fact that $L_\mu \leq \hat{L}$, and (5.32), we obtain the following

$$\begin{aligned}
\mathbb{E}\|Az^{r+1}\|^2 &= \frac{1}{\rho^2}\mathbb{E}\|\lambda^{r+1}-\lambda^r\|^2 \\
&\leq \frac{9\tilde{\sigma}^2}{J\rho^2\sigma_{\min}} + \frac{6\hat{L}^2}{\rho^2\sigma_{\min}}\mathbb{E}\|z^r - z^{r-1}\|^2 + \frac{3\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2. \qquad (5.88)
\end{aligned}$$

Summing up (5.87) and (5.88), we have the following bound for the optimality gap

$$\begin{aligned}
\Phi^{r+1} &\leq \alpha_1\mathbb{E}\|z^{r+1}-z^r\|^2 + \alpha_2\mathbb{E}\|z^r - z^{r-1}\|^2 + \alpha_3\mathbb{E}\|w^r\|_{L^+}^2 \\
&\quad + \left(\frac{9+4\rho^2\sigma_{\min}}{\rho^2\sigma_{\min}}\right)\frac{\tilde{\sigma}^2}{J} + 2\mu^2\hat{L}^2(Q+3)^3, \qquad (5.89)
\end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3$ are positive constants given by

$$\alpha_1 = 8\hat{L}^2 + 2\rho^2\|L^+\|^2,\ \alpha_2 = \frac{6\hat{L}^2}{\rho^2\sigma_{min}},\ \alpha_3 = \frac{3\|L^+\|}{\sigma_{min}}.$$

Summing both sides of (5.89), we obtain the following

$$\begin{aligned}
\sum_{r=1}^{T}\Phi^{r+1} &\leq \sum_{r=1}^{T-1}(\alpha_1+\alpha_2)\mathbb{E}\|z^{r+1}-z^r\|^2 + \sum_{r=1}^{T}\alpha_3\mathbb{E}\|w^r\|_{L^+}^2 \\
&\quad + \alpha_2\mathbb{E}\|z^1 - z^0\|^2 + \alpha_1\mathbb{E}\|z^{T+1}-z^T\|^2 \\
&\quad + 2T\mu^2\hat{L}^2(Q+3)^3 + T\left(\frac{9+4\rho^2\sigma_{\min}}{\rho^2\sigma_{\min}}\right)\frac{\tilde{\sigma}^2}{J}. \qquad (5.90)
\end{aligned}$$

Applying Lemma 24 and summing both sides of (5.35) over T iterations, we obtain

$$\mathbb{E}\Big[P^1 - P^{T+1}\Big] \geq \sum_{r=1}^{T-1} \frac{c_1 - k}{2} \mathbb{E}\|z^{r+1} - z^r\|^2 + \sum_{r=1}^{T} c_2 \mathbb{E}\|w^r\|_{L^+}^2 + \frac{c_1 - k}{2} \mathbb{E}\|z^{T+1} - z^T\|^2 - \frac{3T\mu^2(Q+3)^3}{8} - \frac{Tc_3\tilde{\sigma}^2}{J}. \tag{5.91}$$

Let us set $\zeta = \frac{\max(\alpha_1+\alpha_2,\alpha_3)}{\min(\frac{c_1-k}{2},c_2)}$. Combining the two inequalities (5.90) and (5.91), and utilizing the fact that $\mathbb{E}[P^{T+1}]$ is lower bounded by $\underline{P}$, we arrive at the following inequality

$$\sum_{r=1}^{T} \Phi^{r+1} \leq \zeta \mathbb{E}[P^1 - \underline{P}] + \alpha_2 \mathbb{E}\|z^1 - z^0\|^2 + T\left(\zeta c_3 + \frac{9 + 4\rho^2\sigma_{\min}}{\rho^2\sigma_{\min}}\right)\frac{\tilde{\sigma}^2}{J} + T\left(\frac{3\zeta}{8} + 2\hat{L}^2\right)\mu^2(Q+3)^3. \tag{5.92}$$

Since u is a uniformly random variable in the set $\{1, 2, \cdots, T\}$ we have

$$\mathbb{E}_u[\Phi^u] = \frac{1}{T}\sum_{r=1}^{T} \Phi^{r+1}. \tag{5.93}$$

Dividing both sides of (5.92) on T and using (5.93) implies the following

$$\mathbb{E}_u[\Phi^u] \leq \frac{\zeta \mathbb{E}[P^1 - \underline{P}] + \alpha_2 \mathbb{E}\|z^1 - z^0\|^2}{T} + \left(\zeta c_3 + \frac{9 + 4\rho^2\sigma_{\min}}{\rho^2\sigma_{\min}}\right)\frac{\tilde{\sigma}^2}{J} + \left(\frac{3\zeta}{8} + 2\hat{L}^2\right)\mu^2(Q+3)^3$$

By setting

$$\gamma_1 = \zeta \mathbb{E}[P^1 - \underline{P}] + \alpha_2 \mathbb{E}\|z^1 - z^0\|^2,$$
$$\gamma_2 = \zeta c_3 + \frac{9 + 4\rho^2\sigma_{\min}}{\rho^2\sigma_{\min}}, \ \gamma_3 = \left(\frac{3\zeta}{8} + 2\hat{L}^2\right)(Q+3)^3, \tag{5.94}$$

we conclude the proof. **Q.E.D.**

5.7 Appendix. Proofs for ZONE-S

This appendix contains the proof of the lemmas in Section 5.3 which are related to ZONE-S algorithm.

In order to facilitate the derivations, in the following let us present some key properties of ZONE-S algorithm. Let us define $r(j) := \max\{t \mid t < r+1, j = i_t\}$ which is the last iteration in

which agent j is picked before iteration $r+1$. From this definition we can see that $r(i_r) = r$. Let us repeat the update equations of ZONE-S algorithm

$$z_{i_r}^{r+1} = x^r - \frac{1}{\alpha_{i_r}\rho_{i_r}}\Big[\lambda_{i_r}^r + \bar{G}_{\mu,i_r}(x^r, \phi^r, \xi^r)\Big]; \tag{5.95}$$

$$\lambda_{i_r}^{r+1} = \lambda_{i_r}^r + \alpha_{i_r}\rho_{i_r}\Big(z_{i_r}^{r+1} - x^r\Big); \tag{5.96}$$

$$\lambda_j^{r+1} = \lambda_j^r, \quad z_j^{r+1} = x^r, \quad \forall\ j \neq i_r. \tag{5.97}$$

Property 1: Compact form for dual update. Combining (5.95), (5.96), and using the definition of $r(j)$ we get

$$\lambda_{i_r}^{r+1} = -\bar{G}_{\mu,i_r}(x^r, \phi^r, \xi^r), \tag{5.98}$$

$$\lambda_j^{r+1} = \lambda_j^r = -\bar{G}_{\mu,j}(x^{r(j)}, \phi^{r(j)}, \xi^{r(j)}), \quad j \neq i_r. \tag{5.99}$$

Using the definition of sequence y^r [$y^0 = x^0,\ y_j^r = y_j^{r-1},$ if $j \neq i_r,$ else $y_{i_r}^r = x^r,\ \forall\ r \geq 1$] we have $y_i^r = x^{r(i)}$ for all $i = 1, 2, \cdots N$. Using this we get the following compact form

$$\lambda_i^{r+1} = -\bar{G}_{\mu,i}(y_i^r, \phi^{r(i)}, \xi^{r(i)}),\ i = 1, \cdots, N. \tag{5.100}$$

Property 2: Compact form for primal update. From (5.97) for $j \neq i_r$ we have

$$\begin{aligned} z_j^{r+1} = x^r &\overset{(5.100)}{=} x^r - \frac{1}{\alpha_j\rho_j}[\lambda_j^{r+1} + \bar{G}_{\mu,j}(y_j^r, \phi^{r(j)}, \xi^{r(j)})] \\ &\overset{(5.97)}{=} x^r - \frac{1}{\alpha_j\rho_j}[\lambda_j^r + \bar{G}_{\mu,i}(y_j^r, \phi^{r(j)}, \xi^{r(j)})]. \end{aligned} \tag{5.101}$$

Considering (5.95), and (5.101) we can express the update equation for z in ZONE-S algorithms in the following compact form

$$z_i^{r+1} = x^r - \frac{1}{\alpha_i\rho_i}\Big[\lambda_i^r + \bar{G}_{\mu,i}(y_i^r, \phi^{r(i)}, \xi^{r(i)})\Big],\ i = 1, \cdots, N. \tag{5.102}$$

Property 3: Bound the distance between update direction and the gradient direction. Let us define

$$u^{r+1} := \beta\Big(\sum_{i=1}^N \rho_i z_i^{r+1} + \sum_{i=1}^N \lambda_i^r\Big), \tag{5.103}$$

where we set $\beta := 1/\sum_{i=1}^{N} \rho_i$. Using (5.103), it is easy to check that x-update (5.52) is equivalent to solving the following problem

$$\begin{aligned} x^{r+1} &= \arg\min_x \ \frac{1}{2\beta}\|x - u^{r+1}\|^2 + h(x) \\ &= \text{prox}_h^{1/\beta}(u^{r+1}). \end{aligned} \tag{5.104}$$

The optimality condition for this problem is given by

$$x^{r+1} - u^{r+1} + \beta\eta^{r+1} = 0, \tag{5.105}$$

where $\eta^{r+1} \in \partial h(x^{r+1})$ is a subgradient of h at x^{r+1}. [When there is no confusion we use the shorthand notation $\bar{G}_{\mu,i}^r$ to denote $\bar{G}_{\mu,i}(x^r, \phi^r, \xi^r)$]

$$\begin{aligned} u^{r+1} &= \beta\left(\sum_{i=1}^{N} \rho_i z_i^{r+1} + \sum_{i=1}^{N} \lambda_i^r\right) \\ &\overset{(5.97)}{=} \beta\left(\sum_{i=1}^{N} \rho_i x^r - \rho_{i_r}(x^r - z_{i_r}^{r+1}) + \sum_{i=1}^{N} \lambda_i^r\right) \\ &\overset{(5.100),(5.95)}{=} x^r - \frac{\beta}{\alpha_{i_r}}\left[\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})\right] \\ &\quad - \beta\sum_{i=1}^{N} \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)}). \end{aligned} \tag{5.106}$$

Let us further define

$$\begin{aligned} v_{i_r}^r &:= \sum_{i=1}^{N} \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)}) \\ &\quad + \frac{1}{\alpha_{i_r}}\left[\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})\right]. \end{aligned} \tag{5.107}$$

We conclude that

$$u^{r+1} = x^r - \beta v_{i_r}^r. \tag{5.108}$$

Plugging (5.108) into (5.105) we obtain

$$x^{r+1} = x^r - \beta(v_{i_r}^r + \eta^{r+1}). \tag{5.109}$$

Now let us bound $\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\|$. Using the definition of $v_{i_r}^r$ we have

$$\begin{aligned}
&\left\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\right\|^2 \\
&\overset{(5.107)}{=} \left\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - \sum_{i=1}^N \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right. \\
&\quad \left. - \frac{1}{\alpha_{i_r}}\left[\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})\right]\right\|^2. \qquad (5.110)
\end{aligned}$$

Let us set $\mathcal{J}^r := \{i_r, \phi^r, \xi^r\}$. Setting $\alpha_i = p_i$ and taking conditional expectation on both sides, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\left\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\right\|^2 |\mathcal{F}^r\right] \\
&= \mathbb{E}_{\mathcal{J}^r}\left[\left\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - \sum_{i=1}^N \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right.\right. \\
&\quad \left.\left. - \frac{1}{\alpha_{i_r}}\left[\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})\right]\right\|^2 \mathcal{F}^r\right] \\
&\overset{\text{(i)}}{\leq} \mathbb{E}_{\mathcal{J}^r}\left[\left\|\frac{\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})}{\alpha_{i_r}}\right\|^2 \mathcal{F}^r\right],
\end{aligned}$$

where (i) is true because $\mathbb{E}[\|x - \mathbb{E}[x]\|^2] = \mathbb{E}[\|x\|^2] - \|\mathbb{E}[x]\|^2 \leq \mathbb{E}[\|x\|^2]$ and the following identity

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\frac{1}{\alpha_{i_r}}\left[\bar{G}_{\mu,i_r}^r - \bar{G}_{\mu,i_r}(y_{i_r}^{r-1}, \phi^{(r-1)(i_r)}, \xi^{(r-1)(i_r)})\right] \mid \mathcal{F}^r\right] \\
&= \sum_{i=1}^N \nabla f_{\mu,i}(x^r) - \sum_{i=1}^N \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)}).
\end{aligned}$$

Now if we take expectation with respect to i_r, (given $\mathcal{F}^r$)

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\left\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\right\|^2 |\mathcal{F}^r\right] \\
&\leq \sum_{i=1}^N \frac{1}{p_i}\mathbb{E}_{\phi^r,\xi^r}\left[\left\|\bar{G}_{\mu,i}^r - \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right\|^2 |\mathcal{F}^r\right] \\
&= \sum_{i=1}^N \frac{1}{p_i}\mathbb{E}_{\phi^r,\xi^r}\left[\left\|\bar{G}_{\mu,i}^r - \nabla f_{\mu,i}(x^r) + \nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right.\right. \\
&\quad \left.\left. + \nabla f_{\mu,i}(y_i^{r-1}) - \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right\|^2 |\mathcal{F}^r\right].
\end{aligned}$$

Then utilizing (5.31) and (5.9), we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\left\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\right\|^2|\mathcal{F}^r\right]\\
&\leq 3\sum_{i=1}^{N}\frac{1}{p_i}\Bigg(\frac{\tilde{\sigma}^2}{J}+\left\|\nabla f_{\mu,i}(x^r)-\nabla f_{\mu,i}(y_i^{r-1})\right\|^2\\
&\quad+\left\|\nabla f_{\mu,i}(y_i^{r-1})-\bar{G}_{\mu,i}(y_i^{r-1},\phi^{(r-1)(i)},\xi^{(r-1)(i)})\right\|^2\Bigg).
\end{aligned}\tag{5.111}$$

Using the definition of $\tilde{p}=\sum_{i=1}^{N}\frac{1}{p_i}$, overall we have the following

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\left\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\right\|^2\mid\mathcal{F}^r\right]\\
&\leq\frac{3\tilde{p}\tilde{\sigma}^2}{J}+\sum_{i=1}^{N}\frac{3}{p_i}\Bigg(\left\|\nabla f_{\mu,i}(x^r)-\nabla f_{\mu,i}(y_i^{r-1})\right\|^2\\
&\quad+\left\|\nabla f_{\mu,i}(y_i^{r-1})-\bar{G}_{\mu,i}(y_i^{r-1},\phi^{(r-1)(i)},\xi^{(r-1)(i)})\right\|^2\Bigg).
\end{aligned}\tag{5.112}$$

Using the tower property of conditional expectation we have

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\right\|^2&=\mathbb{E}_{\mathcal{F}^r,\mathcal{J}^r}\left\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\right\|^2\\
&=\mathbb{E}_{\mathcal{F}^r}\left[\mathbb{E}_{\mathcal{J}^r}\left[\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\|^2\mid\mathcal{F}^r\right]\right].
\end{aligned}\tag{5.113}$$

Now let us break the filtration as $\mathcal{F}^r=\mathcal{F}_1^r\cup\mathcal{F}_2^r$ where $\mathcal{F}_1^r:=\{i_t\}_{t=1}^{r-1}$, and $\mathcal{F}_2^r:=\{\phi^t,\xi^t\}_{t=1}^{r-1}$. Using these notations we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}^r}\left\|\nabla f_{\mu,i}(y_i^{r-1})-\bar{G}_{\mu,i}(y_i^{r-1},\phi^{(r-1)(i)},\xi^{(r-1)(i)})\right\|^2\\
&=\mathbb{E}_{\mathcal{F}_1^r}\left[\mathbb{E}_{\mathcal{F}_2^r}\left\|\nabla f_{\mu,i}(y_i^{r-1})-\bar{G}_{\mu,i}(y_i^{r-1},\phi^{(r-1)(i)},\xi^{(r-1)(i)})\right\|^2\mid\mathcal{F}_1^r\right]\\
&\overset{(5.9)}{\leq}\frac{\tilde{\sigma}^2}{J}.
\end{aligned}\tag{5.114}$$

Combining (5.112), (5.113), (5.114), we obtain

$$\mathbb{E}\left\|\sum_{i=1}^{N}\nabla f_{\mu,i}(x^r)-v_{i_r}^r\right\|^2\leq\sum_{i=1}^{N}\frac{3}{p_i}\mathbb{E}\left\|\nabla f_{\mu,i}(x^r)-\nabla f_{\mu,i}(y_i^{r-1})\right\|^2+\frac{6\tilde{p}\tilde{\sigma}^2}{J}.\tag{5.115}$$

5.7.1 Proof of Lemma 26

By assumption $\alpha_i = p_i$, according to the definition of potential function $\tilde{Q}^r$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}[\tilde{Q}^{r+1} - \tilde{Q}^r \mid \mathcal{F}^r] \\
&= \mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \Big(f_{\mu,i}(x^{r+1}) - f_{\mu,i}(x^r)\Big) + h(x^{r+1}) - h(x^r) \mid \mathcal{F}^r\right] \\
&\quad + \mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \frac{4}{p_i\rho_i}\left\|\nabla f_{\mu,i}(x^{r+1}) - \nabla f_{\mu,i}(y_i^r)\right\|^2 - \frac{4}{p_i\rho_i}\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2 \mid \mathcal{F}^r\right]. \quad (5.116)
\end{aligned}$$

The proof consists of the following steps:

Step 1). We bound the first term in (5.116) as follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \Big(f_{\mu,i}(x^{r+1}) - f_{\mu,i}(x^r)\Big) + h(x^{r+1}) - h(x^r) \mid \mathcal{F}^r\right] \\
&\overset{\text{(i)}}{\le} \mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \langle \nabla f_{\mu,i}(x^r), x^{r+1} - x^r\rangle + \langle \eta^{r+1}, x^{r+1} - x^r\rangle + \frac{\sum_{i=1}^N L_{\mu,i}}{2}\|x^{r+1} - x^r\|^2 \mid \mathcal{F}^r\right] \\
&= \mathbb{E}_{\mathcal{J}^r}\left[\Big\langle \sum_{i=1}^N \nabla f_{\mu,i}(x^r) + \eta^{r+1} + \frac{1}{\beta}(x^{r+1} - x^r), x^{r+1} - x^r\Big\rangle \mid \mathcal{F}^r\right] \\
&\quad - \left(\frac{1}{\beta} - \frac{\sum_{i=1}^N L_{\mu,i}}{2}\right)\mathbb{E}_{\mathcal{J}^r}\left[\|x^{r+1} - x^r\|^2 \mid \mathcal{F}^r\right].
\end{aligned}$$

Then from (5.109) we further have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \Big(f_{\mu,i}(x^{r+1}) - f_{\mu,i}(x^r)\Big) + h(x^{r+1}) - h(x^r) \mid \mathcal{F}^r\right] \\
&\le \mathbb{E}_{\mathcal{J}^r}\left[\Big\langle \sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r, x^{r+1} - x^r\Big\rangle \mid \mathcal{F}^r\right] \\
&\quad - \left(\frac{1}{\beta} - \frac{\sum_{i=1}^N L_{\mu,i}}{2}\right)\mathbb{E}_{\mathcal{J}^r}\left[\|x^{r+1} - x^r\|^2 \mid \mathcal{F}^r\right] \\
&\overset{\text{(ii)}}{\le} \sum_{i=1}^N \frac{3\beta}{2p_i}\left(\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2\right. \\
&\quad \left. + \left\|\nabla f_{\mu,i}(y_i^{r-1}) - \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right\|^2\right) \\
&\quad - \left(\frac{1}{2\beta} - \frac{\sum_{i=1}^N L_{\mu,i}}{2}\right)\mathbb{E}_{\mathcal{J}^r}\left[\|x^{r+1} - x^r\|^2 \mid \mathcal{F}^r\right] + \frac{3\tilde{p}\beta\tilde{\sigma}^2}{2J}, \quad (5.117)
\end{aligned}$$

where in (i) we have used the Lipschitz continuity of the gradients of $f_{\mu,i}$, and the convexity of function h; in (ii) we utilize (5.30) with $\epsilon = \frac{1}{\beta}$, and (5.112).

Step 2). In this step we bound the second term in equation (5.116) as follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\left\|\nabla f_{\mu,i}(x^{r+1}) - \nabla f_{\mu,i}(y_i^r)\right\|^2 \mid \mathcal{F}^r\right] \\
&\overset{\text{(i)}}{\leq} (1+\epsilon_i)\mathbb{E}_{\mathcal{J}^r}\left[\left\|\nabla f_{\mu,i}(x^{r+1}) - \nabla f_{\mu,i}(x^r)\right\|^2 \mid \mathcal{F}^r\right] \\
&+ \left(1+\frac{1}{\epsilon_i}\right)\mathbb{E}_{\mathcal{J}^r}\left[\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^r)\right\|^2 \mid \mathcal{F}^r\right] \qquad (5.118) \\
&\overset{\text{(ii)}}{=} (1+\epsilon_i)\mathbb{E}_{\mathcal{J}^r}\left[\left\|\nabla f_{\mu,i}(x^{r+1}) - \nabla f_{\mu,i}(x^r)\right\|^2 \mid \mathcal{F}^r\right] \\
&+ (1-p_i)\left(1+\frac{1}{\epsilon_i}\right)\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2, \qquad (5.119)
\end{aligned}$$

where in (i) we first apply (5.30). Note that when $\mathcal{F}^r$ is given the randomness of the first and second term in (5.118) come from x^{r+1} and y_i^r respectively. Therefore, equality (ii) is true because $y_i^r = x^r$, with probability p_i, and $y_i^r = y_i^{r-1}$, with probability $1-p_i$. Setting $\epsilon_i = \frac{2}{p_i}$, the second part of (5.116) can be bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}\left[\sum_{i=1}^N \frac{4}{p_i\rho_i}\left\|\nabla f_{\mu,i}(x^{r+1}) - \nabla f_{\mu,i}(y_i^r)\right\|^2 - \frac{4}{p_i\rho_i}\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2 \mid \mathcal{F}^r\right] \\
&\leq \sum_{i=1}^N \frac{4L_{\mu,i}^2(2+p_i)}{p_i^2\rho_i}\mathbb{E}_{\mathcal{J}^r}\|x^{r+1}-x^r\|^2 - \sum_{i=1}^N \frac{4(1+p_i)}{2\rho_i}\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2. \qquad (5.120)
\end{aligned}$$

Step 3). In this step we combine the results from the previous steps to obtain the desired descent estimate. Combining (5.117) and (5.120) eventually we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{J}^r}[\tilde{Q}^{r+1} - \tilde{Q}^r \mid \mathcal{F}^r] \\
&\leq \sum_{i=1}^N \left(\frac{3\beta}{2p_i} - \frac{4(1+p_i)}{2\rho_i}\right)\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2 \\
&+ \sum_{i=1}^N \left(\frac{4L_{\mu,i}^2(2+p_i)}{p_i^2\rho_i} + \frac{L_{\mu,i}}{2} - \frac{\rho_i}{2}\right)\mathbb{E}_{\mathcal{J}^r}\left[\|x^{r+1}-x^r\|^2 \mid \mathcal{F}^r\right] \\
&+ \sum_{i=1}^N \frac{3\beta}{2p_i}\left\|\nabla f_{\mu,i}(y_i^{r-1}) - \bar{G}_{\mu,i}(y_i^{r-1}, \phi^{(r-1)(i)}, \xi^{(r-1)(i)})\right\|^2 + \frac{3\tilde{p}\beta\tilde{\sigma}^2}{2J}. \qquad (5.121)
\end{aligned}$$

Using the properties of conditional expectation we have

$$\mathbb{E}[\tilde{Q}^{r+1} - \tilde{Q}^r] = \mathbb{E}_{\mathcal{F}^r}\Big[\mathbb{E}_{\mathcal{J}^r}[\tilde{Q}^{r+1} - \tilde{Q}^r \mid \mathcal{F}^r]\Big]. \tag{5.122}$$

Plugging (5.121) in this relationship and utilizing (5.114), and the definition of $\beta := 1/\sum_{i=1}^N \rho_i$, yield

$$\begin{aligned}
&\mathbb{E}[\tilde{Q}^{r+1} - \tilde{Q}^r] \\
&\leq \sum_{i=1}^N \left(\frac{3\beta}{2p_i} - \frac{4(1+p_i)}{2\rho_i}\right)\mathbb{E}\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2 \\
&\quad + \sum_{i=1}^N \left(\frac{4L_{\mu,i}^2(2+p_i)}{p_i^2\rho_i} + \frac{L_{\mu,i}}{2} - \frac{\rho_i}{2}\right)\mathbb{E}\Big[\|x^{r+1} - x^r\|^2\Big] + \frac{3\tilde{p}\beta\tilde{\sigma}^2}{J}.
\end{aligned} \tag{5.123}$$

Let us define $\{\tilde{c}_i\}$ and $\hat{c}$ as following

$$\tilde{c}_i = \frac{3\beta}{2p_i} - \frac{4(1+p_i)}{2\rho_i}, \quad \hat{c} = \sum_{i=1}^N \left(\frac{4L_{\mu,i}^2(2+p_i)}{p_i^2\rho_i} + \frac{L_{\mu,i}}{2} - \frac{\rho_i}{2}\right).$$

In order to prove the lemma it remains to prove that $\tilde{c}_i < -\frac{1}{2\rho_i}\ \forall\ i$, and $\hat{c} < -\sum_{i=1}^N \frac{\rho_i}{100}$. If we set $p_i = \frac{\rho_i}{\sum_{i=1}^N \rho_i}$, then we have the following

$$\tilde{c}_i = \frac{3}{2\rho_i} - \frac{4(1+p_i)}{2\rho_i} \leq \frac{3}{2\rho_i} - \frac{4}{2\rho_i} = -\frac{1}{2\rho_i}.$$

To show that $\hat{c} \leq -\sum_{i=1}^N \frac{\rho_i}{100}$, it is sufficient to have

$$\frac{4L_{\mu,i}^2(2+p_i)}{p_i^2\rho_i} + \frac{L_{\mu,i}}{2} - \frac{\rho_i}{2} \leq -\frac{\rho_i}{100}. \tag{5.124}$$

It is easy to check that this inequality holds true for $\rho_i \geq \frac{5.5L_{\mu,i}}{p_i}$. The lemma is proved. **Q.E.D.**

5.7.2 Proof of Theorem 10

Here we only prove the first part of the theorem. Similar steps can be followed to prove the second part. First let us define the smoothed version of optimality gap as follows

$$\Psi_\mu^r = \frac{1}{\beta^2}\mathbb{E}\left\|x^r - \text{prox}_h^{1/\beta}[x^r - \beta\nabla f_\mu(x^r)]\right\|^2. \tag{5.125}$$

We bound the gap in the following way

$$\begin{aligned}
&\frac{1}{\beta^2}\left\|x^r - \text{prox}_h^{1/\beta}[x^r - \beta\nabla f_\mu(x^r)]\right\|^2 \\
&\overset{\text{(i)}}{=} \frac{1}{\beta^2}\left\|x^r - x^{r+1} + \text{prox}_h^{1/\beta}(u^{r+1}) - \text{prox}_h^{1/\beta}[x^r - \beta\nabla f_\mu(x^r)]\right\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{2}{\beta^2}\|x^{r+1} - x^r\|^2 + \frac{2}{\beta^2}\|\beta\nabla f_\mu(x^r) + u^{r+1} - x^r\|^2 \\
&\overset{(5.108)}{=} \frac{2}{\beta^2}\|x^{r+1} - x^r\|^2 + 2\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\|^2, \qquad (5.126)
\end{aligned}$$

where (i) is true due to (5.104); (ii) is true due to the nonexpansivness of the prox operator, and equation (5.31). Taking expectation on both sides yields

$$\begin{aligned}
\Psi_\mu^r &\leq \frac{2}{\beta^2}\mathbb{E}\|x^{r+1} - x^r\|^2 + 2\mathbb{E}\|\sum_{i=1}^N \nabla f_{\mu,i}(x^r) - v_{i_r}^r\|^2 \\
&\overset{\text{(i)}}{\leq} \frac{2}{\beta^2}\mathbb{E}\|x^{r+1} - x^r\|^2 + \frac{6}{\beta}\sum_{i=1}^N \frac{1}{\rho_i}\mathbb{E}\left\|\nabla f_{\mu,i}(x^r) - \nabla f_{\mu,i}(y_i^{r-1})\right\|^2 + \frac{12\tilde{p}\tilde{\sigma}^2}{J} \\
&\overset{(5.55)}{\leq} \frac{200}{\beta}\mathbb{E}[\tilde{Q}^r - \tilde{Q}^{r+1}] + \frac{612\tilde{p}\tilde{\sigma}^2}{J} \\
&\overset{\text{(ii)}}{=} 1100\left(\sum_{i=1}^N \sqrt{L_{\mu,i}}\right)^2 \mathbb{E}[\tilde{Q}^r - \tilde{Q}^{r+1}] + \frac{612\tilde{p}\tilde{\sigma}^2}{J}, \qquad (5.127)
\end{aligned}$$

where in (i) we utilize (5.115). To get (ii) let us pick $\rho_i = \frac{5.5L_{\mu,i}}{p_i}$, therefore we have $\rho_i = 5.5L_{\mu,i}\frac{\sum_{i=1}^N \rho_i}{\rho_i}$, which leads to $\rho_i = \sqrt{5.5L_{\mu,i}\sum_{j=1}^N \rho_j} = \sqrt{5.5L_{\mu,i}}\sqrt{\sum_{j=1}^N \rho_j}$. Summing both sides over $i = 1, 2, \cdots N$, and simplifying the result we get

$$\sqrt{\sum_{i=1}^N \rho_i} = \sum_{i=1}^N \sqrt{5.5L_{\mu,i}}.$$

Finally, squaring both sides and set $\beta := 1/\sum_{i=1}^N \rho_i$ we reach $\frac{1}{\beta} = 5.5(\sum_{i=1}^N \sqrt{L_{\mu,i}})^2$. Let us sum both sides of (5.127) over T iterations, use telescopic property, and divide both sides by T, we obtain

$$\frac{1}{T}\sum_{r=1}^T \Psi_\mu^r \leq 1100\left(\sum_{i=1}^N \sqrt{L_{\mu,i}}\right)^2 \frac{\mathbb{E}[\tilde{Q}^1 - \tilde{Q}^{T+1}]}{T} + \frac{612\tilde{p}\tilde{\sigma}^2}{J}.$$

Since u is uniformly random number in $\{1, 2, \cdots, T\}$, we finally have

$$\mathbb{E}_u[\Psi_\mu^u] \leq 1100\left(\sum_{i=1}^N \sqrt{L_{\mu,i}}\right)^2 \frac{\mathbb{E}[\tilde{Q}^1 - \tilde{Q}^{T+1}]}{T} + \frac{612\tilde{p}\tilde{\sigma}^2}{J}. \qquad (5.128)$$

Now let us bound the gap Ψ^r. Using the definition of Ψ^r we have

$$\begin{aligned}\Psi^r &= \frac{1}{\beta^2}\mathbb{E}\Big[\|x^r - \mathrm{prox}_h^{1/\beta}[x^r - \beta\nabla f(x^r)]\Big] \\ &= \frac{1}{\beta^2}\mathbb{E}\Big[\|x^r - \mathrm{prox}_h^{1/\beta}[x^r - \beta\nabla f(x^r)] - \mathrm{prox}_h^{1/\beta}[x^r - \beta\nabla f_\mu(x^r)] + \mathrm{prox}_h^{1/\beta}[x^r - \beta\nabla f_\mu(x^r)]\|^2\Big] \\ &\overset{\text{(i)}}{\leq} 2\Psi_\mu^r + \frac{\mu^2L^2(M+3)^3}{2},\end{aligned}$$

where in (i) we use (5.31); the nonexpansiveness of the prox operator; and inequality (5.5). Next because r is a uniformly random number picked form $\{1, 2, \cdots, T\}$ we have

$$\begin{aligned}\mathbb{E}_u[\Psi^u] &\leq 2\mathbb{E}_u[\Psi_\mu^u] + \frac{\mu^2L^2(M+3)^3}{2} \\ &\overset{(5.128)}{\leq} 2200\Big(\sum_{i=1}^N \sqrt{L_{\mu,i}}\Big)^2\frac{\mathbb{E}[\tilde{Q}^1 - \tilde{Q}^{T+1}]}{T} + \frac{\mu^2L^2(M+3)^3}{2} + \frac{1024\tilde{p}\tilde{\sigma}^2}{J}. \qquad (5.129)\end{aligned}$$

The proof is complete. **Q.E.D.**

Table 5.1: Comparison results for ZONE-M and RGF

	opt-gap			**cons-error**		
N	ZONE-M(C)	ZONE-M(I)	RGF	ZONE-M(C)	ZONE-M(I)	RGF
10	6.8E-6	**8.8E-6**	1.7E-4	2.5E-5	**2.0E-5**	0.002
20	4.2E-5	**2.2E-5**	5.3E-3	3.1E-5	**2.2E-5**	0.003
40	7.0E-5	**3.0E-5**	1.8E-3	3.8E-4	**2.8E-4**	0.017
80	5.7E-4	**7.5E-5**	0.014	5.4E-4	**3.0E-4**	0.09

Bibliography

[1] (2011). Decentralized multi-agent optimization via dual decomposition. {*IFAC*} *Proceedings Volumes*, 44(1):11245 – 11251.

[2] Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40.

[3] Allen-Zhu, Z. and Hazan, E. (2016a). Variance Reduction for Faster Non-Convex Optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML.

[4] Allen-Zhu, Z. and Hazan, E. (2016b). Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 699–707.

[5] Ames, B. and Hong, M. (2016). Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis. *Computational Optimization and Applications*, 64(3):725–754.

[6] Andreani, R., Haeser, G., and Martinez, J. M. (2011). On sequential optimality conditions for smooth constrained optimization. *Optimization*, 60(5):627–641.

[7] Antoniadis, A., Gijbels, I., and Nikolova, M. (2009). Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615.

[8] Asteris, M., Papailiopoulos, D., and Dimakis, A. (2014). Nonnegative sparse pca with provable guarantees. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, pages 1728–1736.

[9] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457.

[10] Aybat, N. and Hamedani, E. Y. (2016a). A primal-dual method for conic constrained distributed optimization problems. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 5049–5057.

[11] Aybat, N.-S. and Hamedani, E.-Y. (2016b). A primal-dual method for conic constrained distributed optimization problems. *Advances in Neural Information Processing Systems*.

[12] Aybat, N. S., Wang, Z., Lin, T., and Ma, S. (2015). Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. *arXiv preprint arXiv:1512.08122*.

[13] Bertsekas, D. (2000). Incremental gradient, subgradient, and proximal methods f or convex optimization: A survey. LIDS Report 2848.

[14] Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Method*. Academic Press.

[15] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[16] Bertsekas, D. P. and Tsitsiklis, J. N. (1997). *Parallel and Distributed Computation: Numerical Methods, 2nd ed*. Athena Scientific, Belmont, MA.

[17] Bianchi, P. and Jakubowicz, J. (2013). Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Transactions on Automatic Control*, 58(2):391–405.

[18] Bjornson, E. and Jorswieck, E. (2013). Optimal resource allocation in coordinated multi-cell systems. *Foundations and Trends in Communications and Information Theory*, 9.

[19] Blatt, D., Hero, A. O., and Gauchman, H. (2007). A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51.

[20] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

[21] Burachik, R. S., Kaya, C. Y., and Mammadov, M. (2008). An inexact modified subgradient algorithm for nonconvex optimization. *Computational Optimization and Applications*, 45(1):1–24.

[22] Candès, E., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905.

[23] Cevher, V., Becker, S., and Schmidt, M. (2014). Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43.

[24] Chang, T.-H., Hong, M., and Wang, X. (2015a). Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497.

[25] Chang, T.-H., Hong, M., and Wang, X. (2015b). Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497.

[26] Chen, J. and Sayed, A. H. (2012). Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305.

[27] Conn, A., Scheinberg, K., and Vicente, L. (2009). *Introduction to derivative-free optimization*. MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics/Mathematical Programming Society, Philadelphia.

[28] Cressie, N. (2015). *Statistics for spatial data*. ohn Wiley & Sons.

[29] Curtis, F. E., Gould, N. I. M., Jiang, H., and Robinson, D. P. (2016). Adaptive augmented lagrangian methods: algorithms and practical numerical experience. *Optimization Methods and Software*, 31(1):157–186.

[30] d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448.

[31] Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *The Proceeding of NIPS*.

[32] Deng, W. and Yin, W. (2015). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, pages 1–28.

[33] Duchi, J., Jordan, M., Wainwright, M., and Wibisono, A. (2015). Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806.

[34] Duchi, J. C., Agarwal, A., and Wainwright, M. J. (2012). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606.

[35] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[36] Fernndez, D. and Solodov, M. V. (2012). Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition. *SIAM Journal on Optimization*, 22(2):384–407.

[37] Flaxman, A., Kalai, A., and McMahan, H. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics.

[38] Forero, P. A., Cano, A., and Giannakis, G. B. (2010). Consensus-based distributed support vector machines. *Journal of Machine Learning Research*, 11(May):1663–1707.

[39] Forero, P. A., Cano, A., and Giannakis, G. B. (2011a). Distributed clustering using wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):707–724.

[40] Forero, P. A., Cano, A., and Giannakis, G. B. (2011b). Distributed clustering using wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):707–724.

[41] Fu, M. C. (2015). *Stochastic gradient estimation, Handbook of simulation optimization.* Springer.

[42] Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2:17–40.

[43] Gao, X., Jiang, B., and Zhang, S. (2014). On the information-adaptive variants of the admm: An iteration complexity perspective. Preprint.

[44] Ghadimi, S. and Lan, G. (2013a). Stochastic first- and zeroth-order methods for nonconvx stochastic programming. *SIAM Journal on Optimizatnoi*, 23(4):2341–2368.

[45] Ghadimi, S. and Lan, G. (2013b). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

[46] Giannakis, G. B., Ling, Q., Mateos, G., Schizas, I. D., and Zhu, H. (2015). Decentralized learning for wireless communications and networking. In *Splitting Methods in Communication and Imaging*. Springer New York.

[47] Giannakis, G. B., Ling, Q., Mateos, G., Schizas, I. D., and Zhu, H. (2016). *Decentralized Learning for Wireless Communications and Networking*, pages 461–497. Springer International Publishing, Cham.

[48] Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par elements finis d'ordre un,et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares. *Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle*, 9:41–76.

[49] Gu, Q., Z. Wang, Z., and Liu, H. (2014). Sparse pca with oracle property. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 1529–1537.

[50] Guan, Y. and Dy, J. G. (2009). Sparse probabilistic principal component analysis. In *Proceedings of 12th International Conference on Artifcial Intelligence and Statistics*, pages 185–192.

[51] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P. (2015). On the convergence rate of incremental aggregated gradient algorithms. *arXiv preprint arXiv:1506.02081*.

[52] H.-T. Wai, T.-H. C. and Scaglione, A. (2015). A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning. In *the Proceedings of the IEEE ICASSP*.

[53] Haeser, G. and Melo, V. (2013). On sequential optimality conditions for smooth constrained optimization. Preprint.

[54] Hajinezhad, D., Chang, T. H., Wang, X., Shi, Q., and Hong, M. (2016a). Nonnegative matrix factorization using admm: Algorithm and convergence analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4742–4746.

[55] Hajinezhad, D. and Hong, M. (2015a). Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE.

[56] Hajinezhad, D. and Hong, M. (2015b). Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*.

[57] Hajinezhad, D., Hong, M., Zhao, T., and Wang, Z. (2016b). Nestt: A nonconvex primal-dual splitting method for distributed and stochastic optimization. In *Advances in Neural Information Processing Systems 29*, pages 3215–3223.

[58] Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of Optimization and Application*, (4):303 – 320.

[59] Hong, M. (2016). Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv preprint arXiv:1604.00543*.

[60] Hong, M. and Chang, T. H. (2017). Stochastic proximal gradient consensus over random networks. *IEEE Transactions on Signal Processing*, 65(11):2933–2948.

[61] Hong, M., Hajinezhad, D., and Zhao, M.-M. (2017). Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1529–1538.

[62] Hong, M. and Luo, Z.-Q. (2016). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming Series A*. to appear, available at arXiv:1208.3922.

[63] Hong, M., Luo, Z.-Q., and Razaviyayn, M. (2014). Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. technical report, University of Minnesota.

[64] Hong, M., Luo, Z.-Q., and Razaviyayn, M. (2016a). Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364.

[65] Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. (2016b). A unified algorithmic framework for block-structured optimization involving big data. *IEEE Signal Processing Magazine*, 33(1):57–77.

[66] Huang, K., Fu, X., and Sidiropoulos, N. D. (2016). Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pages 1786–1794.

[67] Huo, Z. and Huang, H. (2016). Asynchronous stochastic gradient descent with variance reduction for non-convex optimization. *arXiv preprint arXiv:1604.03584*.

[68] I. Schizas, G. M. and Giannakis, G. (2009). Distributed LMS for consensus-based in-network adaptive processing,. *IEEE Transactions on Signal Processing*, 57(6):2365–2382.

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

[70] J.L. Fleiss, B. Levin, M. C. P. J. F. (2003). *Statistical Methods for Rates & Proportions*. Wiley.

[71] Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *the Proceedings of the Neural Information Processing (NIPS)*.

[72] Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York.

[73] J.Wright, S. (1990). Implementing proximal point methods for linear programming. *Journal of Optimization Theory and Applications*, 65(3):531–554.

[74] K. J. Arrow, L. H. and Uzawa, H. (1958). *Studies in Linear and Non-linear Programming*. Stanford University Press.

[75] Koppel, A., Sadler, B. M., and Ribeiro, A. (2016). Proximity without consensus in online multi-agent optimization. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3726–3730.

[76] L. Deng, D. Y. (2014). Deep learning: Methods and applications. Technical report.

[77] Lan, G. and Monteiro, R. D. C. (2015). Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1):511–547.

[78] Lan, G. and Zhou, Y. (2017). An optimal randomized incremental gradient method. *Mathematical Programming*.

[79] Li, G. and Pong, T. K. (2015). Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460.

[80] Li, M., Andersen, D. G., and Smola, A. (2013). Distributed delayed proximal gradient methods. In *NIPS Workshop on Optimization for Machine Learning*.

[81] Lian, X., Zhang, H., Hsieh, C.-J., Huang, Y., and Liu, J. (2016). A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *arXiv preprint arXiv:1606.00498*.

[82] Liao, W.-C., Hong, M., Farmanbar, H., and Luo, Z.-Q. (2015a). Semi-asynchronous routing for large-scale hierarchical networks. In *The Proceedings of IEEE ICASSP*.

[83] Liao, W. C., Hong, M., Farmanbar, H., and Luo, Z. Q. (2015b). Semi-asynchronous routing for large scale hierarchical networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2894–2898.

[84] Liavas, A. P. and Sidiropoulos, N. D. (2015). Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(20):5450–5463.

[85] Ling, Q., Shi, W., Wu, G., and Ribeiro, A. (2015). DLM: Decentralized linearized alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064.

[86] Ling, Q., Xu, Y., Yin, W., and Wen, Z. (2012). Decentralized low-rank matrix completion. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2925–2928.

[87] Liu, Y.-F., Liu, X., and Ma, S. (2016). On the non-ergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming. *arXiv preprint arXiv:1603.05738*.

[88] Lobel, I. and Ozdaglar, A. (2011). Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306.

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

[90] Loh, P.-L. and Wainwright, M. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

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

[92] Lorenzo, P. D. and Scutari, G. (2016b). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136.

[93] Luo, Z.-Q. and Tseng, P. (1992a). Error bounds and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem. *SIAM Journal on Optimization*, pages 43–54.

[94] Luo, Z.-Q. and Tseng, P. (1992b). On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425.

[95] Mateos, G., Bazerque, J. A., and Giannakis, G. B. (2010a). Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276.

[96] Mateos, G., Bazerque, J. A., and Giannakis, G. B. (2010b). Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276.

[97] Max L.N. Goncalves, J. G. M. and Monteiro, R. D. (2017). Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. Preprint, available at: arXiv:1702.01850.

[98] Mokhtari, A., Shi, W., Ling, Q., and Ribeiro, A. (2016). Dqm: Decentralized quadratically approximated alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 64(19):5158–5173.

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

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

[101] Nedić, A. and Ozdaglar, A. (2009). Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228.

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

[103] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

[104] Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*. Springer.

[105] Nesterov, Y. and Spokoiny, V. (2011). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40.

[106] Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer.

[107] Powell, M. M. D. (1969). An efficient method for nonlinear constraints in minimization problems. In *Optimization*. Academic Press.

[108] Rahmani, M. and Atia, G. (2015). A decentralized approach to robust subspace recovery. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 802–807. IEEE.

[109] Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153.

[110] Razaviyayn, M., Hong, M., Luo, Z.-Q., and Pang, J. S. (2014). Parallel successive convex approximation for nonsmooth nonconvex optimization. In *the Proceedings of NIPS*.

[111] Reddi, S., Sra, S., Poczos, B., and Smola, A. (2016). Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*.

[112] Robbins, H. and Monro, S. (1951). Stochastic approximation methods. *Annals of Mathematical Statistics*, 22:400–407.

[113] Rockafellar, R. T. (1976). Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116.

[114] Ruszczynski, A. (2011). *Nonlinear optimization*. Princeton University.

[115] Scardapane, S., Fierimonte, R., Lorenzo, P. D., Panella, M., and Uncini, A. (2016). Distributed semi-supervised support vector machines. *Neural Networks*, 80:43–52.

[116] Scardapane, S. and Lorenzo, P. D. (2016). A framework for parallel and distributed training of neural networks. *arXiv preprint arXiv:1610.07448*.

[117] Schizas, I., Ribeiro, A., and Giannakis, G. (2008). Consensus in ad hoc wsns with noisy links - part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350 – 364.

[118] Schmidt, M., Roux, N. L., and Bach., F. (2013). Minimizing finite sums with the stochastic average gradient. Technical report, INRIA.

[119] Scutari, G., Facchinei, F., Song, P., Palomar, D. P., and Pang, J.-S. (2014). Decomposition by partial linearization: Parallel optimization of multi-agent systems. *IEEE Transactions on Signal Processing*, 63(3):641–656.

[120] Shalev-Shwartz, S. and Zhang, T. (2013). Proximal stochastic dual coordinate ascent methods for regularzied loss minimization. *Journal of Machine Learning Rsearch*, 14:567–599.

[121] Sharp, K. and Rattray, M. (2010). Dense message passing for sparse principal component analysis. In *Proceedings of 13th International Conference on Artifcial Intelligence and Statistics*, pages 725–732.

[122] Shi, W., Ling, Q., Wu, G., and Yin, W. (2014a). Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966.

[123] Shi, W., Ling, Q., Wu, G., and Yin, W. (2015). A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023.

[124] Shi, W., Ling, Q., Yuan, K., Wu, G., and Yin, W. (2014b). On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761.

[125] Sigg, C. and Buhmann, J. (2008). Expectation-maximization for sparse and nonnegative pca. In *Proceedings of the 25th International Conference on Machine Learning*, pages 960–967.

[126] Spall, J. C. (2003a). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley.

[127] Spall, J. C. (2003b). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. Wiley, New York.

[128] Sra, S. (2012). Scalable nonconvex inexact proximal splitting. In *Advances in Neural Information Processing Systems (NIPS)*.

[129] Srivastava, K. and Nedic, A. (2011). Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772–790.

[130] Sun, Y., Scutari, G., and Palomar, D. (2016). Distributed nonconvex multiagent optimization over time-varying networks. In *50th Asilomar Conference on Signals, Systems and Computers*, pages 788–794.

[131] Tatarenko, T. and Touri, B. (2017). Non-convex distributed optimization. *IEEE Transactions on Automatic Control*, 62(8):3744–3757.

[132] Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423.

[133] Tsitsiklis, J. (1984). Problems in decentralized decision making and computation. Ph.D. thesis, Massachusetts Institute of Technology.

[134] Tychogiorgos, G., Gkelias, A., and Leung, K. K. (2013). A non-convex distributed optimization framework and its application to wireless ad-hoc networks. *IEEE Transactions on Wireless Communications*, 12(9):4286–4296.

[135] Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2670–2678.

[136] Wang, Y. and W. Yin, J. Z. (2015). Global convergence of admm in nonconvex nonsmooth optimization. arXiv Preprint, arXiv:1511.06324.

[137] Wang, Z., Liu, H., and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, 42(6):2164–2201.

[138] Wei, E. and Ozdaglar, A. (2013). On the o (1/k) convergence of asynchronous distributed alternating direction method of multipliers. In *Global Conference on Signal and Information Processing (GlobalSIP)*, pages 551–554.

[139] Wen, Z., Yang, C., Liu, X., and Marchesini, S. (2012). Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems*, 28(11):1–18.

[140] Yan, F., Sundaram, S., Vishwanathan, S. V. N., and Qi, Y. (2013). Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493.

[141] Yildiz, M. E. and Scaglione, A. (2008). Coding with side information for rate-constrained consensus. *IEEE Transactions on Signal Processing.*, 56(8):3753–3764.

[142] Yuan, D. and Ho, D. (2015). Randomized gradient-free method for multiagent optimization over time-varying networks. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6):1342–1347.

[143] Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

[144] Zhang, Y. (2010b). Convergence of a class of stationary iterative methods for saddle point problems. Preprint.

[145] Zhang, Y. and Lin, X. (2015). Disco: Distributed optimization for self-concordant empirical loss. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 362–370. JMLR Workshop and Conference Proceedings.

[146] Zhao, M., Hong, M., and Shi, Q. (2016). A distributed algorithm for dictionary learning over networks. In *Proceedings of 2016 IEEE Global Signal Processing (GlobalSIP)*.

[147] Zhu, H., Cano, A., and Giannakis, G. (2010). Distributed consensus-based demodulation: algorithms and error analysis. *IEEE Transactions on Wireless Communications*, 9(6):2044–2054.

[148] Zhu, M. and Martinez, S. (2010). An approximate dual subgradient algorithm for multi-agent non-convex optimization. In *49th IEEE Conference on Decision and Control (CDC)*, pages 7487–7492.

[149] Zlobec, S. (2005a). On the Liu - Floudas convexification of smooth programs. *Journal of Global Optimization*, 32:401 – 407.

[150] Zlobec, S. (2005b). On the liu–floudas convexification of smooth programs. *Journal of Global Optimization*, 32(3):401–407.