Linearized ADMM for Non-convex Non-smooth Optimization with Convergence Analysis

Qinghua Liu, Xinyue Shen, and Yuantao Gu*

May 6, 2017

Abstract

Linearized alternating direction method of multipliers (ADMM) as an extension of ADMM has been widely used to solve linearly constrained problems in signal processing, machine leaning, communications, and many other fields. Despite its broad applications in non-convex optimization, for a great number of non-convex and non-smooth objective functions, its theoretical convergence guarantee is still an open problem. In this paper, we study the convergence of an existing two-block linearized ADMM and a newly proposed multi-block parallel linearized ADMM for problems with non-convex and non-smooth objectives. Mathematically, we present that the algorithms can converge for a broader class of objective functions under less strict assumptions compared with previous works. Our proposed algorithm can update coupled variables in parallel and work for general non-convex problems, where the traditional ADMM may have difficulties in solving subproblems.

Keywords: Linearized ADMM, non-convex optimization, multi-block ADMM, parallel computation.

1 Introduction

1.1 Background

In signal processing [1], machine learning [2], and communication [3], many of the recently most concerned problems, such as compressed sensing [4], dictionary learning [5], and channel estimation [6], can be cast as optimization problems. In doing so, not only has the design of the solving methods been greatly facilitated, but also a more mathematically understandable and manageable description of the problems has been given. While convex optimization has been relatively well studied [7], non-convex optimization has also appeared in numerous

*The authors are with the Department of Electronic Engineering and Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing 100084, CHINA. The corresponding author of this paper is Yuantao Gu (gyt@tsinghua.edu.cn).

topics such as nonnegative matrix factorization [8], phase retrieval [9], distributed matrix factorization [10], and distributed clustering [11].

The alternating direction method of multipliers (ADMM) is widely used in linearly constrained optimization problems arising in machine learning [12, 13], signal processing [14], as well as other fields [15, 16, 17]. First proposed in the early 1970s, it has been studied extensively [18, 19, 20]. At the very beginning, ADMM was mainly applied in solving linearly constrained convex problems [21] in the following form

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) + h(\mathbf{y}) \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0}, \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}^q$ are variables, and $\mathbf{A} \in \mathbb{R}^{n\times p}, \mathbf{B} \in \mathbb{R}^{n\times q}$ are given. With an augmented Lagrangian function defined as

$$L_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\gamma}) = f(\mathbf{x}) + h(\mathbf{y}) + \langle \boldsymbol{\gamma}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}\|_2^2, \tag{2}$$

where $\boldsymbol{\gamma}$ is the Lagrangian dual variable, the ADMM method updates variables iteratively as the following

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}} L_\beta(\mathbf{x}, \mathbf{y}^k, \boldsymbol{\gamma}^k), \\ \mathbf{y}^{k+1} &= \arg\min_{\mathbf{y}} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}, \boldsymbol{\gamma}^k), \\ \boldsymbol{\gamma}^{k+1} &= \boldsymbol{\gamma}^k + \beta(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}). \end{aligned}$$

For ADMM applied in non-convex problems, although the theoretical convergence guarantee is still an open problem, it can converge fast in many cases [22, 23]. Under certain assumptions on the objective function and linear constraints, researchers have studied the convergence of ADMM for non-convex optimization [24, 25, 26].

In ADMM a subproblem is not necessarily easy or computationally efficient to solve, so linearized ADMM [27, 28] was proposed, by linearizing the objective function or the augmentation term, to make the subproblems solvable. It has been applied in sparse recovery [29], low-rank matrix completion [30], and image restoration [31].

When the problem scale is so large that a two-block ADMM method may no longer be efficient or practical [32, 33], distributed algorithms are in demand to exploit parallel computing resources. Multi-block ADMM was proposed to solve problems in the following form [34]

$$\begin{aligned} &\text{minimize} \quad f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \ldots + f_K(\mathbf{x}_K) \\ &\text{subject to} \quad \mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 + \ldots + \mathbf{A}_K\mathbf{x}_K = \mathbf{0}. \end{aligned} \tag{3}$$

It allows parallel computation [18, 25], [35, 36, 37], and has been used in problems such as sparse statistic machine learning [38] and total variation regularized image reconstruction [39].

1.2 Contribution

In this paper, we study linearized ADMM algorithms for problems with non-convex and non-smooth objective functions. First, we provide convergence analysis of an existing two-block linearized ADMM for problems in form (1) with h differentiable, but under more general assumptions. Then we propose a parallel multi-block ADMM method for problems with *coupled* variables, and provide theoretical convergence guarantee under wide assumptions as well.

Our work has the following improvements compared with some latest works based on ADMM for non-convex optimization.

- This is the first work to study theoretical convergence guarantees for linearized ADMM in non-convex optimization. The strategy we used in convergence analysis is new and apparently different from the previous works.
- We propose, as far as we know, the first ADMM-based algorithm which updates coupled variables in parallel with convergence guarantee for non-convex optimization. It is designed to solve problems of the following form

$$\begin{aligned} &\text{minimize} \quad g(\mathbf{x}_1,\ldots,\mathbf{x}_K)+\textstyle\sum_{i=1}^K f_i(\mathbf{x}_i)+h(\mathbf{y}) \\ &\text{subject to} \quad \mathbf{A}_1\mathbf{x}_1+\cdots+\mathbf{A}_K\mathbf{x}_K+\mathbf{B}\mathbf{y}=\mathbf{0}, \end{aligned} \tag{4}$$

  where h and g are differentiable. The augmented Lagrangian function of this problem is defined as below

$$L_\beta(\mathbf{x},\mathbf{y},\boldsymbol{\gamma})=g(\mathbf{x})+\sum_{i=1}^K f_i(\mathbf{x}_i)+h(\mathbf{y})+\left\langle\boldsymbol{\gamma},\sum_{i=1}^K \mathbf{A}_i\mathbf{x}_i+\mathbf{B}\mathbf{y}\right\rangle+\frac{\beta}{2}\left\|\sum_{i=1}^K \mathbf{A}_i\mathbf{x}_i+\mathbf{B}\mathbf{y}\right\|_2^2. \tag{5}$$

  Compared with previous works [18, 25], [35, 36, 37] on parallel ADMM, it can be seen that in (5) we consider variables $\mathbf{x}_1,\ldots,\mathbf{x}_K$ coupled in the function g and $\frac{\beta}{2}\|\sum_{i=1}^K \mathbf{A}_i\mathbf{x}_i+\mathbf{B}\mathbf{y}\|_2^2$ but still update every block in parallel in our proposed algorithm.
- In contrast with the most prior works focusing on ADMM in non-convex optimization, our assumptions for the convergence analysis are much weaker than the ones in [25, 26], and not stronger than the ones in [24]. We study the convergence of linearized ADMM algorithms for non-convex problems in a broader scope than previous works have revealed. Detailed comparisons on assumptions will be given in Section 8.

1.3 Outline

The remainder of this paper is organized as follows. In Section 2 some preliminaries are introduced. Section 3 provides convergence analysis for a two-block linearized ADMM for non-convex problems under certain broad assumptions. In Section 4 we propose a parallel muti-block linearized ADMM, and provide convergence guarantee under general assumptions as well. Section 5 gives detailed discussions on the objective functions of conforming problems and the update rules to demonstrate the advantages of this work. Section 6 and 7 prove the convergence theorems for the considered two-block linearized ADMM and the proposed parallel multi-block linearized ADMM, respectively. Section 8 compares our results with some related works. We conclude this work in Section 9.

2 Preliminary

2.1 Notation

We use bold capital letters for matrices, bold small case letters for vectors, and non-bold letters for scalars. We use $\mathbf{x}^k$ to denote the value of $\mathbf{x}$ after kth iteration and $\mathbf{x}_i$ to denote its ith block. We use $\mathbf{A}_{<i}\mathbf{x}_{<i}$ and $\mathbf{A}_{>i}\mathbf{x}_{>i}$ to denote $\sum_{t=1}^{i-1}\mathbf{A}_t\mathbf{x}_t$ and $\sum_{t=i+1}^{K}\mathbf{A}_t\mathbf{x}_t$, respectively. The gradient of function f at $\mathbf{x}$ for the ith component is denoted as $\nabla_i f(\mathbf{x})$, and the *regular subgradient* of f for the ith component which is defined at a point $\mathbf{x}$ [40], is denoted as $\partial_i f(\mathbf{x})$. The smallest eigenvalue of matrix $\mathbf{x}$ is denoted as $\lambda_{\mathbf{X}}$. Without specification, $\|\cdot\|$ denotes ℓ_2 norm. $\mathbf{Im}(\mathbf{X})$ denotes the image of matrix $\mathbf{X}$. In multi-block ADMM, $\mathbf{x} = \left[\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_K^{\mathrm{T}}\right]^{\mathrm{T}}$ denotes the collection of variables.

2.2 Definition

Definition 1. ***(Lower Semi-continuous)*** *[41] Assume $f(\mathbf{x})$ satisfies*

$$\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0),$$

then $f(\mathbf{x})$ is said to be lower semi-continuous at $\mathbf{x}_0$.

Lower semi-continuity is a more generalized concept than continuity. An example can be seen in Remark 6, where function f is discontinuous anywhere but lower semi-continuous over some points.

Definition 2. ***(Regular Subgradient)*** *[40] The regular subgradient of function f at $\mathbf{x}_0$ is defined as*

$$\partial f(\mathbf{x}_0) = \{\mathbf{v} : f(\mathbf{x}) \geq f(\mathbf{x}_0) + \langle \mathbf{v}, \mathbf{x} - \mathbf{x}_0\rangle + o(\|\mathbf{x} - \mathbf{x}_0\|)\}.$$

Definition 3. ***(Restricted Quasi-convex)*** *Given a function f, $\mathcal{X}$ is a set such that for any $\mathbf{x} \in \mathcal{X}$ the regular subgradient $\partial f(\mathbf{x})$ is not empty. If there exists $\alpha > 0$ such that*

$$f(\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|^2 \geq f(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \mathbf{d}\rangle$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\mathbf{d} \in \partial f(\mathbf{y})$, then function f is called restricted quasi-convex over $\mathcal{X}$.

Remark 1. *A linear combination of lower semi-continuous (or restricted quasi-convex) functions over $\mathcal{X}$ is still lower semi-continuous (or restricted quasi-convex) over the set.*

The ρ-convex function is formally defined as follows.

Definition 4. *(**ρ-convex Function**)[42] Assume that a function f satisfies that for any $\mathbf{x}_1$, $\mathbf{x}_2$, and $\lambda \in [0, 1]$,*

$$f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2) - \lambda(1-\lambda)\rho\|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

If $\rho < 0$, f is said to be weakly convex.

Note that if $\rho > 0$, the basic inequality of convex functions is strengthened, and f is said to be strongly convex.

Remark 2. *All convex, weakly convex, and Lipschitz differentiable functions are restricted quasi-convex over $\mathbb{R}^n$ and therefore, restricted quasi-convex over any sequence.*

Lemma 1. *If $h(\mathbf{y})$ is L_h-Lipschitz differentiable, i.e., there exists $L_h > 0$ such that for all $\mathbf{y}, \mathbf{y}'$,*

$$\|\nabla h(\mathbf{y}) - \nabla h(\mathbf{y}')\|_2 \leq L_h\|\mathbf{y} - \mathbf{y}'\|_2,$$

then

$$h(\mathbf{y}) \geq h(\mathbf{y}') + \langle \mathbf{y} - \mathbf{y}', \nabla h(\mathbf{z}) \rangle - \frac{L_h}{2}\|\mathbf{y} - \mathbf{y}'\|_2^2, \tag{6}$$

where $\mathbf{z}$ denotes $\mathbf{y}$ or $\mathbf{y}'$.

Proof. The proof is postponed to Appendix 10.1. □

Definition 5. *(**Coercive Function**) Assume that function f is defined on $\mathcal{X}$, and for any $\|\mathbf{x}^k\| \to +\infty$ and $\mathbf{x}^k \in \mathcal{X}$, we have $f(\mathbf{x}^k) \to +\infty$, then function f is said to be coercive over $\mathcal{X}$.*

Remark 3. *Any function is coercive over bounded set.*

3 Two-block Linearized ADMM

In this section, we study a linearized ADMM to solve the two-block non-convex problem (1) possibly with function f non-smooth. Its convergence assumptions are, as far as we know, the broadest among the current ADMM algorithms for non-convex optimization.

Algorithm 1: Two-block linearized ADMM algorithm

Initialize $\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\gamma}^0$.

While $\max\{\|\mathbf{y}^k - \mathbf{y}^{k-1}\|, \|\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}\|\} > \varepsilon$

1) $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} L_\beta(\mathbf{x}, \mathbf{y}^k, \boldsymbol{\gamma}^k)$;

2) $\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} \bar{h}(\mathbf{y})$, where

$$\bar{h}(\mathbf{y}) = \langle \mathbf{y} - \mathbf{y}^k, \nabla h(\mathbf{y}^k)\rangle + \tfrac{L_{\bar{h}}}{2}\|\mathbf{y} - \mathbf{y}^k\|^2 + \langle \boldsymbol{\gamma}^k, \mathbf{B}\mathbf{y}\rangle + \tfrac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}\|^2; \tag{7}$$

3) $\boldsymbol{\gamma}^{k+1} = \boldsymbol{\gamma}^k + \beta(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1})$;

4) $k = k + 1$.

end While

3.1 Updating Rules

We try to solve problem (1) using a linearized ADMM algorithm. Being different from the traditional ADMM, in the update of $\mathbf{y}$, the algorithm replaces h by its approximate,

$$\langle \mathbf{y} - \mathbf{y}^k, \nabla h(\mathbf{y}^k)\rangle + \frac{L_{\bar{h}}}{2}\|\mathbf{y} - \mathbf{y}^k\|^2$$

which is a linearized term plus a regularization term ($L_{\bar{h}} > 0$). In this way, the subproblem is converted into a proximal problem and usually easier to solve. The update rules are summarized in Algorithm 1.

3.2 Assumptions

To guarantee the convergence of the linearized ADMM in Algorithm 1, we make the following assumption.

Assumption 1. *Assume that problem* (1) *satisfies the conditions below.*

1. *Function* $f(\mathbf{x})$ *is lower semi-continuous.*
2. *Function* $h(\mathbf{y})$ *is* L_h*-Lipschitz differentiable.*
3. *Function* $f(\mathbf{x}) + h(\mathbf{y})$ *is coercive over the feasible set* $\{[\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{p+q} : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0}\}$.
4. *Matrix* $\mathbf{B}$ *has full column rank, and* $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$.

Remark 4. *In fact, our convergence analysis only requires function* f *to be lower semi-continuous at the limit points of* $\{\mathbf{x}^k\}$. *This will be shown in Section 6.4.*

It can be seen that we impose stronger requirements on the function h of $\mathbf{y}$ than on f of $\mathbf{x}$. We will show in Section 6 that under these assumptions the objective function will have sufficient decent in the update of $\mathbf{y}$ to counteract the ascent in the dual update. As a benefit, function f can be either non-convex or non-smooth. We will discuss the plausibility of our assumptions in detail by analyzing examples in Section 5 and comparing with assumptions in previous works in Section 8.

3.3 Theoretical Analysis

We will give convergence analysis under the above set of assumptions. Note that in this section, we refer L_β to the augmented Lagrangian function defined in (2). To begin with, we show that L_β and the primal and dual residues are able to converge in the following theorem.

Theorem 1. *For the linearized ADMM in Algorithm 1, the sequence $\{L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ is convergent, and both $\|\mathbf{y}^{k+1} - \mathbf{y}^k\|$ and $\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\|$ converge to zero as k approaches infinity.*

Proof. The proof is postponed to Section 6.1. □

Theorem 1 illustrates that the function L_β will converge, and the changes of $\mathbf{y}$ and $\boldsymbol{\gamma}$ after one iteration, which are the primal and the dual residues, will converge to zero as well. From such result we can easily have the following knowledge on the sequence of iterates.

Corollary 1. *For the linearized ADMM in Algorithm 1, the generated sequences $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, and $\{\boldsymbol{\gamma}^k\}$ are bounded.*

Proof. The proof is postponed to Section 6.2. □

The boundedness in Corollary 1 insures that there is at least one limit point in sequence $\{(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$. In the following theorems, we further claim that there exits at least one limit point satisfying certain properties.

Theorem 2. *For the linearized ADMM in Algorithm 1, the sequence $\{(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ has at least one limit point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ such that*

$$\nabla_{\boldsymbol{\gamma}} L_\beta(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*) = \mathbf{0},$$

and

$$\nabla_{\mathbf{y}} L_\beta(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*) = \mathbf{0}.$$

Proof. The proof is postponed to Section 6.3. □

Theorem 2 illustrates that a limit point of $\{(\mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ is a stationary point of the function $L_\beta(\mathbf{x}^*, \cdot, \cdot)$.

Theorem 3. *For the linearized ADMM in Algorithm 1, $\mathbf{x}^*$ is a minimum point of the function $L_\beta(\cdot, \mathbf{y}^*, \boldsymbol{\gamma}^*)$, where $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ is the one defined in Theorem 2.*

Proof. The proof is postponed to Section 6.4. □

The assumption that function f is lower semi-continuous ensures that f will not have a spike-like arise at the limit points of $\{\mathbf{x}^k\}$.

Both Theorem 2 and 3 answer the question of where the function L_β converges to. Because of the linearization of function h in the update of $\mathbf{y}$, the algorithm only utilizes the first order information of $h(\mathbf{y})$. Therefore, in Theorem 2, unlike Theorem 3 where we can have that $\mathbf{x}^*$ is a minimum point of the function $L_\beta(\cdot, \mathbf{y}^*, \boldsymbol{\gamma}^*)$, the best result we can expect for variable $\mathbf{y}$ is converging to a stationary point characterized merely by the first order information.

4 Multi-block Parallel Linearized ADMM

In this section, we turn to multi-block optimization problems in form of (4) and propose a multi-block linearized ADMM which can update blocks of variables in parallel even when they are coupled in the objective function.

4.1 Updating Rules

We try to solve problem (4) using a multi-block parallel linearized ADMM algorithm. The function h is linearized and regularized in the update of $\mathbf{y}$, exactly as that in the two-block algorithm. In the update of $\mathbf{x}$ for each $\mathbf{x}_i$, we further linearize the function g, the augmented item $\frac{\beta}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}\|^2$ and introduce regularization as follows

$$\langle \mathbf{x}_i - \mathbf{x}_i^k, \nabla_i g(\mathbf{x}^k) + \beta \mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k)\rangle + \frac{L_{f_i}}{2}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2$$

where L_{f_i} is a positive constant. In this way, the two subproblems for variable update all becomes proximal problems and usually easier to solve. Besides, each block $\mathbf{x}_i$ can be optimized in parallel, although they are originally coupled in function g and the augmented item. The update rules are listed in Algorithm 2.

4.2 Assumptions

To guarantee the convergence of the multi-block parallel linearized ADMM algorithm, we make the following assumption.

Assumption 2. *We assume that problem* (4) *satisfies the conditions below.*

1. *Function g is L_g-Lipschitz differentiable. Functions f_i are restricted quasi-convex over $\{\mathbf{x}_i^k\}$ for $i = 1, \ldots, K$.*

2. *Function h is L_h-Lipschitz differentiable.*

3. *Matrix $\mathbf{B}$ has full column rank, and $\mathbf{Im}(\mathbf{B}) \subset \mathbf{Im}(\mathbf{A})$.*

Algorithm 2: Multi-block parallel linearized ADMM algorithm

Initialize $\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\gamma}^0$.

While $\max\{\|\mathbf{x}^k - \mathbf{x}^{k-1}\|, \|\mathbf{y}^k - \mathbf{y}^{k-1}\|, \|\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^{k-1}\|\} > \varepsilon$

1) **For** $i = 1, \dots, K$ **in parallel**:

$\mathbf{x}_i^{k+1} = \arg\min_{\mathbf{x}_i} \bar{f}_i(\mathbf{x}_i)$, where

$$\bar{f}_i(\mathbf{x}_i) = \left\langle \mathbf{x}_i - \mathbf{x}_i^k, \nabla_i g(\mathbf{x}^k) + \beta \mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k)\right\rangle + \frac{L_{f_i}}{2}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2 + f_i(\mathbf{x}_i) + \langle \boldsymbol{\gamma}^k, \mathbf{A}_i\mathbf{x}_i\rangle; \quad (8)$$

end For

2) $\mathbf{y}^{k+1} = \arg\min_{\mathbf{y}} \bar{h}(\mathbf{y})$, where

$$\bar{h}(\mathbf{y}) = \langle \mathbf{y} - \mathbf{y}^k, \nabla h(\mathbf{y}^k)\rangle + \frac{L_{\bar{h}}}{2}\|\mathbf{y} - \mathbf{y}^k\|^2 + \langle \boldsymbol{\gamma}^k, \mathbf{B}\mathbf{y}\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}\|^2; \quad (9)$$

3) $\boldsymbol{\gamma}^{k+1} = \boldsymbol{\gamma}^k + \beta(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1})$;

4) $k = k + 1$.

end While

---

4. *Function $f(\mathbf{x}) + h(\mathbf{y})$ is coercive over the feasible set $\{[\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{p+q} : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0}\}$.*

The second, third, and forth assumptions are the same as the first algorithm. The first assumption is newly added which is necessary in the proof of convergence.

4.3 Theoretical analysis

In this section we refer L_β to the function in (5). The following theorems illustrate that the objective function will converge, the increments of all variables after one iteration will converge to zero, and that the limit point of the iterate sequence is a stationary point.

Theorem 4. *For the multi-block parallel linearized ADMM in Algorithm 2, if L_{f_i} is chosen to satisfy*

$$L_{f_i} > \frac{1}{2}\left((K - i)L_g^2 + L_g + (K + 1)\beta L_{\mathbf{A}_i} + 1\right), \quad (10)$$

where $L_{\mathbf{A}_i}$ is the largest singular value of $\mathbf{A}_i^T\mathbf{A}_i$, then we have the following.

1. *The sequence $\{L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ is convergent. The primal residue $\|\mathbf{y}^{k+1} - \mathbf{y}^k\|$, $\|\mathbf{x}^k - \mathbf{x}^{k+1}\|$ and dual residue $\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\|$ converge to zero as k approaches infinity.*

2. *The generated sequences $\{\mathbf{x}^k\}$, $\{\mathbf{y}^k\}$, and $\{\boldsymbol{\gamma}^k\}$ are bounded.*

3. *The sequence $\{(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ has a limit point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ such that*

$$\nabla_{\boldsymbol{\gamma}} L_\beta(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*) = \mathbf{0},$$

and

$$\nabla_{\mathbf{y}} L_\beta(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*) = \mathbf{0}.$$

Proof. The proof is postponed to Section 7.1. □

Theorem 4 is exactly in parallel to the theories of two-block Linearized ADMM. Specifically, the first claim in Theorem 4 corresponds to Theorem 1, the second claim corresponds to Corollary 1, and the third claim corresponds to Theorem 2.

Theorem 4 illustrates that the function L_β will converge and the primal and the dual residues will converge to zero as well. It also shows that a limit point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ exists which is also a stationary point of the function $L_\beta(\mathbf{x}^*, \cdot, \cdot)$.

Remark 5. *Notice that* (10) *contains index i in $(K-i)$, which contradicts with the intuitive idea that, in parallel ADMM, every block is symmetric and a bound on L_{f_i} should not depend on index i. This phenomenon is caused by the fact that the blocks are indeed coupled in g and to update them separately is not equivalent to updating them as one. In a novel decomposition* (46) *we make in the convergence analysis, the variable blocks are no longer symmetric, and the blocks with smaller index affect the ones with larger index.*

Theorem 5. *For the multi-block parallel linearized ADMM in Algorithm 2, there exists a sequence $\{\bar{\mathbf{d}}_i^{k+1}\}$, where $\bar{\mathbf{d}}_i^{k+1} \in \partial_{\mathbf{x}_i} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1})$, such that $\lim\limits_{k\to+\infty} \|\bar{\mathbf{d}}_i^{k+1}\| = 0$ for $i = 1, \dots, K$.*

Proof. The proof is postponed to Appendix 10.10. □

Theorem 5 illustrates that there exists a regular subgradient sequence converging to zero in a neighbor of $\mathbf{x}^*$. If function f also satisfies $\lim\limits_{k\to+\infty} f(\mathbf{x}^k) = f(\mathbf{x}^*)$, then $\mathbf{0}$ belongs to the general subgradient of function $L_\beta(\cdot, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ at $\mathbf{x}^*$ by Definition 8.4 in [40].

5 Discussions

5.1 Function f (or f_i) in the objective

In this part we present that the following general classes of functions can meet the requirements in the above assumptions. Therefore, our theorems guarantee the convergence of the algorithms if f (or f_i) belongs to one of the following commonly encountered classes.

5.1.1 ρ-convex function

ρ-convex functions are often encountered in signal processing such as sparse recovery [23, 43].

According to Definition 4, ρ-convex function includes both convex and weakly convex (non-convex) functions. From Proposition 4.3 in [42], all ρ-convex functions are continuous and therefore lower semi-continuous at any point, which means Assumption 1.1 is satisfied. By assigning α to $-\rho$ in Proposition 4.8 in [42], we get that all weakly ρ-convex functions are restricted quasi-convex and therefore Assumption 2.1 is satisfied.

5.1.2 Indicator function of compact manifold

The indicator function of compact manifold $\mathcal{M}$ is defined as follows

$$f(\mathbf{x}) = \begin{cases} +\infty & \mathbf{x} \notin \mathcal{M}, \\ 0 & \mathbf{x} \in \mathcal{M}. \end{cases}$$

For the two-block linearized ADMM Algorithm 1, by the $\mathbf{x}$-updating rule, $\{\mathbf{x}^k\} \subset \mathcal{M}$. Because $\mathcal{M}$ is closed, any limit point of $\{\mathbf{x}^k\}$ belongs to $\mathcal{M}$. Therefore, by directly referring to Definition 1, f is lower semi-continuous at limit points of $\{\mathbf{x}^k\}$. By Remark 4 we can see that it satisfies the convergence assumptions.

For the multi-block linearized ADMM Algorithm 2, if function f_i is an indicator function of a compact manifold $\mathcal{M}_i$, then by the $\mathbf{x}_i$-updating rule, $\{\mathbf{x}_i^k\}$ belongs to $\mathcal{M}_i$. Besides, for any $\mathbf{x}_i$ in $\mathcal{M}_i$, the regular subgradient of f_i at point $\mathbf{x}_i$ is $\{\mathbf{0}\}$, so by Definition 3 f_i is restricted quasi-convex over $\mathcal{M}_i$ and therefore restricted quasi-convex over $\{\mathbf{x}_i^k\}$.

Remark 6. *Consider the following general form of problem, where $\mathcal{M}$ is a finite subset of $\mathbb{Z}$ and f is restricted quasi-convex*

$$\textit{minimize } f(\mathbf{y}) \quad \textit{subject to } \mathbf{y} \in \mathcal{M}. \tag{11}$$

*This problem is called **integer programming** which is widely used in network design [44], smart grid [45], statistic learning [46], and other fields [47]. Problem (11) can be converted to the following*

$$\begin{aligned} &\textit{minimize} \quad \tau(\mathbf{x}) + \big(f(\mathbf{x}) - h(\mathbf{x})\big) + h(\mathbf{y}) \\ &\textit{subject to} \quad \mathbf{x} = \mathbf{y}, \end{aligned} \tag{12}$$

where $\tau(\mathbf{x})$ is the indicator function of $\mathcal{M}$, and function h can be any nonzero Lipschitz function. It can be verified that problem (12) satisfies both the assumptions for Algorithm 1 and 2. In practice, function h is chosen appropriately for solving the subproblems.

5.2 Update rules

5.2.1 Update of x

The two-block linearized ADMM can still converge, even if we do not find a minimum in updating $\mathbf{x}$. When updating $\mathbf{x}$ in Algorithm 1, we find a minimum point of $L_\beta(\mathbf{x}, \mathbf{y}^k, \boldsymbol{\gamma}^k)$. If instead of finding a minimum we accept any update $\mathbf{x}^{k+1}$ such that the function L_β decreases, i.e., $L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) \leq L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)$, all theoretical results including Theorem 1, Corollary 1, and Theorem 2, will still hold. This statement can be derived from the proof. It means that $L_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\gamma})$ still converges, and the limit point $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ is a stationary point of $L_\beta(\mathbf{x}, \mathbf{y}, \boldsymbol{\gamma})$ for $\mathbf{y}$ and $\boldsymbol{\gamma}$, but not necessarily a minimum point for $\mathbf{x}$.

In scenarios where finding a minimum of $L_\beta(\mathbf{x}, \mathbf{y}^k, \boldsymbol{\gamma}^k)$ is difficult but finding a descent point is relatively much easier, a simple modification of Algorithm 1 that replaces the minimum with a descent point will still work.

In multi-block linearized ADMM Algorithm 2, we break the update of $\mathbf{x}$ into many subproblems by linearizing the differentiable part and the augmented item in Lagrangian function. By doing so, the subproblems are more solvable than the original subproblem, and parallel computation can be used to accelerate the algorithm.

5.2.2 Linearization in update of $\mathbf{y}$

Traditional ADMM updates $\mathbf{y}$ by finding a solution to $\arg\min L_\beta(\mathbf{x}^{k+1}, \mathbf{y}, \boldsymbol{\gamma}^k)$. However, the difficulties in solving this subproblem restricts the application of ADMM. To address this issue, the linearized ADMM converts the original subproblem into a proximal problem that has closed solution. In this way, the subproblem in the update of $\mathbf{y}$ is solvable for any differentiable $h(\mathbf{y})$.

6 Convergence Proof for Linearized ADMM Algorithm 1

In this section, all notations $\mathbf{x}^k$, $\mathbf{y}^k$, and $\boldsymbol{\gamma}^k$ refer to the ones in Algorithm 1.

6.1 Proof of Theorem 1

We need two important inequalities for the proof, which are derived from the update rules and Assumption 1.

Lemma 2. *There exits positive constant $C_\mathbf{R}$ merely determined by matrix $\mathbf{B}$ such that for any $l > k$*

$$\|\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k\|^2 \leq C_\mathbf{R}\|\mathbf{B}^\mathrm{T}(\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k)\|^2.$$

Proof. The proof is postponed to Appendix 10.2. □

Lemma 3. *The following equality holds for $\boldsymbol{\gamma}^{k+1}$, $\mathbf{y}^k$, and $\mathbf{y}^{k+1}$*

$$\mathbf{B}^\mathrm{T}\boldsymbol{\gamma}^{k+1} = -\nabla h(\mathbf{y}^k) - L_{\bar{h}}(\mathbf{y}^{k+1} - \mathbf{y}^k).$$

Proof. The proof is postponed to Appendix 10.3. □

Lemma 3 provides a way to express $\boldsymbol{\gamma}^{k+1}$ using $\mathbf{y}^k$ and $\mathbf{y}^{k+1}$, which is a technique widely used in the convergence proof for non-convex ADMM algorithms [24, 25].

Now we are ready to prove Theorem 1. We first give bounds on the descent or ascent of the Lagrangian function (2) after every update by using the quadratic form of the primal residual. Specifically, in the following, Lemma 4 presents that L_β does not increase after the $\mathbf{x}$-updating step, Lemma 5 shows that the descent of L_β is lower bounded after the $\mathbf{y}$-updating step, and Lemma 6 demonstrates that the ascent of L_β is upper bounded after the $\boldsymbol{\gamma}$-updating step.

Lemma 4. *The following inequality holds for the update of* $\mathbf{x}$

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) \leq L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k). \tag{13}$$

Proof. It can be directly derived from the $\mathbf{x}$-updating rule.

□

Lemma 5. *The following inequality holds for the update of* $\mathbf{y}$

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^k) \geq C_0\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2,$$

where C_0 *is a positive constant.*

Proof. The proof is postponed to Appendix 10.4. □

Lemma 6. *The following inequality holds for the update of* $\boldsymbol{\gamma}$

$$\begin{aligned} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^k) =& \frac{1}{\beta}\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\|^2 \\ \leq& C_1\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + C_2\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2, \end{aligned} \tag{14}$$

where C_1 *and* C_2 *are positive constants.*

Proof. The proof is postponed to Appendix 10.5. □

Then we design a sequence $\{m_k\}_{k=1}^{+\infty}$ by

$$m_k = L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k) + C_3\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2, \tag{15}$$

where C_3 is a positive constant. It will be proved convergent.

Lemma 7. *The sequence* $\{m_k\}$ *defined in* (15) *is convergent. In addition, the parameters satisfying its convergence exist.*

Proof. We will first present that $\{m_k\}$ is monotonically decreasing by

$$m_k - m_{k+1} \geq (C_0 - C_1 - C_3)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + (C_3 - C_2)\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2,$$

and then prove that $\{m_k\}$ is lower bounded by

$$m_k \geq f(\mathbf{x}^k) + h(\mathbf{y}'_k) + \frac{1}{2}\|\mathbf{y}^k - \mathbf{y}'_k\|^2, \tag{16}$$

where $\mathbf{y}'_k$ is defined by $\mathbf{B}\mathbf{y}'_k = -\mathbf{A}\mathbf{x}^k$. Notice that $\mathbf{y}'_k$ always exists because of the assumption $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$.

The detailed proof is postponed to Appendix 10.6. The parameters satisfying the convergence are discussed in Appendix 10.7. □

By the convergence of $\{m_k\}$, $\|\mathbf{y}^{k+1} - \mathbf{y}^k\|$ converges to zero. By the definition of $\{m_k\}$ and its convergence, we readily get the convergence of $L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)$. According to Lemma 6, $\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k\|$ converges to zero as well.

6.2 Proof of Corollary 1

Recall (16) in the proof of Lemma 7. Because $f(\mathbf{x})+h(\mathbf{y})$ is coercive over the feasible set, if $\{(\mathbf{x}^k, \mathbf{y}'_k)\}$ diverges, then the RHS of (16) diverges to positive infinity, which contradicts with the convergence of $\{m_k\}$.

Because of the term $\frac{1}{2}\|\mathbf{y}^k - \mathbf{y}'_k\|^2$ on the RHS of (16), the boundedness of $\{\mathbf{y}^k\}$ can be derived from the boundedness of $\{\mathbf{y}'_k\}$.

In order to prove that $\{\boldsymbol{\gamma}^k\}$ is bounded, we only need to prove $\{\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^0\}$ is bounded. By Lemma 2, it is equivalent to the boundedness of $\{\mathbf{B}^T(\boldsymbol{\gamma}^k - \boldsymbol{\gamma}^0)\}$ and further equivalent to the boundedness of $\{\mathbf{B}^T\boldsymbol{\gamma}^k\}$ which can be deduced from Lemma 3 and the boundedness of $\{\mathbf{y}^k\}$.

Remark 7. *The coercivity of $f(\mathbf{x})+h(\mathbf{y})$ over the feasible set implies that it is lower bounded over the feasible set.*

6.3 Proof of Theorem 2

By Corollary 1, $\{(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\}$ has at least one limit point. Besides, when k approaches infinity, we have

$$\begin{aligned}\nabla_{\boldsymbol{\gamma}} L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k) =&\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}\\ =&\frac{1}{\beta}(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k) \to \mathbf{0}.\end{aligned}$$

By Lemma 3, when k approaches infinity, we have

$$\begin{aligned}&\nabla_{\mathbf{y}} L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)\\ =&\nabla h(\mathbf{y}^k) + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta\mathbf{B}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k)\\ \to&\nabla h(\mathbf{y}^k) + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^{k+1} + \mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k)\\ =&-L_{\bar{h}}(\mathbf{y}^{k+1} - \mathbf{y}^k) + \mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^k) \to \mathbf{0}.\end{aligned}$$

Considering that $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$ is a limit point of $(\mathbf{x}^k, \mathbf{y}^k,$
$\boldsymbol{\gamma}^k)$, we get the result.

6.4 Proof of Theorem 3

Firstly, the fact that $f(\mathbf{x})$ is lower semi-continuous can imply the following conclusion:

$$\liminf_k f(\mathbf{x}_k) \geq f(\mathbf{x}^*), \qquad \forall \mathbf{x}_k \to \mathbf{x}^*.$$

Besides, function L_β is continuous for variables $\mathbf{y}$ and $\boldsymbol{\gamma}$.

By the definition of $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$, there exists a subsequence of $(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k)$, which we define as $(\mathbf{x}^{n_k}, \mathbf{y}^{n_k}, \boldsymbol{\gamma}^{n_k})$, such that $(\mathbf{x}^{n_k}, \mathbf{y}^{n_k}, \boldsymbol{\gamma}^{n_k})$ converges to $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*)$. So $\forall \mathbf{x}_0 \in \mathbb{R}^p$, we have

$$
\begin{aligned}
L_\beta(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*) &= \lim_{k\to+\infty} L_\beta(\mathbf{x}^*, \mathbf{y}^{n_k}, \boldsymbol{\gamma}^{n_k}) \\
&\le \lim_{k\to+\infty} L_\beta(\mathbf{x}^{n_k}, \mathbf{y}^{n_k}, \boldsymbol{\gamma}^{n_k}) \\
&= \lim_{k\to+\infty} L_\beta(\mathbf{x}^{n_k}, \mathbf{y}^{n_k-1}, \boldsymbol{\gamma}^{n_k-1}) \\
&\le \lim_{k\to+\infty} L_\beta(\mathbf{x}_0, \mathbf{y}^{n_k-1}, \boldsymbol{\gamma}^{n_k-1}) \qquad (17) \\
&= L_\beta(\mathbf{x}_0, \mathbf{y}^*, \boldsymbol{\gamma}^*),
\end{aligned}
$$

where (17) is a result of the $\mathbf{x}$-updating rule.

7 Convergence Proof for Multi-block Parallel Linearized ADMM Algorithm 2

In this section, all notations $\mathbf{x}^k$, $\mathbf{y}^k$, and $\boldsymbol{\gamma}^k$ refer to the ones in Algorithm 2.

7.1 Proof of Theorem 4

The following lemma proves that the subproblems in variable updates are lower bounded.

Lemma 8. *By choosing $L_{f_i} > L_g$ and β large enough, $L_{\bar{h}} > 0$, the subproblems in the updates of $\mathbf{x}$ and $\mathbf{y}$ are lower bounded.*

Proof. The proof is postponed to Appendix 10.8. □

Because the multi-block ADMM Algorithm 2 only differs from the two-block ADMM Algorithm 2 in the $\mathbf{x}$-updating rule, we only show the descent of the Lagrangian function (5) from updating $\mathbf{x}$ in this proof. The descent from updating $\mathbf{y}$ and the ascent from updating $\boldsymbol{\gamma}$ are the same as the two-block ADMM, which have been proved respectively in Lemma 5 and Lemma 6.

The following lemma is to verify that the descent of L_β is lower bounded. Regular subgradient is introduced, which is appropriate for our restricted quasi-convex assumption.

Lemma 9. *The following equality holds for the update of $\mathbf{x}$*

$$
\begin{aligned}
&L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) \\
\ge &\sum_{i=1}^{K} \left(f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) + \frac{\bar{\alpha}_i}{2}\|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 - \left\langle \mathbf{d}_i^{k+1}, \mathbf{x}_i^k - \mathbf{x}_i^{k+1} \right\rangle \right), \qquad (18)
\end{aligned}
$$

where $\mathbf{d}_i^{k+1}$ is in $\partial f_i(\mathbf{x}_i^{k+1})$ and

$$
\bar{\alpha}_i := 2L_{f_i} - \left((K-i)L_g^2 + L_g + (K+1)\beta L_{\mathbf{A}_i} + 1 \right), \qquad (19)
$$

for all $1 \le i \le K$.

Proof. The proof is postponed to Appendix 10.9. □

According to Lemma 9 and the condition of (10), if function f_i is restricted quasi-convex over $\{\mathbf{x}_i^k\}$ with constant $\alpha_i < \bar{\alpha}_i$ for $i = 1, \dots, K$, by choosing L_{f_i} large enough, which is always possible.

$$f_i(\mathbf{x}_i^k) + \frac{\alpha_i}{2}\|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2 \geq f_i(\mathbf{x}_i^{k+1}) + \langle \mathbf{d}_i^{k+1}, \mathbf{x}_i^k - \mathbf{x}_i^{k+1} \rangle, \tag{20}$$

then

$$L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k) \geq L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) + \sum_{i=1}^{K} \frac{\bar{\alpha}_i - \alpha_i}{2}\|\mathbf{x}_i^k - \mathbf{x}_i^{k+1}\|^2. \tag{21}$$

Now we can see that all the lemmas in Section 6 hold for multi-block linearized ADMM Algorithm 2, and all the theorems and corollaries except Theorem 3 derived for two-block linearized ADMM also hold here. Specifically, Theorem 1 implies the first claim in Theorem 4, Corollary 1 implies the second claim, and Theorem 2 implies the third claim. The conclusion $\|\mathbf{x}^k - \mathbf{x}^{k+1}\| \to 0$, which Algorithm 1 doesn't have, can be derived from the convergence of L_β and (21).

8 Related Works

The contribution of this work will be highlighted by comparing with some most recent works on ADMM for non-convex optimization.

8.1 Multi-block ADMM for Nonconvex Non-smooth Optimization

The paper [24] studied ADMM for minimizing a non-convex and possibly non-smooth objective function subject to linear constraints. The algorithm sequentially updates the primal variables in the order $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{y}$, followed by updating the dual variable. Notice that $\mathbf{y}$ is separated from $\mathbf{x}_i$ as it has a special role in convergence analysis, which inspired our works. Under the following assumptions, they prove that the algorithm is able to converge to some stationary point.

1) The objective function $f(\mathbf{x}) + h(\mathbf{y})$ is coercive over the feasible set $\{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q} : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{0}\}$.

2) $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$.

3) Function h is Lipschitz differentiable.

4) Function f has the form

$$f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^{p} f_i(\mathbf{x}_i),$$

where g is Lipschitz differentiable, and functions f_i are restricted prox-regular or piecewise linear for $i = 1, \dots, p$.

5) For any $i = 1, \dots, p$, $\mathbf{x}_{<i}$, and $\mathbf{x}_{>i}$, there exists a Lipschitz continuous map $F_i : \mathbf{Im}(\mathbf{A}_i) \to \mathbb{R}^{n_i}$ obeying $F_i(\mathbf{u}) = \arg\min_{\mathbf{x}_i}\{f(\mathbf{x}_{<i}, \mathbf{x}_i, \mathbf{x}_{>i}) : \mathbf{A}_i\mathbf{x}_i = \mathbf{u}\}$.

6) There exists a Lipschitz continuous map $H : \mathbf{Im}(\mathbf{B}) \to \mathbb{R}^p$ obeying $H(\mathbf{u}) = \arg\min_{\mathbf{y}}\{h(\mathbf{y}) : \mathbf{B}\mathbf{y} = \mathbf{u}\}$.

Let us compare the above assumptions with our Assumption 2. The first three assumptions are equivalent to ours. The forth assumption and our assumption on f cannot include each other. We put no restrictions on $\mathbf{A}_i$, while they add complex requirements on $\mathbf{A}_i$ in the fifth assumption, which is a strong one. Their sixth assumption is weaker than our full column-rank assumption on $\mathbf{B}$.

Although we cannot claim which set of assumptions is weaker, a notable advantage of our Algorithm 2 is that, by linearizing the differentiable part and the augmented item in Lagrangian function, the subproblems are easier to solve and can be solved in parallel.

8.2 ADMM for Nonconvex Sharing and Consensus Problems

The paper [25] studied ADMM for non-convex consensus and sharing problem. Several ADMM algorithms were proposed to solve these two problems and were proved converging to some stationary points. Their analysis covered some variants of ADMM including linearization and flexible block selection.

The assumptions for the non-convex consensus problem are

1) Function f is convex with a compact convex domain.

2) Function $h(\mathbf{x}) = \sum_{i=1}^{p} h_i(\mathbf{x}_i)$ with h_i Lipschitz differentiable.

3) Matrix $\mathbf{A} = [\mathbf{I}; \dots; \mathbf{I}]$, and $\mathbf{B}$ is identity matrix.

Compare the above assumptions with Assumption 1, we can see that the first assumption implies the coercivity of $f(\mathbf{x})+g(\mathbf{y})$ over the feasible set and our assumption of function f being restricted quasi-convex. The second assumption implies that h is Lipschitz differentiable. The third assumption implies $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$. In summary, assumptions on f, $\mathbf{A}$, and $\mathbf{B}$ in Assumption 1 are weaker.

The assumptions for the non-convex sharing problem are as follows.

1) Function $f(\mathbf{x}) = \sum_{i=1}^{p} f_i(\mathbf{x}_i)$, where f_i is either Lipschitz differentiable or convex. Besides, the domain of $f(\mathbf{x})$ is a compact convex set.

2) Function h is Lipschitz differentiable.

3) Matrix $\mathbf{A}_i$ has full column rank, and $\mathbf{B}$ is identity matrix.

In comparison between the above assumptions and Assumption 2, the first and third assumptions implies the coercivity of $f(\mathbf{x}) + g(\mathbf{y})$ over the feasible set and the restricted quasi-convexity of $f(\mathbf{x})$. The second assumption is the same as ours. The third assumption implies $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$. In summary, Assumption 2 on f, $\mathbf{A}$ and $\mathbf{B}$ is weaker.

Compared with this reference, in general our assumptions are broader, and the algorithms we study can work with convergence guarantee for the consensus and sharing problems satisfying the corresponding assumptions. Besides, our algorithm allows updating $\mathbf{x}_i$ in parallel, even though the coupled parts $f(\mathbf{x})$ and $\frac{\beta}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}\|_2^2$ are inseparable for all blocks, while the parallel ADMM in the reference requires the problem to be totally separable for every block.

8.3 Multi-block Bregman ADMM

The paper [26] studied the 3-block Bregman ADMM for non-convex optimization. The Bregman ADMM was proved converging to some stationary points in a 3-block case, which could be further extended to a multi-block case.

By setting the Bregman distance to zero, this algorithm degenerates to the standard ADMM, and the assumptions are as follows.

1) Function $f(\mathbf{x}) = \sum_{i=1}^{p} f_i(\mathbf{x}_i)$, where f_i is strongly-convex.

2) Function $h(\mathbf{y})$ is Lipschitz differentiable and lower-bounded. There exists $\beta_0 > 0$ such that $h(\mathbf{y}) - \beta_0 \nabla h(\mathbf{y})$ is lower bounded.

3) Matrix $\mathbf{B}$ is invertible.

In comparison between the above assumptions and Assumption 2, the first and second assumptions imply the coercivity of $f(\mathbf{x}) + g(\mathbf{y})$ over the feasible set, the restricted quasi-convexity of $f(\mathbf{x})$, and the Lipschitz differentiability of $h(\mathbf{y})$. The third assumption implies $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$.

8.4 Regularization in linearization

In the two linearized ADMM algorithms we study, a square regularization term is added. In this way we get rid of the strong assumptions on $\mathbf{A}_i$ for $i = 1, \ldots, K$. Specifically, [25] requires every $\mathbf{A}_i$ to have full column rank in the non-convex consensus problem, and [24] requires $\mathbf{A}_i$ to have a complex relationship with function f, which in most cases is equivalent to the requirements in [25]. In comparison, we have no assumptions on $\mathbf{A}_i$.

9 Conclusion

In this work we study linearized ADMM algorithms for non-convex optimization problems with non-smooth objective function. First we provide convergence analysis for a two-block linearized ADMM algorithm under Assumption 1. Then we propose a multi-block parallel ADMM algorithm which can update coupled variables in parallel and render subproblems easier to solve, and prove its convergence under Assumption 2. By arguing that both Assumption 1 and Assumption 2 are not only plausible, but also relatively broad compared to other recent works on ADMM for non-convex optimization, we show that the algorithms and their convergence analyses are general enough to work for many interesting problems such as integer programming.

10 Appendix

10.1 Proof of Lemma 1

Define l as the straight line starting from $\mathbf{y}'$ to $\mathbf{y}$, then

$$\begin{aligned} h(\mathbf{y}) - h(\mathbf{y}') &= \int_l \langle \nabla h(\mathbf{s}), \mathrm{d}\mathbf{s} \rangle \\ &= \int_l \langle \nabla h(\mathbf{s}) - \nabla h(\mathbf{z}), \mathrm{d}\mathbf{s} \rangle + \int_l \langle \nabla h(\mathbf{z}), \mathrm{d}\mathbf{s} \rangle \\ &\geq \int_l -L_h |\langle \mathbf{s} - \mathbf{z}, \mathrm{d}\mathbf{s} \rangle| + \int_l \langle \nabla h(\mathbf{z}), \mathrm{d}\mathbf{s} \rangle \\ &= -\frac{L_h}{2} \|\mathbf{y} - \mathbf{y}'\|^2 + \left\langle \mathbf{y} - \mathbf{y}', \nabla h(\mathbf{z}) \right\rangle . \end{aligned}$$

10.2 Proof of Lemma 2

By the $\boldsymbol{\gamma}$-updating rule and the assumption $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$, for two integers $l > k$, we have

$$\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k = \sum_{i=k+1}^{l} \beta(\mathbf{A}\mathbf{x}^i + \mathbf{B}\mathbf{y}^i) \in \mathbf{Im}(\mathbf{B}).$$

Because $\mathbf{B} \in \mathbb{R}^{n \times q}$ has full column rank, there exists $\mathbf{R} \in \mathbb{R}^{q \times q}, \mathbf{Q} \in \mathbb{R}^{q \times n}$ such that $\mathbf{R}$ is invertible, $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}_{n \times n}$, and $\mathbf{B}^{\mathrm{T}} = \mathbf{R}\mathbf{Q}$. Noticing that $\mathbf{Im}(\mathbf{B}) = \mathbf{Im}(\mathbf{Q}^{\mathrm{T}})$, we get $\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k \in \mathbf{Im}(\mathbf{Q}^{\mathrm{T}})$. Thus, $\|\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k\|^2 = \|\mathbf{Q}(\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k)\|^2$. Consequently, we have

$$\begin{aligned} \|\mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k)\|^2 &= \|\mathbf{R}\mathbf{Q}(\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k)\|^2 \\ &\geq \lambda_{\mathbf{R}^{\mathrm{T}}\mathbf{R}} \|\mathbf{Q}(\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k)\|^2 \\ &= \lambda_{\mathbf{R}^{\mathrm{T}}\mathbf{R}} \|\boldsymbol{\gamma}^l - \boldsymbol{\gamma}^k\|^2, \end{aligned}$$

where $\lambda_{\mathbf{R}^{\mathrm{T}}\mathbf{R}}$ denotes the minimum eigenvalue of $\mathbf{R}^{\mathrm{T}}\mathbf{R}$. Therefore, any $C_{\mathbf{R}} > \frac{1}{\lambda_{\mathbf{R}^{\mathrm{T}}\mathbf{R}}}$ satisfies the requirement.

10.3 Proof of Lemma 3

By calculating the derivative of $\bar{h}(\mathbf{y})$ defined in (7), we have

$$\nabla \bar{h}(\mathbf{y}) = \nabla h(\mathbf{y}^k) + L_{\bar{h}}(\mathbf{y} - \mathbf{y}^k) + \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta \mathbf{B}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}).$$

Plug $\mathbf{y} = \mathbf{y}^{k+1}$ into it, and by the $\mathbf{y}$-updating rule we have

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta\mathbf{B}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}) = -\nabla h(\mathbf{y}^k) - L_{\bar{h}}(\mathbf{y}^{k+1} - \mathbf{y}^k). \tag{22}$$

Besides, by the $\boldsymbol{\gamma}$-updating rule, we have

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^{k+1} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta\mathbf{B}^{\mathrm{T}}(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}). \tag{23}$$

By replacing the RHS of (23) with (22), we get

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^{k+1} = -\nabla h(\mathbf{y}^k) - L_{\bar{h}}(\mathbf{y}^{k+1} - \mathbf{y}^k). \tag{24}$$

10.4 Proof of Lemma 5

According to that $\bar{h}(\mathbf{y})$ is $L_{\bar{h}}$-convex, by Proposition 4.8 in [42] we have

$$\bar{h}(\mathbf{y}^k) \geq \bar{h}(\mathbf{y}^{k+1}) + \left\langle \mathbf{y}^k - \mathbf{y}^{k+1}, \nabla\bar{h}(\mathbf{y}^{k+1}) \right\rangle + \frac{L_{\bar{h}}}{2}\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2.$$

According to the updating rule of $\mathbf{y}$, i.e., $\nabla\bar{h}(\mathbf{y}^{k+1}) = \mathbf{0}$, the above inequality is reshaped to

$$\bar{h}(\mathbf{y}^k) \geq \bar{h}(\mathbf{y}^{k+1}) + \frac{L_{\bar{h}}}{2}\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2. \tag{25}$$

Recall that $h(\mathbf{y})$ is Lipschitz-differentiable, and by Lemma 1 we have

$$h(\mathbf{y}^k) \geq h(\mathbf{y}^{k+1}) + \left\langle \mathbf{y}^k - \mathbf{y}^{k+1}, \nabla h(\mathbf{y}^k) \right\rangle - \frac{L_h}{2}\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2. \tag{26}$$

Now we consider the descent of L_β in $\mathbf{y}$-updating step.

$$\begin{aligned} &L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^k) \\ =\ & h(\mathbf{y}^k) - h(\mathbf{y}^{k+1}) + \left\langle \boldsymbol{\gamma}^k, \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \right\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^k\|^2 - \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}\|^2. \end{aligned} \tag{27}$$

By plugging (26) into (27), we have

$$\begin{aligned} \text{RHS of (27)} \geq & \left\langle \mathbf{y}^k - \mathbf{y}^{k+1}, \nabla h(\mathbf{y}^k) \right\rangle - \frac{L_h}{2}\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2 \\ & + \langle \boldsymbol{\gamma}^k, \mathbf{B}(\mathbf{y}^k - \mathbf{y}^{k+1}) \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^k\|^2 - \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}\|^2. \end{aligned} \tag{28}$$

By the definition of $\bar{h}(\mathbf{y})$ in (7), we further derive

$$\text{RHS of (28)} \geq \bar{h}(\mathbf{y}^k) - \bar{h}(\mathbf{y}^{k+1}) + \frac{L_{\bar{h}} - L_h}{2}\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2. \tag{29}$$

By inserting (25) into (29), we finally reach

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^k) \geq C_0\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2,$$

where

$$C_0 := \frac{2L_{\bar{h}} - L_h}{2} \tag{30}$$

and the proof is completed. We will show that C_0 is positive in 10.7.

10.5 Proof of Lemma 6

By definition the ascent of L_β is

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k}) = \left\langle \boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^{k}, \mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1} \right\rangle. \tag{31}$$

By inserting the $\boldsymbol{\gamma}$-updating rule in (31) and applying Lemma 2, we have

$$\begin{aligned} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k}) =& \frac{1}{\beta}\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^{k}\|^2 \\ \leq& \frac{C_{\mathbf{R}}}{\beta}\|\mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^{k})\|^2. \end{aligned} \tag{32}$$

By Lemma 3 and AM-GM Inequality we have

$$\begin{aligned} \|\mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\gamma}^{k})\|^2 =& \|\nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1}) + L_{\bar{h}}(\mathbf{y}^{k+1} - \mathbf{y}^k) - L_{\bar{h}}(\mathbf{y}^k - \mathbf{y}^{k-1})\|^2 \\ \leq& 3\Big(\|\nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1})\|^2 + L_{\bar{h}}^2\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \quad + L_{\bar{h}}^2\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2 \Big). \end{aligned} \tag{33}$$

Because h is Lipschitz differentiable, we have

$$\|\nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1})\|^2 \leq L_h^2\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2,$$

and together with (32) and (33) we have

$$L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k}) \leq \ C_1\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + C_2\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2,$$

where

$$C_1 := \frac{3C_{\mathbf{R}}L_{\bar{h}}^2}{\beta}, \tag{34}$$

$$C_2 := \frac{3C_{\mathbf{R}}(L_h^2 + L_{\bar{h}}^2)}{\beta}. \tag{35}$$

10.6 Proof of Lemma 7

By using Lemma 4, Lemma 5, and Lemma 6, we have

$$\begin{aligned} & L_\beta(\mathbf{x}^{k}, \mathbf{y}^{k}, \boldsymbol{\gamma}^{k}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) \\ \geq\ & L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k}, \boldsymbol{\gamma}^{k}) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) \\ \geq\ & L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k}) + C_0\|\mathbf{y}^k - \mathbf{y}^{k+1}\|^2 - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}) \\ \geq\ & (C_0 - C_1)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 - C_2\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2. \end{aligned} \tag{36}$$

By combining (36) with the definition of m_k, we have

$$m_k - m_{k+1} \geq (C_0 - C_1 - C_3)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + (C_3 - C_2)\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2. \tag{37}$$

By the definition of C_0, C_1, and C_2, if we choose $L_{\bar{h}}$, C_3, and β satisfying

$$C_0 - C_1 - C_3 = \frac{2L_{\bar{h}} - L_h}{2} - \frac{3C_{\mathbf{R}}L_{\bar{h}}^2}{\beta} - C_3 > 0, \tag{38}$$

$$C_3 - C_2 = C_3 - \frac{3C_{\mathbf{R}}(L_h^2 + L_{\bar{h}}^2)}{\beta} > 0, \tag{39}$$

then $\{m_k\}$ is monotonically decreasing. We will verify that (38) and (39) can be satisfied in Appendix 10.7.

Next we will argue that $\{m_k\}$ is also lower bounded. By the assumption $\mathbf{Im}(\mathbf{A}) \subset \mathbf{Im}(\mathbf{B})$, there exists $\mathbf{y}'_k$ such that $\mathbf{B}\mathbf{y}'_k = -\mathbf{A}\mathbf{x}^k$, so we have

$$m_k = f(\mathbf{x}^k) + h(\mathbf{y}^k) + \langle \boldsymbol{\gamma}^k, \mathbf{B}(\mathbf{y}^k - \mathbf{y}'_k)\rangle + \frac{\beta}{2}\|\mathbf{B}(\mathbf{y}^k - \mathbf{y}'_k)\|^2 + C_3\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2. \tag{40}$$

By applying Lemma 3 to the third item in the RHS of (40), we have

$$\begin{aligned}\langle \boldsymbol{\gamma}^k, \mathbf{B}(\mathbf{y}^k - \mathbf{y}'_k)\rangle &= \langle \mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}^k, \mathbf{y}^k - \mathbf{y}'_k\rangle \\ &= \langle -\nabla h(\mathbf{y}^{k-1}) - L_{\bar{h}}(\mathbf{y}^k - \mathbf{y}^{k-1}), \mathbf{y}^k - \mathbf{y}'_k\rangle \\ &= \langle \nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1}) - L_{\bar{h}}(\mathbf{y}^k - \mathbf{y}^{k-1}), \mathbf{y}^k - \mathbf{y}'_k\rangle - \langle \nabla h(\mathbf{y}^k), \mathbf{y}^k - \mathbf{y}'_k\rangle.\end{aligned} \tag{41}$$

By AM-GM Inequality, we bound the first item in the RHS of (41)

$$\begin{aligned}&\langle \nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1}) - L_{\bar{h}}(\mathbf{y}^k - \mathbf{y}^{k-1}), \mathbf{y}^k - \mathbf{y}'_k\rangle \\ \geq& -\frac{1}{2}\left(\|\nabla h(\mathbf{y}^k) - \nabla h(\mathbf{y}^{k-1})\|^2 + \|\mathbf{y}^k - \mathbf{y}'_k\|^2 + L_{\bar{h}}\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2 + L_{\bar{h}}\|\mathbf{y}^k - \mathbf{y}'_k\|^2\right) \\ \geq& -\frac{1}{2}\left((L_h^2 + L_{\bar{h}})\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2 + (L_{\bar{h}} + 1)\|\mathbf{y}^k - \mathbf{y}'_k\|^2\right),\end{aligned} \tag{42}$$

where the last inequality is from the Lipschitz differentiability of $h(\mathbf{y})$.

Considering that $\mathbf{B}$ has full rank and $\|\mathbf{B}\mathbf{z}\|^2 \geq \lambda_{\mathbf{B}^{\mathrm{T}}\mathbf{B}}\|\mathbf{z}\|^2$, for all $\mathbf{z}$, the fourth item in the RHS of (40) can be bounded by

$$\|\mathbf{B}(\mathbf{y}^k - \mathbf{y}'_k)\|^2 \geq \frac{1}{C_{\mathbf{B}}}\|\mathbf{y}^k - \mathbf{y}'_k\|^2, \tag{43}$$

where

$$C_{\mathbf{B}} > \frac{1}{\lambda_{\mathbf{B}^{\mathrm{T}}\mathbf{B}}}.$$

By plugging (41), (42), and (43) into (40), we get

$$m_k \geq Q_1^k + Q_2^k,$$

where

$$\begin{aligned}Q_1^k &:= f(\mathbf{x}^k) + h(\mathbf{y}^k) - \langle \nabla h(\mathbf{y}^k), \mathbf{y}^k - \mathbf{y}'_k\rangle + \frac{1}{2}\left(\frac{\beta}{C_{\mathbf{B}}} - L_{\bar{h}} - 1\right)\|\mathbf{y}^k - \mathbf{y}'_k\|^2, \\ Q_2^k &:= \left(C_3 - \frac{L_{\bar{h}}}{2} - \frac{L_h^2}{2}\right)\|\mathbf{y}^k - \mathbf{y}^{k-1}\|^2.\end{aligned}$$

If both Q_1^k and Q_2^k are lower bounded, the proof will be completed. Let us first check Q_2^k. If we choose $L_{\bar{h}}$ and C_3 satisfying

$$C_3 > \frac{L_{\bar{h}}}{2} + \frac{L_{\bar{h}}^2}{2}, \tag{44}$$

which will be checked in Appendix 10.7, then Q_2^k is lower bounded. For Q_1^k, if we choose $L_{\bar{h}}$ and β satisfying

$$\frac{\beta}{C_{\mathbf{B}}} > L_h + L_{\bar{h}} + 2, \tag{45}$$

then by Lemma 1 we have

$$\begin{aligned} Q_1^k >& f(\mathbf{x}^k) + h(\mathbf{y}^k) - \left\langle \nabla h(\mathbf{y}^k), \mathbf{y}^k - \mathbf{y}'_k \right\rangle + \frac{L_h}{2}\|\mathbf{y}^k - \mathbf{y}'_k\|^2 + \frac{1}{2}\|\mathbf{y}^k - \mathbf{y}'_k\|^2 \\ >& f(\mathbf{x}^k) + h(\mathbf{y}'_k) + \frac{1}{2}\|\mathbf{y}^k - \mathbf{y}'_k\|^2, \end{aligned}$$

where $f(\mathbf{x}^k) + h(\mathbf{y}'_k)$ is lower bounded, so $\{m_k\}$ is lower bounded. Together with its monotonic decrease, we get $\{m_k\}$ is convergent.

10.7 Existence of Parameters

In this subsection, we demonstrate that parameters satisfying our requirements (38), (39), (44), and (45) do exist, and they also guarantee the C_0 defined in Lemma 5 to be positive.

First we recall the definitions of these parameters.

1. L_h is a constant determined by the given function h;
2. β and $L_{\bar{h}}$ are defined in (7), i.e., the Lagrange Function and $\mathbf{y}$-updating rule. They are merely required to be positive;
3. $C_{\mathbf{R}} > \frac{1}{\lambda_{\mathbf{R}^{\mathrm{T}}\mathbf{R}}}$ and $C_{\mathbf{B}} > \frac{1}{\lambda_{\mathbf{B}^{\mathrm{T}}\mathbf{B}}}$ are defined in Lemma 2 and (43), respectively;
4. C_3 first occurs in the definition of m_k, and it only needs to be positive.

Let us take

$$\beta \gg L_{\bar{h}} = \frac{3}{2}C_3 = C_{\mathbf{R}} = C_{\mathbf{B}} \gg L_h,$$

then (38) is simplified to $C_3 > L_h$, and (39) is simplified to $C_3 > 0$. Similarly, (44) and (45) hold evidently, and the C_0 defined in Lemma 5 is positive.

10.8 Proof of Lemma 8

To begin with, there exists $\mathbf{y}'_i$ such that $\mathbf{B}\mathbf{y}'_i = \mathbf{A}_{<i}\mathbf{x}_{<i}^k + \mathbf{A}_i\mathbf{x}_i + \mathbf{A}_{>i}\mathbf{x}_{>i}^k$. Then according to (8), we have

$$\begin{aligned} \bar{f}_i(\mathbf{x}_i) =& \left\langle \mathbf{x}_i - \mathbf{x}_i^k, \nabla_i g(\mathbf{x}^k) + \beta\mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k) \right\rangle + \frac{L_{f_i}}{2}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2 + f_i(\mathbf{x}_i) + \langle \boldsymbol{\gamma}^k, \mathbf{A}_i\mathbf{x}_i \rangle \\ =& Q_3^k + Q_4^k + Q_5^k, \end{aligned}$$

where

$Q_3^k := \dfrac{L_{f_i}}{4}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2 - h(\mathbf{y}_i'),$

$Q_4^k := h(\mathbf{y}_i') + g(\mathbf{x}_{<i}^k, \mathbf{x}_i, \mathbf{x}_{>i}^k) + f_i(\mathbf{x}_i),$

$Q_5^k := \left\langle \mathbf{x}_i - \mathbf{x}_i^k, \nabla_i g(\mathbf{x}^k) + \beta \mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k)\right\rangle + \dfrac{L_{f_i}}{4}\|\mathbf{x}_i - \mathbf{x}_i^k\|^2 + \langle \boldsymbol{\gamma}^k, \mathbf{A}_i\mathbf{x}_i\rangle - g(\mathbf{x}_{<i}^k, \mathbf{x}_i, \mathbf{x}_{>i}^k).$

We will demonstrate that all these three parts are lower bounded and then $\bar{f}_i(\mathbf{x}_i)$ is lower bounded.

Firstly, Q_3^k is lower bounded because $h(\mathbf{y})$ is Lipschitz-differentiable, which means it can always be bounded by quadratic function, and $\mathbf{B}$ has full column rank, if L_{f_i} is chosen large enough. Secondly, Q_4^k is lower bounded by the fourth condition in Assumption 2. Finally, because $g(\mathbf{x})$ is Lipschitz-differentiable, by Lemma 1 we have

$$Q_5^k \geq -g(\mathbf{x}_{<i}^k, \mathbf{x}_i^k, \mathbf{x}_{>i}^k) + \left(\frac{L_{f_i}}{4} - \frac{L_g}{2}\right)\|\mathbf{x}_i - \mathbf{x}_i^k\|^2 + \langle \boldsymbol{\gamma}^k, \mathbf{A}_i\mathbf{x}_i\rangle + \left\langle \mathbf{x}_i - \mathbf{x}_i^k, \beta \mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k)\right\rangle,$$

which is lower bounded if L_{f_i} is chosen larger than $2L_g$. As a result, the subproblem for the update of $\mathbf{x}$ is well defined.

The subproblem of the $\mathbf{y}$ update is a minimization of a quadratic function, so it is well defined as well.

10.9 Proof of Lemma 9

We first decompose the descent in the $\mathbf{x}$-updating step by

$$L_\beta(\mathbf{x}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k) - L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^k, \boldsymbol{\gamma}^k) = \sum_{i=1}^{K} r_i, \tag{46}$$

where

$$r_i := L_\beta\left(\mathbf{x}_{<i}^{k+1}, \mathbf{x}_i^k, \mathbf{x}_{>i}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k\right) - L_\beta\left(\mathbf{x}_{<i}^{k+1}, \mathbf{x}_i^{k+1}, \mathbf{x}_{>i}^k, \mathbf{y}^k, \boldsymbol{\gamma}^k\right), \tag{47}$$

and then verify that r_i is lower bounded by the ith item in the RHS of (18).

In order to shorten the formulations, we denote

$$\Delta\mathbf{x}_i := \mathbf{x}_i^{k+1} - \mathbf{x}_i^k,$$

and

$$\begin{aligned}\mathbf{A}\mathbf{x}(m,n) &:= \mathbf{A}_{<i}\mathbf{x}_{<i}^{k+m} + \mathbf{A}_i\mathbf{x}_i^{k+n} + \mathbf{A}_{>i}\mathbf{x}_{>i}^k,\\ g(m,n) &:= g\left(\mathbf{x}_{<i}^{k+m}, \mathbf{x}_i^{k+n}, \mathbf{x}_{>i}^k\right),\end{aligned}$$

where $m, n \in \{0, 1\}$. According to the above notations and the definition of function L_β in (5), we have

$$r_i = g(1,0) - g(1,1) + f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) - \left\langle \boldsymbol{\gamma}^k, \mathbf{A}_i\Delta\mathbf{x}_i\right\rangle + \frac{\beta}{2}Q_6, \tag{48}$$

where

$$Q_6 := \|\mathbf{A}\mathbf{x}(1,0) + \mathbf{B}\mathbf{y}^k\|^2 - \|\mathbf{A}\mathbf{x}(1,1) + \mathbf{B}\mathbf{y}^k\|^2$$
$$= \|\mathbf{A}_i\Delta\mathbf{x}_i\|^2 - 2\left\langle \mathbf{A}\mathbf{x}(1,1) + \mathbf{B}\mathbf{y}^k, \mathbf{A}_i\Delta\mathbf{x}_i \right\rangle, \tag{49}$$

which is obtained by the cosine theorem $\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2 = \|\mathbf{a} - \mathbf{b}\|^2 - 2\langle \mathbf{b}, \mathbf{b} - \mathbf{a}\rangle$.

According to the $\mathbf{x}$-updating rule, there always exists some $\mathbf{d}_i^{k+1} \in \partial f_i(\mathbf{x}_i^{k+1})$ such that

$$\partial \bar{f}_i(\mathbf{x}_i^{k+1}) \ni \nabla_i g(0,0) + L_{f_i}\Delta\mathbf{x}_i + \mathbf{d}_i^{k+1} + \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta\mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}(0,0) + \mathbf{B}\mathbf{y}^k) = \mathbf{0}. \tag{50}$$

The proof of the above claim is postponed to Appendix 10.11. By taking inner product between $\Delta\mathbf{x}_i$ and both sides of the equation (50), we have

$$\left\langle \nabla_i g(0,0) + \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i \right\rangle + L_{f_i}\|\Delta\mathbf{x}_i\|^2 + \left\langle \boldsymbol{\gamma}^k, \mathbf{A}_i\Delta\mathbf{x}_i \right\rangle = -\beta\left\langle \mathbf{A}\mathbf{x}(0,0) + \mathbf{B}\mathbf{y}^k, \mathbf{A}_i\Delta\mathbf{x}_i \right\rangle. \tag{51}$$

By combining (49) and (51), we have

$$-\langle \boldsymbol{\gamma}^k, \mathbf{A}_i\Delta\mathbf{x}_i\rangle + \frac{\beta}{2}Q_6 = \langle \nabla_i g(0,0) + \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i\rangle + L_{f_i}\|\Delta\mathbf{x}_i\|^2 + \frac{\beta}{2}\|\mathbf{A}_i\Delta\mathbf{x}_i\|^2$$
$$- \beta\left\langle \mathbf{A}\mathbf{x}(1,1) - \mathbf{A}\mathbf{x}(0,0), \mathbf{A}_i\Delta\mathbf{x}_i \right\rangle. \tag{52}$$

Plugging (52) into (48) gives

$$r_i = f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) + \left\langle \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i \right\rangle + L_{f_i}\|\Delta\mathbf{x}_i\|^2 + \frac{\beta}{2}\|\mathbf{A}_i\Delta\mathbf{x}_i\|^2 + Q_7 + Q_8, \tag{53}$$

where

$$Q_7 := g(1,0) - g(1,1) + \left\langle \nabla_i g(0,0), \Delta\mathbf{x}_i \right\rangle, \tag{54}$$
$$Q_8 := -\beta\left\langle \mathbf{A}\mathbf{x}(1,1) - \mathbf{A}\mathbf{x}(0,0), \mathbf{A}_i\Delta\mathbf{x}_i \right\rangle. \tag{55}$$

Recall the purpose of this Lemma, we need to show that Q_7 and Q_8 are lowered bounded.

Let us check Q_7 first. Because $\nabla_i g(\mathbf{x})$ is Lipschitz continuous, by AM-GM Inequality we have

$$\langle \nabla_i g(0,0), \Delta\mathbf{x}_i\rangle = \langle \nabla_i g(0,0) - \nabla_i g(1,0), \Delta\mathbf{x}_i\rangle + \langle \nabla_i g(1,0), \Delta\mathbf{x}_i\rangle$$
$$\geq -\frac{L_g^2}{2}\sum_{t=1}^{i-1}\|\Delta\mathbf{x}_t\|^2 - \frac{1}{2}\|\Delta\mathbf{x}_i\|^2 + \langle \nabla_i g(1,0), \Delta\mathbf{x}_i\rangle. \tag{56}$$

Because $g(\mathbf{x})$ is Lipschitz differentiable, by Lemma 1 we have

$$g(1,0) - g(1,1) + \langle \nabla_i g(1,0), \Delta\mathbf{x}_i\rangle \geq -\frac{L_g}{2}\|\Delta\mathbf{x}_i\|^2. \tag{57}$$

By combining (56) and (57), we get

$$Q_7 \geq -\frac{L_g^2}{2}\sum_{t=1}^{i-1}\|\Delta\mathbf{x}_t\|^2 - \frac{1+L_g}{2}\|\Delta\mathbf{x}_i\|^2. \tag{58}$$

Next we check Q_8. By AM-GM Inequality, we have

$$\begin{aligned}\frac{1}{\beta}Q_8 =& - \langle \mathbf{A}_{<i}\mathbf{x}_{<i}^{k+1} - \mathbf{A}_{<i}\mathbf{x}_{<i}^{k} + \mathbf{A}_i\Delta\mathbf{x}_i, \mathbf{A}_i\Delta\mathbf{x}_i\rangle \\ \geq& -\frac{1}{2}\sum_{t=1}^{i-1}\|\mathbf{A}_t\Delta\mathbf{x}_t\|^2 - \frac{i+1}{2}\|\mathbf{A}_i\Delta\mathbf{x}_i\|^2 \\ \geq& -\frac{1}{2}\sum_{t=1}^{i-1}L_{\mathbf{A}_t}\|\Delta\mathbf{x}_t\|^2 - \frac{i+1}{2}L_{\mathbf{A}_i}\|\Delta\mathbf{x}_i\|^2. \end{aligned} \tag{59}$$

Plug (58) and (59) into (53), and we get

$$r_i \geq \sum_{t=1}^{i-1} C_{4,t}\|\Delta\mathbf{x}_t\|^2 + C_{5,i}\|\Delta\mathbf{x}_i\|^2 + \frac{\beta}{2}\|\mathbf{A}_i\Delta\mathbf{x}_i\|^2 + f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) + \left\langle \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i \right\rangle, \tag{60}$$

where

$$\begin{aligned} C_{4,t} :=& -\frac{L_g^2 + \beta L_{\mathbf{A}_t}}{2}, \\ C_{5,i} :=& L_{f_i} - \frac{1 + L_g + (i+1)\beta L_{\mathbf{A}_i}}{2}. \end{aligned}$$

Plug (60) into (46), and we have

$$\sum_{i=1}^{K} r_i \geq \sum_{i=1}^{K}\sum_{t=1}^{i-1} C_{4,t}\|\Delta\mathbf{x}_t\|^2 + \sum_{i=1}^{K} C_{5,i}\|\Delta\mathbf{x}_i\|^2 + \sum_{i=1}^{K}\left(f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) + \left\langle \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i \right\rangle \right), \tag{61}$$

where the item contains $\|\mathbf{A}_i\Delta\mathbf{x}_i\|^2 \geq 0$ is dropped.

By reshaping (61), we complete the proof

$$\sum_{i=1}^{K} r_i \geq \sum_{i=1}^{K}\left(\frac{\bar{\alpha}_i}{2}\|\Delta\mathbf{x}_i\|^2 + f_i(\mathbf{x}_i^k) - f_i(\mathbf{x}_i^{k+1}) + \left\langle \mathbf{d}_i^{k+1}, \Delta\mathbf{x}_i \right\rangle \right),$$

where

$$\frac{\bar{\alpha}_i}{2} = (K-i)C_{4,i} + C_{5,i}$$

is the definition in (19).

10.10 Proof of Theorem 5

By (50) there exists $\mathbf{d}_i^{k+1} \in \partial f_i(\mathbf{x}_i^{k+1})$ such that

$$\nabla_i g(\mathbf{x}^k) + L_{f_i}(\Delta\mathbf{x}_i) + \mathbf{d}_i^{k+1} + \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\gamma}^k + \beta\mathbf{A}_i^{\mathrm{T}}(\mathbf{A}\mathbf{x}^k + \mathbf{B}\mathbf{y}^k) = \mathbf{0}. \tag{62}$$

We further define

$$\bar{\mathbf{d}}_i^{k+1} := \nabla_i g(\mathbf{x}^{k+1}) + \mathbf{d}_i^{k+1} + \mathbf{A}_i^{\mathrm{T}}\boldsymbol{\gamma}^{k+1} + \beta\mathbf{A}_i^{\mathrm{T}}\big(\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}\big), \tag{63}$$

which, one may readily check, satisfies

$$\bar{\mathbf{d}}_i^{k+1} \in \partial_{\mathbf{x}_i} L_\beta(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}, \boldsymbol{\gamma}^{k+1}).$$

According to the second claim in Theorem 4, we have

$$\lim_{k\to+\infty} \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^{k+1} = \lim_{k\to+\infty} \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^{k}. \tag{64}$$

By the first claim in Theorem 4, we have

$$\lim_{k\to+\infty} \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}) = \lim_{k\to+\infty} \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}^{k} + \mathbf{B}\mathbf{y}^{k}). \tag{65}$$

By the first claim in Theorem 4 and the Lipschitz continuity of ∇g, we have $\lim_{k\to+\infty} L_{f_i}(\Delta \mathbf{x}_i) = \mathbf{0}$ and

$$\lim_{k\to+\infty} \nabla_i g(\mathbf{x}^{k+1}) = \lim_{k\to+\infty} \nabla_i g(\mathbf{x}^{k}). \tag{66}$$

Plug (62), (64), (65), and (66) into (63), and we complete the proof by

$$\begin{aligned} \lim_{k\to+\infty} \bar{\mathbf{d}}_i^{k+1} &= \lim_{k\to+\infty} \nabla_i g(\mathbf{x}^{k+1}) + \mathbf{d}_i^{k+1} + \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^{k+1} + \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}^{k+1} + \mathbf{B}\mathbf{y}^{k+1}) \\ &= \lim_{k\to+\infty} \nabla_i g(\mathbf{x}^{k}) + L_{f_i}(\Delta \mathbf{x}_i) + \mathbf{d}_i^{k+1} + \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^{k} + \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}^{k} + \mathbf{B}\mathbf{y}^{k}) = \mathbf{0}. \end{aligned}$$

10.11 Supplement Proof of Lemma 9

Because $\mathbf{x}_i^{k+1}$ is a minimum point of $\bar{f}_i$, there exists $r > 0$ such that $\forall \mathbf{z} \in B_r(\mathbf{x}_i^{k+1})$, $\bar{f}_i(\mathbf{z}) \geq \bar{f}_i(\mathbf{x}_i^{k+1})$. By plugging the definition of $\bar{f}_i$ (8) into it, we get

$$\begin{aligned} f_i(\mathbf{z}) \geq & f_i(\mathbf{x}_i^{k+1}) + \langle -\nabla_i g(0,0) - L_{f_i} \Delta \mathbf{x}_i - \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^k - \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}(0,0) + \mathbf{B}\mathbf{y}^k), \mathbf{z} - \mathbf{x}_i^{k+1} \rangle \\ & + o(\|\mathbf{z} - \mathbf{x}_i^{k+1}\|). \end{aligned}$$

By the definition of regular subgradient in Definition 2, we get

$$\partial f_i(\mathbf{x}_i^{k+1}) \ni -\nabla_i g(0,0) - L_{f_i} \Delta \mathbf{x}_i - \mathbf{A}_i^{\mathrm{T}} \boldsymbol{\gamma}^k - \beta \mathbf{A}_i^{\mathrm{T}} (\mathbf{A}\mathbf{x}(0,0) + \mathbf{B}\mathbf{y}^k) = \mathbf{d}_i^{k+1},$$

where $\mathbf{d}_i^{k+1}$ is exactly the regular subgradient we need in Lemma 9.

References

[1] Michael Zibulevsky and Michael Elad. ℓ_1-ℓ_2 optimization in signal and image processing. *IEEE Signal Processing Magazine*, 27(3):76–88, 2010.

[2] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[3] Zhi-Quan Luo and Wei Yu. An introduction to convex optimization for communications and signal processing. *IEEE Journal on selected areas in communications*, 24(8):1426–1438, 2006.

[4] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[5] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[6] Jin-Jun Xiao, Shuguang Cui, Zhi-Quan Luo, and Andrea J Goldsmith. Linear coherent decentralized estimation. *IEEE Transactions on Signal Processing*, 56(2):757–770, 2008.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[9] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[10] Rainer Gemulla, Erik Nijkamp, Peter J Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 69–77. ACM, 2011.

[11] Eshref Januzaj, Hans-Peter Kriegel, and Martin Pfeifle. Scalable density-based distributed clustering. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 231–244. Springer, 2004.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[13] Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in neural information processing systems*, pages 2101–2109, 2010.

[14] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Processing*, 62(7):1750–1761, 2014.

[15] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging sciences*, 1(1):143–168, 2008.

[16] Xiaoqun Zhang, Martin Burger, and Stanley Osher. A unified primal-dual algorithm framework based on bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

[17] Chen Feng, Hong Xu, and Baochun Li. An alternating direction method approach to cloud traffic management. *IEEE Transactions on Parallel and Distributed Systems*, 2017.

[18] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, pages 1–25, 2014.

[19] Tian-Yi Lin, Shi-Qian Ma, and Shu-Zhong Zhang. On the sublinear convergence rate of multi-block admm. *Journal of the Operations Research Society of China*, 3(3):251–274, 2015.

[20] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

[21] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):41–76, 1975.

[22] Xinyue Shen, Laming Chen, Yuantao Gu, and HC So. Square-root lasso with nonconvex regularization: An admm approach. *IEEE Signal Processing Letters*, 23(7):934–938, 2016.

[23] Laming Chen and Yuantao Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Transactions on Signal Processing*, 62(15):3754–3767, 2014.

[24] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv*, 2015.

[25] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

[26] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *arXiv*, 2015.

[27] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

[28] Raymond H Chan, Min Tao, and Xiaoming Yuan. Linearized alternating direction method of multipliers for constrained linear least-squares problem. *East Asian Journal on Applied Mathematics*, 2(04):326–341, 2012.

[29] Zhen-Zhen Yang and Zhen Yang. Fast linearized alternating direction method of multipliers for the augmented ℓ_1-regularized problem. *Signal, Image and Video Processing*, 9(7):1601–1612, 2015.

[30] Junfeng Yang and Xiaoming Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 82(281):301–329, 2013.

[31] T Jeong, H Woo, and S Yun. Frame-based poisson image restoration using a proximal linearized alternating direction method. *Inverse Problems*, 29(7):075007, 2013.

[32] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93, 2015.

[33] Georgios B Giannakis, Francis Bach, Raphael Cendrillon, Michael Mahoney, and Jennifer Neville. Signal processing for big data [from the guest editors]. *IEEE Signal Processing Magazine*, 31(5):15–16, 2014.

[34] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[35] Huahua Wang, Arindam Banerjee, and Zhi-Quan Luo. Parallel direction method of multipliers. In *Advances in Neural Information Processing Systems*, pages 181–189, 2014.

[36] Tsung-Hui Chang, Mingyi Hong, and Xiangfeng Wang. Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing*, 63(2):482–497, 2015.

[37] Kai Wang, Jitamitra Desai, and Hongjin He. A note on augmented lagrangian-based parallel splitting method. *Optimization Letters*, 9(6):1199–1212, 2015.

[38] Yue Hu, Eric C Chi, and Genevera I Allen. Admm algorithmic regularization paths for sparse statistical machine learning. In *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 433–459. Springer, 2016.

[39] Bo Wahlberg, Stephen Boyd, Mariette Annergren, and Yang Wang. An admm algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16):83–88, 2012.

[40] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[41] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *arXiv preprint arXiv:1505.03063*, 2015.

[42] Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

[43] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[44] Arif Merchant and Bhaskar Sengupta. Assignment of cells to switches in pcs networks. *IEEE/ACM Transactions on Networking (TON)*, 3(5):521–526, 1995.

[45] Tao Li and Mohammad Shahidehpour. Price-based unit commitment: A case of lagrangian relaxation versus mixed integer programming. *IEEE transactions on power systems*, 20(4):2015–2025, 2005.

[46] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers & operations research*, 13(5):533–549, 1986.

[47] Lucia Pallottino, Eric M Feron, and Antonio Bicchi. Conflict resolution problems for air traffic management systems solved with mixed integer programming. *IEEE transactions on intelligent transportation systems*, 3(1):3–11, 2002.