DISTRIBUTED AND ASYNCHRONOUS ALGORITHMS FOR N-BLOCK CONVEX OPTIMIZATION WITH COUPLING CONSTRAINTS

A PREPRINT

Run Chen

School of Industrial Engineering Purdue University West Lafayette, IN 47906 chen885@purdue.edu

Andrew L. Liu

School of Industrial Engineering Purdue University West Lafayette, IN 47906 andrewliu@purdue.edu

March 26, 2021

1 Introduction

In this work, we focus on designing a distributed algorithm for solving block-separable convex optimization problems with both linear and nonlinear coupling constraints. More specifically, we consider the following problem:

minimize
$$f(\mathbf{x}_1, \dots, \mathbf{x}_N) = \sum_{i=1}^N f_i(\mathbf{x}_i)$$

subject to $\mathbf{x}_i \in \mathcal{X}_i, \quad i = 1, \dots, N,$

$$\sum_{i=1}^N A_i \mathbf{x}_i = \mathbf{b},$$

$$g_j(\mathbf{x}_1, \dots, \mathbf{x}_N) = \sum_{i=1}^N g_{ji}(\mathbf{x}_i) \le 0, \quad j = 1, \dots, M,$$
(1)

where each block of decision variable $\mathbf{x}_i \in \mathbb{R}^{n_i}$ is constrained by a closed and convex set $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ for all $i=1,\ldots,N$, and $\sum_{i=1}^N n_i = n$. The objective function $f:\mathbb{R}^n \to \mathbb{R}$ is block-separable, and each function $f_i:\mathbb{R}^{n_i} \to \mathbb{R}$ is assumed to be continuous and convex for all $i=1,\ldots,N$. All blocks \mathbf{x}_i 's are coupled in a linear equality constraint, where each $A_i \in \mathbb{R}^{m \times n_i}$ is a given matrix for all $i=1,\ldots,N$, and $\mathbf{b} \in \mathbb{R}^m$ is a given vector. All blocks \mathbf{x}_i 's are also coupled in a system of nonlinear inequality constraints, where each constraint function $g_j:\mathbb{R}^n \to \mathbb{R}$ is also block-separable and each function $g_{ji}:\mathbb{R}^{n_i} \to \mathbb{R}$ is assumed to be continuous and convex for all $i=1,\ldots,N$ and $j=1,\ldots,M$. A wide range of application problems can be mathematically formulated as optimization problems of the form (1), arising from the areas including optimal control [14], network optimization [7], statistical learning [2] and etc.

The alternating direction method of multipliers (ADMM) [2], as well as its variants [6, 12, 13], is an efficient distributed algorithm for solving convex block-separable optimization problems with linear coupling constraints, but problems of (1) with nonlinear coupling constraints can not be directly handled by ADMM-typed algorithms.

To overcome the above-mentioned limitations of the ADMM-typed algorithms, we first extend the 2-block Predictor Corrector Proximal Multiplier Method (PCPM) algorithm to solve an N-block convex optimization problem with both linear and nonlinear coupling constraints. We further extend the N-block PCPM algorithm to an asynchronous iterative scheme, where a maximum tolerable delay is allowed for each distributed unit, and apply it to solve an N-block convex optimization problem with general linear coupling constraints.

The remainder of the chapter is organized as follows. In Section 2, we present an extended N-block PCPM algorithm for solving general constrained N-block convex optimization problems. We first establish global convergence under mild assumptions, and then prove the linear convergence rate with slightly stronger assumptions. In Section 3, we further extend the N-block PCPM algorithm to an asynchronous scheme with the bounded delay assumption. We establish both convergence and global sub-linear convergence rate under the conditions of strong convexity. Section 4 presents the numerical results of applying the proposed algorithms to solve a graph optimization problem arising from an application of housing price prediction. Finally, Section 5 concludes this chapter with discussions of the limitations of the algorithms and possible future research directions.

2 Extending the PCPM Algorithm to Solving General Constrained N-block Convex Optimization Problems

2.1 PCPM Algorithm

To present our distributed algorithm, we first briefly describe the original PCPM algorithm [5] to make this paper self-contained. For this purpose, it suffices to consider a 2-block linearly constrained convex optimization problem:

where $f_1: \mathbb{R}^{n_1} \to (-\infty, +\infty]$ and $f_2: \mathbb{R}^{n_2} \to (-\infty, +\infty]$ are closed proper convex functions, $A_1 \in \mathbb{R}^{m \times n_1}$ and $A_2 \in \mathbb{R}^{m \times n_2}$ are full row-rank matrices, $\mathbf{b} \in \mathbb{R}^m$ is a given vector, and $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the corresponding Lagrangian multiplier associated with the linear equality constraint. The classic Lagrangian function $\mathcal{L}: \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^m \to \mathbb{R}$ is defined as:

$$\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \boldsymbol{\lambda}^T (A_1 \mathbf{x}_1 + A_2 \mathbf{x}_2 - \mathbf{b}).$$
(3)

It is well-known that for a convex problem of the specific form in (2) (where the linear constraint qualification automatically holds), finding an optimal solution is equivalent to finding a saddle point $(\mathbf{x}_1^*, \mathbf{x}_2^*, \boldsymbol{\lambda}^*)$ such that $\mathcal{L}(\mathbf{x}_1^*, \mathbf{x}_2^*, \boldsymbol{\lambda}) \leq \mathcal{L}(\mathbf{x}_1^*, \mathbf{x}_2^*, \boldsymbol{\lambda}^*) \leq \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda}^*)$. To find such a saddle point, a simple dual decomposition algorithm can be applied to $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda})$. More specifically, at each iteration k, given a fixed Lagrangian multiplier $\boldsymbol{\lambda}^k$, the primal decision variables $(\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1})$ can be obtained, in parallel, by minimizing $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda}^k)$. Then a dual update $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho(A_1\mathbf{x}_1^{k+1} + A_2\mathbf{x}_2^{k+1} - b)$ is performed.

While the above algorithmic idea is simple, it is well-known that convergence cannot be established without more restrictive assumptions, such as strict convexity of f_1 and f_2 (e.g., Theorem 26.3 in [10]). One approach to overcome such difficulties is the proximal point algorithm, which obtains $(\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1})$ by minimizing the proximal augmented Lagrangian function defined as $\mathcal{L}_{\rho}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda}^k) := \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \boldsymbol{\lambda}^k) + \frac{\rho}{2} \|A_1\mathbf{x}_1 + A_2\mathbf{x}_2 - \mathbf{b}\|_2^2 + \frac{1}{2\rho} \|\mathbf{x}_1 - \mathbf{x}_1^k\|_2^2 + \frac{1}{2\rho} \|\mathbf{x}_2 - \mathbf{x}_2^k\|_2^2$. The parameter ρ is given, which determines the step-size for updating both primal and dual variables in each iteration, and plays a key role in the convergence of the overall algorithm. The primal minimization step now becomes:

$$(\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}) = \operatorname*{argmin}_{\mathbf{x}_1 \in \mathbb{R}^{n_1}, \mathbf{x}_2 \in \mathbb{R}^{n_2}} f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + (\boldsymbol{\lambda}^k)^T (A_1 \mathbf{x}_1 + A_2 \mathbf{x}_2 - \mathbf{b})$$

$$+\frac{\rho}{2} \|A_1 \mathbf{x}_1 + A_2 \mathbf{x}_2 - \mathbf{b}\|_2^2 + \frac{1}{2\rho} \|\mathbf{x}_1 - \mathbf{x}_1^k\|_2^2 + \frac{1}{2\rho} \|\mathbf{x}_2 - \mathbf{x}_2^k\|_2^2.$$
 (4)

With (4), however, \mathbf{x}_1^{k+1} and \mathbf{x}_2^{k+1} can no longer be obtained in parallel due to the augmented term $||A_1\mathbf{x}_1 + A_2\mathbf{x}_2 - \mathbf{b}||_2^2$. To overcome this difficulty, the PCPM algorithm introduces a predictor variable $\boldsymbol{\mu}^{k+1}$:

$$\boldsymbol{\mu}^{k+1} \coloneqq \boldsymbol{\lambda}^k + \rho(A_1 \mathbf{x}_1^k + A_2 \mathbf{x}_2^k - \mathbf{b}). \tag{5}$$

Using the predictor variable, the optimization in (4) can be approximated as:

$$(\mathbf{x}_{1}^{k+1}, \mathbf{x}_{2}^{k+1}) = \underset{\mathbf{x}_{1} \in \mathbb{R}^{n_{1}}, \mathbf{x}_{2} \in \mathbb{R}^{n_{2}}}{\operatorname{argmin}} f_{1}(\mathbf{x}_{1}) + f_{2}(\mathbf{x}_{2}) + (\boldsymbol{\mu}^{k+1})^{T} (A_{1}\mathbf{x}_{1} + A_{2}\mathbf{x}_{2} - \mathbf{b})$$

$$+ \frac{1}{2\rho} \|\mathbf{x}_{1} - \mathbf{x}_{1}^{k}\|_{2}^{2} + \frac{1}{2\rho} \|\mathbf{x}_{2} - \mathbf{x}_{2}^{k}\|_{2}^{2},$$
(6)

which allows \mathbf{x}_1^{k+1} and \mathbf{x}_2^{k+1} to be obtained in parallel again. After solving (6), the PCPM algorithm updates the dual variable as follows:

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho (A_1 \mathbf{x}_1^{k+1} + A_2 \mathbf{x}_2^{k+1} - \mathbf{b}), \tag{7}$$

which is referred to as a corrector update.

2.2 N-block PCPM Algorithm for General Constrained Convex Optimization Problems

In the 2-block PCPM algorithm presented in Section 2.1, we observe that the introduction of the predictor variable eliminates the quadratic term in the proximal augmented Lagrangian function and make the primal minimization step parallelizable again, which is a major difference from the ADMM algorithm. For an *N*-block convex optimization problem with additional nonlinear coupling constraints:

minimize
$$\sum_{i=1}^{N} f_i(\mathbf{x}_i)$$
subject to
$$\mathbf{x}_i \in \mathcal{X}_i, \quad i = 1, \dots, N,$$

$$\sum_{i=1}^{N} A_i \mathbf{x}_i = \mathbf{b}, \quad (\lambda)$$

$$\sum_{i=1}^{N} g_{ji}(\mathbf{x}_i) \leq 0, \quad j = 1, \dots, M, \quad (\mu_j)$$
(8)

the potential coupling caused by the quadratic term $\|\sum_{i=1}^{N} g_{ji}(\mathbf{x}_i)\|_2^2$ could be even worse. Using the same technique of introducing the predictor variable, we extend the 2-block PCPM algorithm for solving (8).

First, we make a blanket assumption on problem (8) throughout this chapter that the Slater's constraint qualification (CQ) holds.

Assumption 2.1 (Slater's CQ). There exists a point $(\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_N)$ such that

$$\left\{ \bar{\mathbf{x}}_i \in relint(\mathcal{X}_i), \quad i = 1, \dots, N \middle| \begin{array}{l} \sum_{i=1}^N A_i \bar{\mathbf{x}}_i = \mathbf{b} \\ \sum_{i=1}^N g_{ji}(\bar{\mathbf{x}}_i) < 0, \quad j = 1, \dots, M \end{array} \right\},$$

where $relint(\mathcal{X}_i)$ denotes the relative interior of the convex set \mathcal{X}_i for all i = 1, ..., N.

To apply the PCPM algorithm to (8), at each iteration k, with a given primal dual pair, $(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k \coloneqq (\mu_1^k, \dots, \mu_M^k)^T)$, we start with a predictor update:

(predictor update):

$$\gamma^{k+1} = \lambda^k + \rho \Big(\sum_{i=1}^N A_i \mathbf{x}_i^k - \mathbf{b} \Big),$$

$$\nu_j^{k+1} = \Pi_{\mathbb{R}_+} \Big[\mu_j^k + \rho \sum_{i=1}^N g_{ji}(\mathbf{x}_i^k) \Big], \quad j = 1, \dots, M,$$

$$(9)$$

where $\Pi_{\mathbb{Z}}(\mathbf{z})$ denotes the projection of a vector $\mathbf{z} \in \mathbb{R}^n$ onto a set $\mathbb{Z} \subset \mathbb{R}^n$, and \mathbb{R}_+ refers to the set of all non-negative real numbers.

After the predictor update, we update the primal variables $(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1})$ by minimizing the Lagrangian function $\mathcal{L}(\mathbf{x}_1,\ldots,\mathbf{x}_N,\gamma^{k+1},\boldsymbol{\nu}^{k+1})$ evaluated at the predictor variable $(\gamma^{k+1},\boldsymbol{\nu}^{k+1})$:= $(\nu_1^{k+1},\ldots,\nu_M^{k+1})^T$), plus the proximal terms. The primal minimization step can be decomposed as

(primal minimization):

$$\mathbf{x}_{i}^{k+1} = \underset{\mathbf{x}_{i} \in \mathcal{X}_{i}}{\operatorname{argmin}} f_{i}(\mathbf{x}_{i}) + (\boldsymbol{\gamma}^{k+1})^{T} A_{i} \mathbf{x}_{i} + \sum_{j=1}^{M} \nu_{j}^{k+1} g_{ji}(\mathbf{x}_{i}) + \frac{1}{2\rho} \|\mathbf{x}_{i} - \mathbf{x}_{i}^{k}\|_{2}^{2},$$

$$i = 1, \dots, N.$$

$$(10)$$

A corrector update is then performed for each Lagrangian multiplier $(\lambda^{k+1}, \mu^{k+1})$:

(dual corrector):

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho \Big(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{b} \Big),$$

$$\boldsymbol{\mu}_j^{k+1} = \Pi_{\mathbb{R}_+} \Big[\boldsymbol{\mu}_j^k + \rho \sum_{i=1}^N g_{ji}(\mathbf{x}_i^{k+1}) \Big], \quad j = 1, \dots, M.$$
(11)

The overall structure of N-block PCPM algorithm is presented in Algorithm 1 below.

Algorithm 1 N-PCPM

- 1: **Initialization** choose an arbitrary starting point $(\mathbf{x}_1^0, \dots, \mathbf{x}_N^0, \boldsymbol{\lambda}^0, \boldsymbol{\nu}^0)$.
- 2: $k \leftarrow 0$.
- 3: while termination conditions are not met do
- 4: (Predictor update)
 - **update** $(\gamma^{k+1}, \nu^{k+1})$ according to (9);
- 5: (Primal minimization)
 - **update** $(\mathbf{x}_1^{k+1}, \dots, \mathbf{x}_N^{k+1})$ according to (10);
- 6: (Corrector update)
 - **update** $(\lambda^{k+1}, \mu^{k+1})$ according to (11);
- 7: $k \leftarrow k + 1$
- 8: return $(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$.

2.3 Convergence Analysis

We make the following additional assumptions on the optimization problem (8).

Assumption 2.2 (Lipschitz Continuity). For all $j=1\ldots M$ and $i=1\ldots N$, each single-valued function $g_{ij}:\mathcal{X}_i\to\mathbb{R}$ is Lipschitz continuous with modulus of L_{ji} , i.e., $\|g_{ji}(\mathbf{x}_1)-g_{ji}(\mathbf{x}_2)\|_2\leq L_{ji}\|\mathbf{x}_1-\mathbf{x}_2\|_2$ for any $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}_i$.

Assumption 2.3 (Existence of a Saddle Point). For the Lagrangian function of (8):

$$\mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \sum_{i=1}^N f_i(\mathbf{x}_i) + \boldsymbol{\lambda}^T \Big(\sum_{i=1}^N A_i \mathbf{x}_i - \mathbf{b} \Big) + \sum_{j=1}^M \mu_j \sum_{i=1}^N g_{ji}(\mathbf{x}_i),$$
(12)

we assume that a saddle point $(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ exists; that is, for any $\mathbf{x}_i \in \mathcal{X}_i$, $i = 1, \dots, N$, $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}_+^M$,

$$\mathcal{L}(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \lambda, \mu) \le \mathcal{L}(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \lambda^*, \mu^*) \le \mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \lambda^*, \mu^*). \tag{13}$$

Note that coupled with the blanket Assumption 2.1 that Slater's CQ holds for the optimization problem (8), the above assumption is equivalent to say that an optimal solution of (8) is assumed to exist (see Corollary 28.3.1 in [10]).

Next, we derive some essential lemmas for constructing the main convergence proof. The following lemma is due to Proposition 6 in [9], and for completeness, we provide the detailed statements below.

Lemma 2.1 (Inequality of Proximal Minimization Point). Given a closed, convex set $\mathbb{Z} \subset \mathbb{R}^n$, and a continuous, convex function $F: \mathbb{Z} \to \mathbb{R}$. With a given point $\bar{\mathbf{z}} \in \mathbb{Z}$ and a positive number $\rho > 0$, if $\hat{\mathbf{z}}$ is a proximal minimization point; i.e. $\hat{\mathbf{z}} := \arg\min_{\mathbf{z} \in \mathbb{Z}} F(\mathbf{z}) + \frac{1}{2\rho} \|\mathbf{z} - \bar{\mathbf{z}}\|_2^2$, then we have that

$$2\rho[F(\widehat{\mathbf{z}}) - F(\mathbf{z})] \le \|\bar{\mathbf{z}} - \mathbf{z}\|_2^2 - \|\widehat{\mathbf{z}} - \mathbf{z}\|_2^2 - \|\widehat{\mathbf{z}} - \bar{\mathbf{z}}\|_2^2, \quad \forall \mathbf{z} \in \mathbb{Z}.$$

$$(14)$$

Proof. Denote $\Phi(\mathbf{z}) = F(\mathbf{z}) + \frac{1}{2\rho} \|\mathbf{z} - \bar{\mathbf{z}}\|_2^2$. By the definition of $\hat{\mathbf{z}}$, we have $\mathbf{0} \in \partial_{\mathbf{z}} \Phi(\hat{\mathbf{z}})$. Since $\Phi(\mathbf{z})$ is strongly convex with modulus $\frac{1}{\rho}$, it follows that $2\rho \left[\Phi(\mathbf{z}) - \Phi(\hat{\mathbf{z}})\right] \geq \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2$ for any $\mathbf{z} \in \mathbb{Z}$.

Lemma 2.2. The update steps (9) and (11) are equivalent to obtaining proximal minimization points as follows:

$$(\boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) = \underset{\boldsymbol{\lambda} \in \mathbb{R}^{m}, \boldsymbol{\mu} \in \mathbb{R}_{+}^{M}}{\operatorname{argmin}} - \mathcal{L}(\mathbf{x}_{1}^{k}, \dots, \mathbf{x}_{N}^{k}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

$$+ \frac{1}{2\rho} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \frac{1}{2\rho} \|\boldsymbol{\mu} - \boldsymbol{\mu}^{k}\|_{2}^{2},$$

$$(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) = \underset{\boldsymbol{\lambda} \in \mathbb{R}^{m}, \boldsymbol{\mu} \in \mathbb{R}_{+}^{M}}{\operatorname{argmin}} - \mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

$$+ \frac{1}{2\rho} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \frac{1}{2\rho} \|\boldsymbol{\mu} - \boldsymbol{\mu}^{k}\|_{2}^{2}.$$

$$(15)$$

Similar to the convergence analysis of the PCPM algorithm in [5], we now establish some fundamental estimates of the distance at each iteration k between the solution point $(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1},\boldsymbol{\lambda}^{k+1},\boldsymbol{\mu}^{k+1})$ and the saddle point $(\mathbf{x}_1^*,\ldots,\mathbf{x}_N^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$.

Proposition 2.3. Let $(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be a saddle point of the optimization problem (8). For all $k \geq 0$, we have that

$$\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} \leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} + 2\rho \left[(\boldsymbol{\lambda}^{*} - \boldsymbol{\gamma}^{k+1})^{T} \sum_{i=1}^{N} A_{i} \mathbf{x}_{i}^{k+1} + \sum_{j=1}^{M} (\mu_{j}^{*} - \nu_{j}^{k+1}) \sum_{i=1}^{N} g_{ji}(\mathbf{x}_{i}^{k+1}) \right]$$
(16)

and

$$\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^*\|_{2}^{2} \leq \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^*\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^*\|_{2}^{2}$$

$$-\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} - \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}$$

$$+2\rho \Big[(\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1})^{T} \sum_{i=1}^{N} A_{i} \mathbf{x}_{i}^{k} + \sum_{j=1}^{M} (\boldsymbol{\nu}_{j}^{k+1} - \boldsymbol{\mu}_{j}^{k+1}) \sum_{i=1}^{N} g_{ji}(\mathbf{x}_{i}^{k})$$

$$+ (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^*)^{T} \sum_{i=1}^{N} A_{i} \mathbf{x}_{i}^{k+1} + \sum_{j=1}^{M} (\boldsymbol{\mu}_{j}^{k+1} - \boldsymbol{\mu}_{j}^*) \sum_{i=1}^{N} g_{ji}(\mathbf{x}_{i}^{k+1}) \Big]$$

$$(17)$$

Proof. The details of the proof are provided in Section A.1.

Theorem 2.4 (Global Convergence). Assume that Assumption 2.1 to Assumption 2.3 hold. Given a scalar $0 < \epsilon < 1$, choose a step size ρ satisfying

$$0 < \rho \le \min\left\{\frac{1 - \epsilon}{A_{max} + ML_{max}}, \frac{1 - \epsilon}{NA_{max}}, \frac{1 - \epsilon}{NL_{max}}\right\},\tag{18}$$

where $A_{max} := \max_{i=1}^N \{ \|A_i\|_2 \}$, and $L_{max} := \max_{j=1}^M \{ \max_{i=1}^N \{ L_{ji} \} \}$. Let $\{ \mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k \}$ be the sequence generated by Algorithm 1, with an arbitrary point $(\mathbf{x}_1^0, \dots, \mathbf{x}_N^0, \boldsymbol{\lambda}^0, \boldsymbol{\mu}^0)$; then the sequence converges globally to a saddle point $(\mathbf{x}_1^* \dots \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of the optimization problem (8).

Proof. Please see Section A.2 for details.

To establish convergence rate, we need to make an additional assumption on Problem (8), as follows.

Assumption 2.4 (Lipschitz Inverse Mapping). Assume that there exists a unique saddle point $(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ such that, given an inverse mapping $\mathcal{S}^{-1} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^M \to \mathbb{R}$:

$$S^{-1}(\mathbf{u}_1,\ldots,\mathbf{u}_N,\mathbf{v})$$

$$= \arg \min_{(\mathbf{x}_1, \dots, \mathbf{x}_N) \in \prod_{i=1}^N \mathcal{X}_i} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}_+^M} \left\{ \mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\lambda}, \boldsymbol{\mu}) - \sum_{i=1}^N \mathbf{x}_i^T \mathbf{u}_i + \boldsymbol{\lambda}^T \mathbf{v} + \boldsymbol{\mu}^T \mathbf{w} \right\},$$
(19)

and a fixed positive real number $\tau > 0$, we have

$$\sum_{i=1}^{N} \|\mathbf{x}_{i} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu} - \boldsymbol{\mu}^{*}\|_{2}^{2} \le a^{2} \left(\sum_{i=1}^{N} \|\mathbf{u}_{i}\|_{2}^{2} + \|\mathbf{v}\|_{2}^{2} + \|\mathbf{w}\|_{2}^{2}\right)$$

for some $a \ge 0$, whenever the point $(\mathbf{x}_1, \dots, \mathbf{x}_N, \lambda, \mu) \in \mathcal{S}^{-1}(\mathbf{u}_1, \dots, \mathbf{u}_N, \mathbf{v}, \mathbf{w})$ and

$$\sum_{i=1}^{N} \|\mathbf{u}_i\|_2^2 + \|\mathbf{v}\|_2^2 + \|\mathbf{w}\|_2^2 \le \tau^2.$$

The above assumption states that the inverse mapping S^{-1} is Lipschitz continuous at the origin with modulus a. By Proposition 2 in [8], to obtain a plausible condition for the Lipschitz continuity of S^{-1} at the origin, we appeal to the *strong second-order conditions for optimality* which are comprised of the following properties:

(i) There is a saddle point $(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of the optimization problem (8) such that $\mathbf{x}_i^* \in int(\mathcal{X}_i)$ for all $i=1\dots N$, where $int(\mathcal{X}_i)$ denotes the interior of the convex set \mathcal{X}_i . Moreover, for all $i=1\dots N$ and $j=1,\dots,M$, function $g_{ji}(\mathbf{x}_i)$ is twice continuously differentiable on a neighborhood of \mathbf{x}_i^* .

- (ii) Let $\mathcal J$ denote the set of active constraint indices at the point of $(\mathbf x_1^*,\dots,\mathbf x_N^*)$: $\mathcal J=\left\{j=1,\dots,M\Big|\sum_{i=1}^Ng_{ji}(\mathbf x_i^*)=0\right\}$. Then $\mu_j^*>0$ for all $j\in\mathcal J$, and $\left\{\sum_{i=1}^N\nabla_{\mathbf x_i}g_{ji}(\mathbf x_i^*)\right\}_{j\in\mathcal J}\cup\left\{\sum_{i=1}^NA_i\mathbf x_i^*\right\}$ forms a linearly independent set.
- (iii) The Hessian matrix $H \coloneqq \nabla^2_{(\mathbf{x}_1, \dots, \mathbf{x}_N)} \mathcal{L}(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ satisfies $\mathbf{y}^T H \mathbf{y} > 0$ for any

$$\mathbf{y} \coloneqq \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix} \in \left\{ \mathbf{y} \neq \mathbf{0} \middle| \begin{array}{l} \sum_{i=1}^N A_i \mathbf{y}_i = \mathbf{0} \\ \sum_{i=1}^N \mathbf{y}_i^T \left[\nabla_{\mathbf{x}_i} g_{ji}(\mathbf{x}_i^*) \right] = 0, \quad \forall j \in \mathcal{J} \end{array} \right\}.$$

We now establish the linear convergence rate of Algorithm 1.

Theorem 2.5 (Linear Convergence Rate). Assume that Assumption 2.1 to Assumption 2.4 hold. Let ρ satisfy (18) and let $\{\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k,\boldsymbol{\lambda}^k,\boldsymbol{\mu}^k\}$ be a sequence generated by Algorithm 1, with an arbitrary starting point $(\mathbf{x}_1^0,\ldots,\mathbf{x}_N^0,\boldsymbol{\lambda}^0,\boldsymbol{\mu}^0)$; then the sequence converges linearly to the unique saddle point $(\mathbf{x}_1^*,\ldots,\mathbf{x}_N^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$. More specifically, there exists an integer \bar{k} such that, for all $k\geq \bar{k}$, we have:

$$\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2}
\leq \theta^{2} \Big(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2} \Big),$$
(20)

where $0 < \theta < 1$.

Proof. Please see Section A.3 for details.

2.4 Numerical Experiments

Consider the following 20-dimensioned, block-separable convex optimization problem with 17 nonlinear coupling constraints, modeling a decentralized planning of an economic system and suggested by [1]:

$$14x_1^2 + 35x_{15} - 79x_{16} - 92 \le 0$$

$$15x_2^2 + 11x_{15} - 61x_{16} - 54 \le 0$$

$$5x_1^2 + 2x_2 + 9x_{17}^4 - x_{18} - 68 \le 0$$

$$x_1^2 - x_2 + 19x_{19} - 20x_{20} + 19 \le 0$$

$$7x_1^2 + 5x_2^2 + x_{19}^2 - 30x_{20} \le 0$$
(21)

A minimum function value of 133.723 can be obtained at the point of (2.18, 2.35, 8.77, 5.07, 0.99, 1.43, 1.33, 9.84, 8.29, 8.37, 2.28, 1.36, 6.08, 14.17, 1.00, 0.66, 1.47, 2.00, 1.05, 2.06). Applying Algorithm 1 and decomposing the original problem into 19 small sub-problems, we achieve the following convergence results, presented in Figure 1.

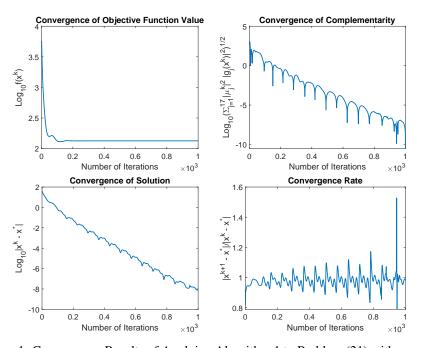


Figure 1: Convergence Results of Applying Algorithm 1 to Problem (21) with $\rho = 0.009$.

Next, by replacing the term of x_{16}^4 in the objective function of (21) with x_{16}^2 and the term of x_{17}^4 in the 15-th constraint with x_{17}^2 , we make the modified problem satisfy Assumption 2.4. A minimum function value of 133.687 can be obtained at the point of (2.18, 2.34, 8.76, 5.07, 0.99, 1.43, 1.34, 9.84, 8.30, 8.36, 2.27, 1.36, 6.08, 14.17, 1.00, 0.64, 2.00, 2.00, 1.04, 2.06). An additional linear convergence rate of applying Algorithm 1 is observed in the following convergence results, presented in Figure 2.

¹The solution is obtained using the nonlinear constrained optimization solver **filter** of Neos Solver at https://neos-server.org/neos/solvers/.

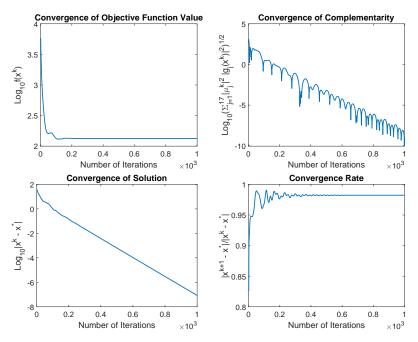


Figure 2: Convergence Results of Applying Algorithm 1 to the Modified Problem (21) with $\rho = 0.009$.

3 Extending the N-block PCPM Algorithm to an Asynchronous Scheme

As a starting point, we first consider the following N-block convex optimization problem with only linear coupling constraints:

minimize
$$\sum_{i=1}^{N} f_i(\mathbf{x}_i)$$
subject to
$$\mathbf{x}_i \in \mathcal{X}_i, \quad i = 1 \dots N,$$

$$\sum_{i=1}^{N} A_i \mathbf{x}_i = \mathbf{b}.$$
(22)

The decision variables, the objective function and the constraints are the same as in the optimization problem (1). When applying Algorithm 1 to solving the above problem, each iteration can be interpreted as main-worker paradigm [11], shown in Figure 3. At each iteration k, a predictor update of γ^{k+1} is first performed on a main processor and is broadcast to each worker processor, which is called a pre-processing task. Upon receiving the updated predictor variable from the main processor, each worker processor solves the decomposed sub-problem in parallel and send its updated primal decision variable \mathbf{x}_i^{k+1} back to the main processor, which is called a worker task. After gathering all updated decision variables, a corrector update is then performed on the main processor, which is called a post-processing task.

The speed of the algorithm is significantly limited by the slowest worker processor, since the post-processing task can not start until all worker tasks are finished and the results are sent back to the main processor. For large-scale problems, with the number of worker processors increasing, the issue of node synchronization can be a major concern for the performance of synchronous distributed algorithms. While in an asynchronous scheme, the main processor can proceed with only part of worker tasks finished. Figure 4 shows an example of 1 main processor and 4 worker processors with different lengths of computation and communication delays. In the asynchronous scheme, the main processor starts a new iteration whenever receives the results from at least 2 worker processors, which leads to much faster iterations than the synchronous scheme.

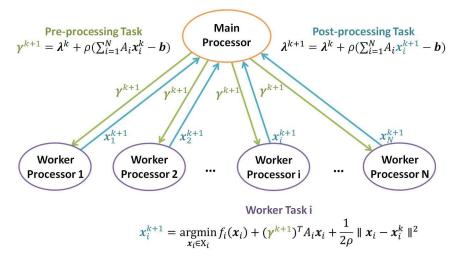


Figure 3: Main-Worker Paradigm for Algorithm 1.

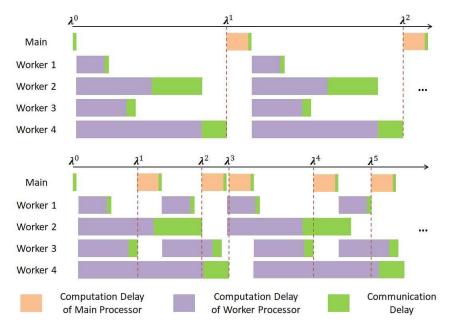


Figure 4: Illustration of how asynchronous scheme iterates faster than synchronous scheme.

In this section, we extend the N-block PCPM algorithm to an asynchronous scheme to solve the linearly constrained N-block convex optimization problem (22).

3.1 Asynchronous N-block PCPM Algorithm for Convex Optimization Problems with Linear Coupling Constraints

To achieve the convergence of the asynchronous N-block PCPM algorithm, similar to [3, 4], we require that the asynchronous delay of each parallel worker processor is bounded. Let $k \geq 0$ denote the iteration index on the main processor. At each iteration k, let $\mathcal{A}_k \subseteq \{1,\ldots,N\}$ denote the subset of worker processors from whom the main processor receives the updated decision variable $\widehat{\mathbf{x}}_i$, and let $\mathcal{A}_k^{\complement} \subseteq \{1,\ldots,N\}$ denote the rest of the worker processors, whose information does not arrive.

Definition 3.1 (Bounded Delay). Let an integer $\tau \geq 1$ denote the maximum tolerable delay. At any iteration $k \geq 0$, with a *bounded delay*, it must holds that $i \in \mathcal{A}_k \cup \mathcal{A}_{k-1} \cup \cdots \cup \mathcal{A}_{k-\tau+1}$ for all $i = 1, \ldots, N$. When $\tau = 1$, it's a synchronous scheme.

At each each iteration k on the main processor, all worker processors are divided into two sets \mathcal{A}_k and $\mathcal{A}_k^{\complement}$, distinguished by whether their information arrives or not at the current moment. Let d_i denote the number of iterations that each worker processor i is delayed. If $d_i < \tau - 1$ for each worker processor $i \in \mathcal{A}_k^{\complement}$, the main processor uses the partially updated decision variables $(\mathbf{x}_1^{k+1}, \dots, \mathbf{x}_N^{k+1})$ to perform both corrector and predictor updates for the Lagrangian multiplier; otherwise, the main processor must wait until the worker processors with $d_i = \tau - 1$ finish their tasks with information received by the master processor. Consequently, new divide of worker processors into \mathcal{A}_k and $\mathcal{A}_k^{\complement}$ is generated, and the bounded delay condition is then satisfied. The overall structure of the Asynchronous N-Block PCPM algorithm for solving the linearly constrained convex optimization problem (22) is presented in Algorithm 2 and Algorithm 3.

Algorithm 2 AN-PCPM (Main Processor)

- 1: **Initialization** choose an arbitrary starting point λ^0 .
- 2: $k \leftarrow 0, d_1, d_2, \dots, d_N \leftarrow 0$
- 3: **wait** until receiving $\{\hat{\mathbf{x}}_i\}_{i=1...N}$
- 4: update:

$$\mathbf{x}_i^0 = \hat{\mathbf{x}}_i, \quad i = 1 \dots N \tag{23}$$

- 5: **broadcast** $\hat{\gamma} = \lambda^0 + \rho(\sum_{i=1}^N A_i \mathbf{x}_i^0 \mathbf{b})$ to all worker processors
- 6: while termination conditions are not met do
- 7: **wait** until receiving $\{\hat{\mathbf{x}}_i\}_{i \in \mathcal{A}_k}$ such that $d_i < \tau$, $\forall i \in \mathcal{A}_k^{\complement}$
- 8: **update**:

$$\mathbf{x}_{i}^{k+1} = \begin{cases} \hat{\mathbf{x}}_{i}, & \forall i \in \mathcal{A}_{k} \\ \mathbf{x}_{i}^{k}, & \forall i \in \mathcal{A}_{k}^{\complement} \end{cases}$$

$$d_{i} = \begin{cases} 0, & \forall i \in \mathcal{A}_{k} \\ d_{i} + 1, & \forall i \in \mathcal{A}_{k}^{\complement} \end{cases}$$
(24)

9: **update**:

$$\lambda^{k+1} = \lambda^k + \rho \left(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{b} \right)$$
 (25)

- 10: **broadcast** $\hat{\gamma} = \lambda^{k+1} + \rho \sum_{i=1}^{N} A_i \mathbf{x}_i^{k+1}$ to the worker processors in A_k
- 11: k + +
- 12: **return** $(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k)$

Algorithm 3 AN-PCPM (Woker Processor)

- 1: **send** $\hat{\mathbf{x}}_i = \mathbf{x}_i^0$ to the main processor
- 2: while not receiving termination signal do
- 3: **wait** until receiving $\hat{\gamma}$
- 4: calculate:

$$\mathbf{y}_i = \underset{\mathbf{x}_i \in \mathcal{X}_i}{\operatorname{argmin}} f_i(\mathbf{x}_i) + \hat{\boldsymbol{\gamma}}^T A_i \mathbf{x}_i + \frac{1}{2\rho} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$
 (26)

- 5: **update**: $\hat{\mathbf{x}}_i = \mathbf{y}_i$
- 6: **send** $\hat{\mathbf{x}}_i$ to the main processor

3.2 Convergence Analysis

Different from the synchronous N-block PCPM algorithm, the convexity of f_i is not enough to achieve the global convergence due to the asynchronous delay in the system. We make the following additional assumption on problem (22).

Assumption 3.1 (Strong Convexity). For all i = 1 ... N, each $f_i : \mathcal{X}_i \to \mathbb{R}$ is a continuous, strongly convex function with modulus $\sigma_i > 0$.

Accordingly, we extend Lemma 2.1 to functions with strong convexity.

Lemma 3.1 (Inequality of Proximal Minimization Point with Strong Convexity). Given a closed, convex set $\mathbb{Z} \subset \mathbb{R}^n$, and a continuous, strongly convex function $F: \mathbb{Z} \to \mathbb{R}$ with modulus σ . With a given point $\bar{\mathbf{z}} \in \mathbb{Z}$ and a positive number $\rho > 0$, if $\hat{\mathbf{z}}$ is a proximal minimization point; i.e. $\hat{\mathbf{z}} := \arg\min_{\mathbf{z} \in \mathbb{Z}} F(\mathbf{z}) + \frac{1}{2\rho} \|\mathbf{z} - \bar{\mathbf{z}}\|_2^2$, then we have that

$$F(\widehat{\mathbf{z}}) - F(\mathbf{z}) \le \frac{1}{2\rho} \|\overline{\mathbf{z}} - \mathbf{z}\|_2^2 - (\frac{\sigma}{2} + \frac{1}{2\rho}) \|\widehat{\mathbf{z}} - \mathbf{z}\|_2^2 - \frac{1}{2\rho} \|\widehat{\mathbf{z}} - \overline{\mathbf{z}}\|_2^2, \quad \forall \mathbf{z} \in \mathbb{Z}.$$
 (27)

Proof. Denote $\Phi(\mathbf{z}) = F(\mathbf{z}) + \frac{1}{2\rho} \|\mathbf{z} - \bar{\mathbf{z}}\|_2^2$. By the definition of $\hat{\mathbf{z}}$, we have $\partial_{\mathbf{z}} \Phi(\hat{\mathbf{z}}) = \mathbf{0}$. Since $\Phi(\mathbf{z})$ is strongly convex with modulus $\sigma + \frac{1}{\rho}$, it follows that $\Phi(\mathbf{z}) - \Phi(\hat{\mathbf{z}}) \geq (\frac{\sigma}{2} + \frac{1}{2\rho}) \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2$ for any $\mathbf{z} \in \mathbb{Z}$. \square

Now, we present the main convergence result.

Theorem 3.2 (Sub-linear Convergence Rate). Let Assumption 2.1, Assumption 2.3 and Assumption 3.1 hold. Choose a step size ρ satisfying:

$$\rho \le \frac{\sigma_{min}}{25N(\tau - 1)^2 A_{max}},\tag{28}$$

where $\sigma_{min} \coloneqq \min_{i=1}^{N} \{\sigma_i\}$. Denote $\bar{\mathbf{x}}_i^k = \frac{1}{k} \sum_{k'=1}^{k} \mathbf{x}_i^{k'}$ for all $i = 1, \dots, N$, where $\{(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k)\}$ is the sequence generated by Algorithm 2 and Algorithm 3, then for all k > 0, it holds that:

$$\left| \sum_{i=1}^{N} f_i(\bar{\mathbf{x}}_i^k) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*) \right| \le \frac{\delta_{\lambda} C_1 + C_2}{k}, \quad \left\| \sum_{i=1}^{N} A_i \bar{\mathbf{x}}^k - \mathbf{b} \right\|_2 \le \frac{C_1}{k}, \tag{29}$$

where $\delta_{\lambda} = \|\lambda^*\|_2$ and C_1 , C_2 are some finite constants.

Proof. Please see Section B.1 for details.

4 Numerical Experiments

4.1 An Optimization Problem on a Graph

In this subsection, we consider an optimization problem on a graph, arising from the training process of regressors with spatial clustering, proposed by [7]. Traditional regressors obtains a parameter vector \mathbf{x} via solving the following optimization problem on a training data set:

$$\underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad \sum_{i=1}^{N} f_i(\mathbf{x}) + r(\mathbf{x}), \tag{30}$$

where N is the number of data points, $\mathcal{X} \subset \mathbb{R}^n$ describes the constraints on the parameter vector, each function $f_i : \mathbb{R}^n \to \mathbb{R}$ denotes the loss function on each training data point for all $i = 1, \ldots, N$, and $r : \mathbb{R}^n \to \mathbb{R}$ denotes some type of regularization function.

When the spatial information is accessible, such as the latitude and longitude data, a map of data points then becomes available. Instead of using a global regressor with a common parameter vector \mathbf{x} for the whole

data set, a local regressor can be built at each data point $i=1,\ldots,N$ with a local parameter vector $\mathbf{x}_i \in \mathbb{R}^n$. Let d_{ij} denote the distance between the data point i and j ($i \neq j$). Different from distributed learning, where a consensus constraint should be satisfied for all local variables, we require that the difference between two local parameter vectors $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ decreases as the distance d_{ij} decreases. Let $\mathcal{N}_{\epsilon}(i)$ denote a set of data points within a neighborhood of point i, i.e., $\mathcal{N}_{\epsilon}(i) \coloneqq \{j=1,\ldots,N|j\neq i,d_{ij}\leq \epsilon\}$. If any data point is regarded as a vertex and any two data points within a neighborhood are connected through an edge, a graph is then constructed as $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where \mathcal{V} denotes the set of vertices with $N=|\mathcal{V}|$ and \mathcal{E} denotes the set of edges with $p=|\mathcal{E}|$. Consider the following optimization problem on the graph:

$$\underset{\mathbf{x}_{1},...,\mathbf{x}_{N}}{\text{minimize}} \quad \sum_{i \in \mathcal{V}} \left[f_{i}(\mathbf{x}_{i}) + r(\mathbf{x}_{i}) \right] + \omega \sum_{(j,k) \in \mathcal{E}} w_{jk} \|\mathbf{x}_{j} - \mathbf{x}_{k}\|_{2}^{2}$$
subject to
$$\mathbf{x}_{i} \in \mathcal{X}_{i}, \quad i = 1,..., N.$$
(31)

Different from [7], we use $\|\cdot\|_2^2$ instead of $\|\cdot\|_2$. The parameter w_{jk} along each edge $(j,k) \in \mathcal{E}$, describing the weight of the penalty term of the difference between the two connected vertices, increases as d_{jk} decreases. The global parameter ω describes the trade-off between minimizing the individual loss function on each data point and agreeing with neighbors. When $\omega=0$, \mathbf{x}_i^* is simply the solution to the optimization problem: minimize $f_i(\mathbf{x}_i)+r(\mathbf{x}_i)$, obtained locally at each vertex i. When $\omega\to+\infty$, the model reduces to a traditional regressor without spatial clustering.

Once the optimal solution $(\mathbf{x}_1^*, \dots, \mathbf{x}_N^*)$ is obtained, for any new node i', the local regressor can be evaluated with the local parameter vector $\mathbf{x}_{i'}$, estimated through the interpolation of the solution:

$$\underset{\mathbf{x}_{i'} \in \mathcal{X}_{i'}}{\text{minimize}} \quad \sum_{j \in \mathcal{N}_{\epsilon}(i')} w_{i'j} \|\mathbf{x}_{i'} - \mathbf{x}_{j}^*\|_{2}^{2}.$$
(32)

4.2 Two Problem Reformulations

To apply N-block PCPM algorithm to solve the graph optimization problem (31), we first need to reformulate it into the block-separable form as in (22).

Similar to [7], for each pair of connected vertices $(\mathbf{x}_j, \mathbf{x}_k)$ along the edge $(j, k) \in \mathcal{E}$, introducing a copy $(\mathbf{z}_{jk}, \mathbf{z}_{kj})$, we can rewrite the graph optimization problem (31) as

• Problem Reformulation 1

minimize
$$\sum_{i \in \mathcal{V}} \left[f_i(\mathbf{x}_i) + r(\mathbf{x}_i) \right] + \omega \sum_{(j,k) \in \mathcal{E}} w_{jk} \|\mathbf{z}_{jk} - \mathbf{z}_{kj}\|_2^2$$
subject to
$$\mathbf{x}_i \in \mathcal{X}_i, \quad \forall i \in \mathcal{V},$$

$$\mathbf{x}_i - \mathbf{z}_{ij} = \mathbf{0}, \quad \forall j \in \mathcal{N}_{\epsilon}(i), \quad \forall i \in \mathcal{V}.$$
(33)

The reformulated problem can be decomposed into N sub-problems on vertices and p sub-problems along edges, using N-block PCPM algorithm. One small issue is that the objective function of the edge sub-problem, $\|\mathbf{z}_{jk} - \mathbf{z}_{kj}\|_2^2$, is not strongly convex.

To overcome this limitation, we propose an alternative way of rewriting (31). For each edge $(j, k) \in \mathcal{E}$, introducing a slack variable $\mathbf{z}_{jk} = \mathbf{x}_j - \mathbf{x}_k$, we can rewrite the graph optimization problem (31) as

• Problem Reformulation 2

minimize
$$\sum_{i \in \mathcal{V}} \left[f_i(\mathbf{x}_i) + r(\mathbf{x}_i) \right] + \omega \sum_{(j,k) \in \mathcal{E}} w_{jk} \|\mathbf{z}_{jk}\|_2^2$$
subject to
$$\mathbf{x}_i \in \mathcal{X}_i, \quad \forall i \in \mathcal{V},$$

$$\mathbf{x}_j - \mathbf{x}_k - \mathbf{z}_{jk} = \mathbf{0}, \quad \forall (j,k) \in \mathcal{E}.$$
(34)

The reformulated problem can also be decomposed into N sub-problems on vertices and p sub-problems along edges. However, in this way of rewriting (31), the objective function of each decomposed sub-problem enjoys the nice property of strong convexity. When applying N-block PCPM algorithm, a linear convergence rate is expected for synchronous iteration scheme, and a sub-linear convergence rate is expected for an asynchronous scheme.

4.3 Housing Price Prediction

In this subsection, we present an application example of the graph optimization problem, where the housing price is predicted based on a set of features, including the number of bedrooms, the number of bathrooms, the number of square feet, and the latitude and longitude of each house. We use the same data set as [7], a list of 985 real estate transactions over a period of one week during May of 2008 in the Greater Sacramento area. All data is standardized with zero mean and unit variance, and all missing data is then set to zero. We randomly select a subset of 193 transactions as our test data set, and use the rest as our training data set.

The graph is constructed based on the latitude and longitude of each house. The rule of selecting neighbors is slightly different from [7]. For each house, we connect it with all the other houses within a distance of 1.0 mile. If the number of connected houses is less then 5, we connect more nearest houses until the number of neighbors reaches 5. The resulting graph has 792 vertices and 4303 edges. Thus, the graph optimization problem can be decomposed into 792 + 4303 = 5095 sub-problems, using N-block PCPM algorithm.

At each data point i = 1...792, the decision variable is $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3})$. The predicted price for each house is:

$$x_{i0} + x_{i1} \times (\text{num_bed})_i + x_{i2} \times (\text{num_bath})_i + x_{i3} \times (\text{num_sq_ft})_i$$

where $(num_bed)_i$, $(num_bath)_i$ and $(num_sq_ft)_i$ are the number of bedrooms, the number of bathrooms and the number of square feets for each house respectively. At each vertex i, the objective function

$$f_i(\mathbf{x}_i) = \|x_{i0} + x_{i1} \times (\text{num_bed})_i + x_{i2} \times (\text{num_bath})_i + x_{i3} \times (\text{num_sq_ft})_i - (\text{price})_i\|_2^2$$

is strongly convex, as well as the regularization function

$$r(\mathbf{x}_i) = \mu(\|x_{i1}\|_2^2 + \|x_{i2}\|_2^2 + \|x_{i3}\|_2^2),$$

where $(\text{price})_i$ is the actual sales price for each house, and μ is a constant regularization parameter, fixed as $\mu = 0.1$.

4.4 Numerical Results of Synchronous N-block PCPM Algorithm

We first apply Algorithm 1 to solve the two reformulated problem (33) and (34), and compare the performance. The convergence results are shown in Figure 5. Due to the strong convexity of the reformulated problem (34), the algorithm converges much faster than solving the reformulated problem (33).

We also plot the mean square error (MSE) on the testing data set using various values of ω , shown in Figure 6. When $\omega=1.0$, a minimum MSE of 0.27 can be obtained on the testing data set. We fixed $\omega=1.0$ for all the numerical experiments.

4.5 Numerical Results of Asynchronous N-block PCPM Algorithm

We apply Algorithm 2 and Algorithm 3 to solve the reformulated problem (34) with a maximum delay $\tau = 4$. The convergence results of are shown in Figure 7. A sub-linear convergence rate is observed.

While implementing the algorithm as a sequential code, we simulate the elapsed wall-clock time on the main processor. As illustrated in Figure 4, the computation delay of the main processor is set as 1.0 second, the computation delay of each worker processor for solving vertex sub-problem is set as 1.2 second,

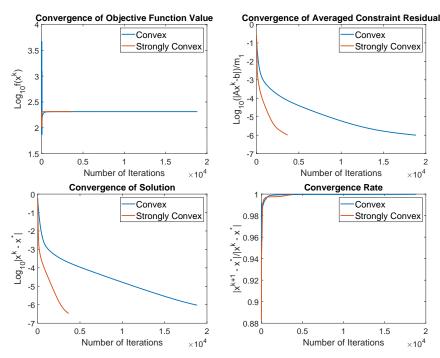


Figure 5: Convergence Results of Applying Algorithm 1 to solve Reformulated Problems (33) and (34) with $\omega=1.0$ and $\rho=0.06$.

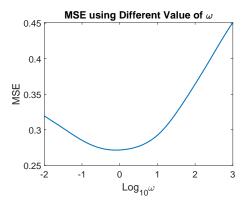


Figure 6: MSE for Testing Data Set with ω Varying from 10^{-2} to 10^3 and $\mu = 0.1$.

the computation delay of each worker processor for solving edge sub-problem is set as 0.6 second, and the communication delay of each worker processor is uniformly drawn from a range of 0.0 to 1.0 second. Under these settings, we simulate the elapsed wall-clock time on the main processor for the different values of maximum delay $\tau=1,2,4,7$, shown in Figure 8. We observe that, with a larger number of maximum delay, the number of iterations, used for the asynchronous algorithm to converge, increases but the simulated elapsed wall-clock time decreases, which implies a faster convergence with more short-time iterations.

5 Conclusion and Future Works

In this paper, we first proposed an N-block PCPM algorithm to solve N-block convex optimization problems with both linear and nonlinear constraints, with global convergence established. A linear convergence rate under the strong second-order conditions for optimality is observed in the numerical experiments. Next,

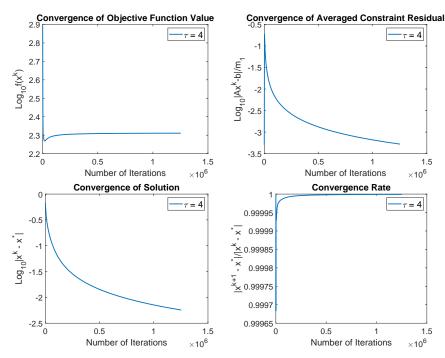


Figure 7: Convergence Results of Applying Algorithm 2 and Algorithm 3 to Solve the Reformulated Problem (34) with $\tau = 4$ and $\rho = 0.0005$.

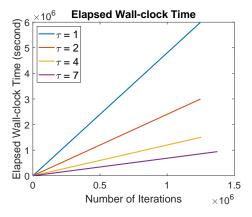


Figure 8: Simulated Elapsed Wall-clock Time on the Main Processor with Various Maximum Delay τ and a Same $\rho = 0.0005$.

for a starting point, we proposed an asynchronous N-block PCPM algorithm to solve linearly constrained N-block convex optimization problems. The numerical results demonstrate the sub-linear convergence rate under the bounded delay assumption, as well as the faster convergence with more short-time iterations than a synchronous iterative scheme.

However, the performance of real asynchronous implementation of N-block PCPM algorithm is unknown, and thus in the future, more experiments (probably with much larger problem sizes) will be conducted on a multi-node computer cluster using MPI functions without blocking communication, such as MPI_Isend and MPI_Irecv. Also, the extension of the asynchronous N-block PCPM algorithm to solve N-block convex optimization problems with both linear and nonlinear constraints is worth to be explored.

References

- [1] J. Asaadi. A computational comparison of some non-linear programs. *Mathematical Programming*, 4(1):144–154, 1973.
- [2] Stephen Boyd, Neal Parikh, and Eric Chu. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [3] Tsung-Hui Chang, Mingyi Hong, Wei-Cheng Liao, and Xiangfeng Wang. Asynchronous distributed ADMM for large-scale optimization—part i: Algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016.
- [4] Tsung-Hui Chang, Wei-Cheng Liao, Mingyi Hong, and Xiangfeng Wang. Asynchronous distributed admm for large-scale optimization—part ii: Linear convergence analysis and numerical performance. *IEEE Transactions on Signal Processing*, 64(12):3131–3144, 2016.
- [5] Gong Chen and Marc Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1-3):81–101, 1994.
- [6] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block ADMM with o (1/k) convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.
- [7] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 387–396. ACM, 2015.
- [8] R Tyrrell Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.
- [9] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
- [10] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 2015.
- [11] Sartaj Sahni and George Vairaktarakis. The master-slave paradigm in parallel computer and industrial settings. *Journal of Global Optimization*, 9(3-4):357–377, 1996.
- [12] Huahua Wang, Arindam Banerjee, and Zhi-Quan Luo. Parallel direction method of multipliers. In *Advances in Neural Information Processing Systems*, pages 181–189, 2014.
- [13] Xiangfeng Wang, Mingyi Hong, Shiqian Ma, and Zhi-Quan Luo. Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers. *arXiv* preprint *arXiv*:1308.5294, 2013.
- [14] Hao Yu and Michael J Neely. A simple parallel algorithm with an o(1/t) convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.

A Proofs in Section 2.3

A.1 Proof of Proposition 2.3

We first prove the inequality (16). From the primal minimization step (10), we know that $(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1})$ is the unique proximal minimization point of the Lagrangian function evaluated at the predictor variable: $\mathcal{L}(\mathbf{x}_1,\ldots,\mathbf{x}_N,\boldsymbol{\gamma}^{k+1},\boldsymbol{\nu}^{k+1})$. Applying Lemma 2.1 with $\hat{\mathbf{z}}=(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1})$, $\bar{\mathbf{z}}=(\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k)$ and

 $\mathbf{z} = (\mathbf{x}_1^*, \dots, \mathbf{x}_N^*)$, we have:

$$2\rho \left[\mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) - \mathcal{L}(\mathbf{x}_{1}^{*}, \dots, \mathbf{x}_{N}^{*}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) \right]$$

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2}.$$
(35)

Since $(\mathbf{x}_1^*,\ldots,\mathbf{x}_N^*,\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*)$ is a saddle point of the Lagrangian function $\mathcal{L}(\mathbf{x}_1,\ldots,\mathbf{x}_N,\boldsymbol{\lambda},\boldsymbol{\mu})$, i.e., $\mathcal{L}(\mathbf{x}_1^*,\ldots,\mathbf{x}_N^*,\boldsymbol{\lambda}^*) \leq \mathcal{L}(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1},\boldsymbol{\lambda}^*)$, we have:

$$2\rho \left[\mathcal{L}(\mathbf{x}_{1}^{*}, \dots, \mathbf{x}_{N}^{*}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) - \mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{*}, \boldsymbol{\mu}^{*}) \right] \leq 0.$$
 (36)

Adding the above two inequalities yields the inequality (16) in Proposition 2.3.

To prove the second inequality, (17), in Proposition 2.3, we use a similar approach as above. By Lemma 2.2, we know that $(\gamma^{k+1}, \boldsymbol{\nu}^{k+1})$ is the unique proximal minimization point of the function $-\mathcal{L}(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}, \boldsymbol{\mu})$. Applying Lemma 2.1 with $\hat{\mathbf{z}} = (\gamma^{k+1}, \boldsymbol{\nu}^{k+1})$, $\bar{\mathbf{z}} = (\boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$ and $\mathbf{z} = (\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1})$, we have:

$$2\rho \left\{ \left[-\mathcal{L}(\mathbf{x}_{1}^{k}, \dots, \mathbf{x}_{N}^{k}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) \right] - \left[-\mathcal{L}(\mathbf{x}_{1}^{k}, \dots, \mathbf{x}_{N}^{k}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) \right] \right\}$$

$$\leq \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{k+1}\|_{2}^{2}$$

$$-\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} - \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}.$$
(37)

By Lemma 2.2, we also know that $(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1})$ is the unique proximal minimization point of the function $-\mathcal{L}(\mathbf{x}_1^{k+1}, \dots, \mathbf{x}_N^{k+1}, \boldsymbol{\lambda}, \boldsymbol{\mu})$. Applying Lemma 2.1 with $\widehat{\mathbf{z}} = (\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1})$, $\bar{\mathbf{z}} = (\boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$ and $\mathbf{z} = (\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, we have:

$$2\rho \left\{ \left[-\mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) \right] - \left[-\mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{*}, \boldsymbol{\mu}^{*}) \right] \right\}$$

$$\leq \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$-\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} - \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2} - \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} - \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}.$$
(38)

Adding the above two inequalities yields the inequality (17) in Proposition 2.3.

A.2 Proof of Theorem 2.4

By adding the two inequalities (16) and (17) in Proposition 2.3, we have:

$$\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$- \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} - \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} - \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}$$

$$+ \sum_{i=1}^{N} \underbrace{2\rho(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\gamma}^{k+1})^{T} A_{i}(\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k})}_{(a)_{i}} + \sum_{j=1}^{M} \sum_{i=1}^{N} \underbrace{2\rho(\boldsymbol{\mu}_{j}^{k+1} - \boldsymbol{\nu}_{j}^{k+1}) \left[g_{ji}(\mathbf{x}_{i}^{k+1}) - g_{ji}(\mathbf{x}_{i}^{k})\right]}_{(b)_{ji}}. (39)$$

Before we continue with the proof, we first show an extension of the Young's inequality² on vector products that will play a key role in the following proof. Given any two vectors $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$, we have that

$$\mathbf{z}_{1}^{T}\mathbf{z}_{2} = \sum_{j=1}^{n} z_{1j}z_{2j} = \sum_{j=1}^{n} \left(\frac{1}{\delta}z_{1j}\right) \left(\delta z_{2j}\right) \leq \sum_{j=1}^{n} \left|\frac{1}{\delta}z_{1j}\right| \left|\delta z_{2j}\right|,$$

where δ is a non-zero real number. Applying Young's inequality on each summation term with p=q=2, we obtain that

$$\mathbf{z}_{1}^{T}\mathbf{z}_{2} \leq \sum_{j=1}^{n} \left[\frac{1}{2} \left(\frac{1}{\delta} z_{1j} \right)^{2} + \frac{1}{2} \left(\delta z_{2j} \right)^{2} \right] = \frac{1}{2\delta^{2}} \|\mathbf{z}_{1}\|_{2}^{2} + \frac{\delta^{2}}{2} \|\mathbf{z}_{2}\|_{2}^{2}. \tag{40}$$

Applying (40) on each term $(a)_i$ yields

$$(a)_{i} \leq 2\rho \left[\frac{1}{2\delta^{2}} \| \boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1} \|_{2}^{2} + \frac{\delta^{2}}{2} \| A_{i} (\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}) \|_{2}^{2} \right]$$

$$\leq 2\rho \left[\frac{1}{2\delta^{2}} \| \boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1} \|_{2}^{2} + \frac{\delta^{2}}{2} \| A_{i} \|_{2}^{2} \| \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k} \|_{2}^{2} \right],$$

$$(41)$$

and letting $\delta^2 = \frac{1}{\|A_i\|_2}$ yields

$$(a)_{i} \leq \rho \|A_{i}\|_{2} \left(\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} \right)$$

$$\leq \rho A_{max} \left(\|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} \right).$$

$$(42)$$

Applying (40) on each term $(b)_{ji}$ yields

$$(b)_{ji} \leq 2\rho \left[\frac{1}{2\delta^{2}} \|\nu_{j}^{k+1} - \mu_{j}^{k+1}\|_{2}^{2} + \frac{\delta^{2}}{2} \|g_{ji}(\mathbf{x}_{i}^{k+1}) - g_{ji}(\mathbf{x}_{i}^{k})\|_{2}^{2} \right]$$

$$\leq 2\rho \left[\frac{1}{2\delta^{2}} \|\nu_{j}^{k+1} - \mu_{j}^{k+1}\|_{2}^{2} + \frac{\delta^{2}}{2} L_{ji}^{2} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} \right],$$

$$(43)$$

and letting $\delta^2 = \frac{1}{L_{ji}}$ yields

$$(b)_{ji} \leq \rho L_{ji} \left(\| \nu_j^{k+1} - \mu_j^{k+1} \|_2^2 + \| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \|_2^2 \right)$$

$$\leq \rho L_{max} \left(\| \nu_j^{k+1} - \mu_j^{k+1} \|_2^2 + \| \mathbf{x}_i^{k+1} - \mathbf{x}_i^k \|_2^2 \right).$$

$$(44)$$

Substituting (42) and (44) into (39) yields

$$\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2}
\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2}
- (1 - \rho A_{max} - \rho M L_{max}) \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} - (1 - \rho N A_{max}) \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2}
- (1 - \rho N L_{max}) \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} - \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} - \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}$$
(45)

Since $0 < \rho \le \min\left\{\frac{1-\epsilon}{A_{max}+ML_{max}}, \frac{1-\epsilon}{NA_{max}}, \frac{1-\epsilon}{NL_{max}}\right\}$, we have:

$$\sum_{i=1}^{N} \lVert \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*} \rVert_{2}^{2} + \lVert \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*} \rVert_{2}^{2} + \lVert \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*} \rVert_{2}^{2}$$

²Young's inequality states that if a and b are two non-negative real numbers, and p and q are real numbers greater than 1 such that $\frac{1}{p} + \frac{1}{q} = 1$, then $ab < \frac{a^p}{p} + \frac{b^q}{q}$.

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2}
-\epsilon \Big(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} + \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2}
+ \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2} \Big).$$
(46)

It implies that for all $k \geq 0$:

$$0 \leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k-1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k-1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k-1} - \boldsymbol{\mu}^{*}\|_{2}^{2}$$

$$\leq \cdots \leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{0} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{0} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{0} - \boldsymbol{\mu}^{*}\|_{2}^{2}.$$

$$(47)$$

It further implies that the sequence $\left\{\sum_{i=1}^{N} \|\mathbf{x}_i^k - \mathbf{x}_i^*\|_2^2 + \|\boldsymbol{\lambda}^k - \boldsymbol{\lambda}^*\|_2^2 + \|\boldsymbol{\mu}^k - \boldsymbol{\mu}^*\|_2^2\right\}$ is monotonically decreasing and bounded below by 0; hence the sequence must be convergent to a limit, denoted by ξ :

$$\lim_{k \to +\infty} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2} = \xi.$$
(48)

Taking the limit on both sides of (46) yields:

$$\lim_{k \to +\infty} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} = 0,$$

$$\lim_{k \to +\infty} \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} = 0, \qquad \lim_{k \to +\infty} \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} = 0,$$

$$\lim_{k \to +\infty} \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} = 0, \qquad \lim_{k \to +\infty} \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2} = 0.$$
(49)

Additionally, (48) also implies that $\{(\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k,\boldsymbol{\lambda}^k,\boldsymbol{\mu}^k)\}$ is a bounded sequence, and thus there exists a sub-sequence $\{(\mathbf{x}_1^{k_j},\ldots,\mathbf{x}_N^{k_j},\boldsymbol{\lambda}^{k_j},\boldsymbol{\mu}^{k_j})\}$ that converges to a limit point $(\mathbf{x}_1^\infty,\ldots,\mathbf{x}_N^\infty,\boldsymbol{\lambda}^\infty,\boldsymbol{\mu}^\infty)$. We next show that the limit point is indeed a saddle point and is also the unique limit point of $\{(\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k,\boldsymbol{\lambda}^k,\boldsymbol{\mu}^k)\}$. Applying Lemma 2.1 with $\widehat{\mathbf{z}}=(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1}), \, \bar{\mathbf{z}}=(\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k)$ and any $\mathbf{z}=(\mathbf{x}_1,\ldots,\mathbf{x}_N)\in\prod_{i=1}^N\mathcal{X}_i,$ we have:

$$2\rho \Big[\mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) - \mathcal{L}(\mathbf{x}_{1}, \dots, \mathbf{x}_{N}, \boldsymbol{\gamma}^{k+1}, \boldsymbol{\nu}^{k+1}) \Big]$$

$$\leq \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2}$$

$$\leq \sum_{i=1}^{N} (\|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{k+1}\|_{2}^{2} + \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}\|_{2}^{2}) - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}\|_{2}^{2} - \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} = 0$$

$$\forall (\mathbf{x}_{1}, \dots, \mathbf{x}_{N}) \in \prod_{i=1}^{N} \mathcal{X}_{i}.$$

$$(50)$$

Taking the limits over an appropriate sub-sequence $\{k_i\}$ on both sides and using (49), we have:

$$\mathcal{L}(\mathbf{x}_1^{\infty}, \dots, \mathbf{x}_N^{\infty}, \boldsymbol{\lambda}^{\infty}, \boldsymbol{\mu}^{\infty}) \leq \mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\lambda}^{\infty}, \boldsymbol{\mu}^{\infty}), \quad \forall (\mathbf{x}_1, \dots, \mathbf{x}_N) \in \prod_{i=1}^N \mathcal{X}_i.$$
 (51)

Similarly, applying Lemma 2.1 with $\hat{\mathbf{z}} = (\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}), \bar{\mathbf{z}} = (\boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$ and any $\mathbf{z} = (\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathbb{R}^{m_2}_+)$, we have:

$$\begin{split} & 2\rho \left[\mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}, \boldsymbol{\mu}) - \mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) \right] \\ \leq & \| \boldsymbol{\lambda}^{k} - \boldsymbol{\lambda} \|^{2} - \| \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda} \|^{2} - \| \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k} \|^{2} \\ & + \| \boldsymbol{\mu}^{k} - \boldsymbol{\mu} \|^{2} - \| \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu} \|^{2} - \| \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k} \|^{2} \\ \leq & \left(\| \boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{k+1} \|^{2} + \| \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda} \|^{2} \right) - \| \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda} \|^{2} - \| \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k} \|^{2} \\ & + \left(\| \boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{k+1} \|^{2} + \| \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu} \|^{2} \right) - \| \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu} \|^{2} - \| \boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k} \|^{2} = 0, \quad \forall \boldsymbol{\mu} \in \mathbb{R}_{+}^{m_{2}}. \end{split} \tag{52}$$

Taking the limits over an appropriate sub-sequence $\{k_i\}$ on both sides and using (49), we have:

$$\mathcal{L}(\mathbf{x}_1^{\infty}, \dots, \mathbf{x}_N^{\infty}, \lambda, \mu) \le \mathcal{L}(\mathbf{x}_1^{\infty}, \dots, \mathbf{x}_N^{\infty}, \lambda^{\infty}, \mu^{\infty}), \quad \forall \mu \in \mathbb{R}_+^{m_2}.$$
 (53)

Therefore, we show that $(\mathbf{x}_1^{\infty}, \dots, \mathbf{x}_N^{\infty}, \boldsymbol{\lambda}^{\infty}, \boldsymbol{\mu}^{\infty})$ is indeed a saddle point of the Lagrangian function $\mathcal{L}(\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\lambda}, \boldsymbol{\mu})$. Then (48) implies that

$$\lim_{k \to +\infty} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{\infty}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{\infty}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{\infty}\|_{2}^{2} = \xi.$$
 (54)

Since we have already argued (after Eq. (49)) that there exists a bounded sequence of $\{(\mathbf{x}_1^k, \dots, \mathbf{x}_N^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)\}$ that converges to 0; that is, there exists $\{k_i\}$ such that

$$\lim_{k_j \to +\infty} \sum_{i=1}^N \|\mathbf{x}_i^{k_j} - \mathbf{x}_i^{\infty}\|_2^2 + \|\boldsymbol{\lambda}^{k_j} - \boldsymbol{\lambda}^{\infty}\|_2^2 + \|\boldsymbol{\mu}^{k_j} - \boldsymbol{\mu}^{\infty}\|_2^2 = 0,$$

which then implies that $\xi=0$. Therefore, we show that $\{(\mathbf{x}_1^k,\ldots,\mathbf{x}_N^k,\boldsymbol{\lambda}^k,\boldsymbol{\mu}^k)\}$ converges globally to a saddle point $(\mathbf{x}_1^\infty,\ldots,\mathbf{x}_N^\infty,\boldsymbol{\lambda}^\infty,\boldsymbol{\mu}^\infty)$.

A.3 Proof of Theorem 2.5

Letting

$$\mathbf{u}_{i}^{k} = A_{i}^{T} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\gamma}^{k+1}) + \sum_{j=1}^{m_{2}} (\mu_{j}^{k+1} - \nu_{j}^{k+1}) \nabla_{\mathbf{x}_{i}} g_{ji}(\mathbf{x}_{i}^{k+1}) - \frac{1}{\rho} (\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}), \quad \forall i = 1 \dots N,$$

$$\mathbf{v}^{k} = -\frac{1}{\rho} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}),$$

$$\mathbf{w}^{k} = -\frac{1}{\rho} (\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k}),$$
(55)

we first show that $(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1},\boldsymbol{\lambda}^{k+1},\boldsymbol{\mu}^{k+1}) \in S^{-1}(\mathbf{u}_1^k,\ldots,\mathbf{u}_N^k,\mathbf{v}^k,\mathbf{w}^k)$. By the primal minimization step (10), we have, for all $i=1\ldots N$:

$$-\nabla_{\mathbf{x}_{i}} f_{i}(\mathbf{x}_{i}^{k+1}) - \underbrace{\left[A_{i}^{T} \boldsymbol{\gamma}^{k+1} + \sum_{j=1}^{M} \nu_{j}^{k+1} \nabla_{\mathbf{x}_{i}} g_{ji}(\mathbf{x}_{i}^{k+1}) + \frac{1}{\rho} (\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k})\right]}_{\Delta_{u_{i}}} \in \mathcal{N}_{\mathcal{X}_{i}}(\mathbf{x}_{i}^{k+1}), \tag{56}$$

where $\mathcal{N}_{\mathcal{X}_i}(\mathbf{x}^{k+1}) \coloneqq \{\mathbf{y} \in \mathbb{R}^{n_1} | \mathbf{y}^T(\mathbf{x} - \mathbf{x}^{k+1}) \leq \mathbf{0}, \forall \mathbf{x} \in \mathcal{X}_i \}$ denotes the normal cone to the set \mathcal{X}_i at the point \mathbf{x}_i^{k+1} for all $i = 1, \dots, N$. Plugging

$$\Delta_{u_i} = A_i^T \boldsymbol{\lambda}^{k+1} + \sum_{i=1}^M \mu_j^{k+1} \nabla_{\mathbf{x}_i} g_{ji}(\mathbf{x}_i^{k+1}) - \mathbf{u}_i^k$$

into the above expression, we have that, for all i = 1, ..., N:

$$-\nabla_{\mathbf{x}_i} f_i(\mathbf{x}_i^{k+1}) - A_i^T \boldsymbol{\lambda}^{k+1} - \sum_{j=1}^M \mu_j^{k+1} \nabla_{\mathbf{x}_i} g_{ji}(\mathbf{x}_i^{k+1}) + \mathbf{u}_i^k \in \mathcal{N}_{\mathcal{X}_i}(\mathbf{x}_i^{k+1}), \tag{57}$$

which implies

$$(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1})$$

$$\in \underset{(\mathbf{x}_{1}, \dots, \mathbf{x}_{N}) \in \prod_{i=1}^{N} \mathcal{X}_{i}}{\operatorname{argmin}} \mathcal{L}(\mathbf{x}_{1}, \dots, \mathbf{x}_{N}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) - \sum_{i=1}^{N} \mathbf{x}_{i}^{T} \mathbf{u}_{i}^{k} + (\boldsymbol{\lambda}^{k+1})^{T} \mathbf{v}^{k} + (\boldsymbol{\mu}^{k+1})^{T} \mathbf{w}^{k}.$$
(58)

Similarly, by the interpretation of $(\lambda^{k+1}, \mu^{k+1})$ in Lemma 2.2, we have:

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) + \underbrace{\left[-\frac{1}{\rho} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}) \right]}_{\mathbf{v}^{k}} = \mathbf{0},$$

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}(\mathbf{x}_{1}^{k+1}, \dots, \mathbf{x}_{N}^{k+1}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) + \underbrace{\left[-\frac{1}{\rho} (\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k}) \right]}_{\mathbf{w}^{k}} \in \mathcal{N}_{\mathbb{R}_{+}^{m_{2}}}(\boldsymbol{\mu}^{k+1}),$$
(59)

which imply

$$(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\mu}^{k+1}) \in \underset{\boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^M}{\operatorname{argmax}} \mathcal{L}(\mathbf{x}_1^{k+1}, \dots, \mathbf{x}_N^{k+1}, \boldsymbol{\lambda}, \boldsymbol{\mu}) - \sum_{i=1}^N (\mathbf{x}_i^{k+1})^T \mathbf{u}_i^k + \boldsymbol{\lambda}^T \mathbf{v}^k + \boldsymbol{\mu}^T \mathbf{w}^k.$$
(60)

The first-order optimality conditions (58) and (60) together imply that

$$(\mathbf{x}_1^{k+1},\ldots,\mathbf{x}_N^{k+1},\boldsymbol{\lambda}^{k+1},\boldsymbol{\mu}^{k+1}) \in S^{-1}(\mathbf{u}_1^k,\ldots,\mathbf{u}_N^k,\mathbf{v}^k,\mathbf{w}^k)$$

By (49), we have $\lim_{k\to\infty}(\mathbf{u}_1^k,\ldots,\mathbf{u}_N^k,\mathbf{v}^k,\mathbf{w}^k)\to \mathbf{0}$. Choose in integer \bar{k} such that, for all $k\geq \bar{k}$, $\|(\mathbf{u}_1^k,\ldots,\mathbf{u}_N^k,\mathbf{v}^k,\mathbf{w}^k)\|_2\leq \tau$, then by Assumption 2.4, we have:

$$\begin{split} &\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2} \\ &\leq a^{2} \Big(\sum_{i=1}^{N} \|\mathbf{u}_{i}^{k}\|_{2}^{2} + \|\mathbf{v}^{k}\|_{2}^{2} + \|\mathbf{w}^{k}\|_{2}^{2} \Big) \\ &\leq a^{2} \Big(NA_{max}^{2} \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + NL_{max}^{2} \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} + \frac{1}{\rho^{2}} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} \\ &\quad + \frac{1}{\rho^{2}} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \frac{1}{\rho^{2}} \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2} \Big) \\ &\leq a^{2} \Bigg[\frac{1}{\rho^{2}} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} + NA_{max}^{2} \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + NL_{max}^{2} \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} \\ &\quad + \frac{1}{\rho^{2}} \Big(\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\gamma}^{k+1}\|_{2}^{2} + \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} \Big) \\ &\quad + \frac{1}{\rho^{2}} \Big(\|\boldsymbol{\mu}^{k+1} - \boldsymbol{\nu}^{k+1}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2} \Big) \Bigg] \\ \leq a^{2} \Bigg[\frac{1}{\rho^{2}} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} \\ &\quad + (NA_{max}^{2} + \frac{1}{\rho^{2}}) \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + (NL_{max}^{2} + \frac{1}{\rho^{2}}) \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} \\ &\quad + \frac{1}{\rho^{2}} \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \frac{1}{\rho^{2}} \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2} \Bigg] \\ \leq a^{2} (N\alpha^{2} + \frac{1}{\rho^{2}}) \Big(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{k}\|_{2}^{2} + \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k+1}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k+1}\|_{2$$

$$+ \|\boldsymbol{\gamma}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \|\boldsymbol{\nu}^{k+1} - \boldsymbol{\mu}^{k}\|_{2}^{2}$$

$$\leq \frac{a^{2}(N\alpha^{2} + \frac{1}{\rho^{2}})}{\epsilon} \left[\left(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2} \right)$$

$$- \left(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2} \right) \right].$$

$$(61)$$

The last inequality is due to (46), and $\alpha := \max\{A_{max}, L_{max}\}$. We further derive

$$\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^{*}\|_{2}^{2}
\leq \theta^{2} \Big(\sum_{i=1}^{N} \|\mathbf{x}_{i}^{k} - \mathbf{x}_{i}^{*}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k} - \boldsymbol{\lambda}^{*}\|_{2}^{2} + \|\boldsymbol{\mu}^{k} - \boldsymbol{\mu}^{*}\|_{2}^{2} \Big),$$
(62)

where $\theta = \sqrt{\frac{1}{1+\beta}} < 1$ and $\beta = \frac{\epsilon}{a^2(N\alpha^2 + \frac{1}{\rho^2})} > 0$.

B Proofs in Section 3.2

B.1 Proof of Theorem 3.2

For all $k \ge 0$, we can equivalently write the update steps in Algorithm 2 and Algorithm 3 as

$$\mathbf{x}_{i}^{k+1} = \begin{cases} \underset{\mathbf{x}_{i} \in \mathcal{X}_{i}}{\operatorname{argmin}} f_{i}(\mathbf{x}_{i}) + (2\boldsymbol{\lambda}^{\hat{k}_{i}+1} - \boldsymbol{\lambda}^{\hat{k}_{i}})^{T} A_{i} \mathbf{x}_{i} + \frac{1}{2\rho} \|\mathbf{x}_{i} - \mathbf{x}_{i}^{\hat{k}_{i}+1}\|^{2}, & \forall i \in \mathcal{A}_{k} \\ \mathbf{x}_{i}^{k}, & \forall i \in \mathcal{A}_{k}^{\complement}, \end{cases}$$
(63)

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho \left(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{b} \right), \tag{64}$$

$$\hat{\gamma} = \lambda^{k+1} + \rho \left(\sum_{i=1}^{N} A_i \mathbf{x}_i^{k+1} - \mathbf{b} \right) = 2\lambda^{k+1} - \lambda^k, \tag{65}$$

where \hat{k}_i is the last iteration when the main processor receives $\hat{\mathbf{x}}_i$ from the worker processor $i \in \mathcal{A}_k$ before iteration k. For each worker processor $i \in \mathcal{A}_k^0$, we denote $\bar{k}_i \in (k-\tau,k)$ as the last iteration when the main processor receives $\hat{\mathbf{x}}_i$ from the worker processor i before iteration k, and further denote $\bar{k}_i \in [\bar{k}_i - \tau, \bar{k}_i)$ as the last iteration when the main processor receives $\hat{\mathbf{x}}_i$ from the worker processor i before iteration \bar{k}_i . We can rewrite the primal minimization step as

$$\mathbf{x}_{i}^{k+1} = \begin{cases} \arg\min_{\mathbf{x}_{i}} f_{i}(\mathbf{x}_{i}) + (2\boldsymbol{\lambda}^{\hat{k}_{i}+1} - \boldsymbol{\lambda}^{\hat{k}_{i}})^{T} A_{i} \mathbf{x}_{i} + \frac{1}{2\rho} \|\mathbf{x}_{i} - \mathbf{x}_{i}^{\hat{k}_{i}+1}\|^{2}, & \forall i \in \mathcal{A}_{k} \\ \mathbf{x}_{i}^{\bar{k}_{i}+1} = \arg\min_{\mathbf{x}_{i}} f_{i}(\mathbf{x}_{i}) + (2\boldsymbol{\lambda}^{\bar{k}_{i}+1} - \boldsymbol{\lambda}^{\bar{k}_{i}})^{T} A_{i} \mathbf{x}_{i} + \frac{1}{2\rho} \|\mathbf{x}_{i} - \mathbf{x}_{i}^{\bar{k}_{i}+1}\|^{2}, & \forall i \in \mathcal{A}_{k}^{\complement} \end{cases}$$
(66)

At each iteration $k \geq 0$, for any $i \in \mathcal{A}_k$, applying Lemma 3.1 with $\hat{\mathbf{z}} = \mathbf{x}_i^{k+1}$, $\bar{\mathbf{z}} = \mathbf{x}_i^{\hat{k}_i+1}$, and $\mathbf{z} = \mathbf{x}_i^*$, we have:

$$f_{i}(\mathbf{x}_{i}^{k+1}) - f_{i}(\mathbf{x}_{i}^{*}) + \boldsymbol{\lambda}^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) + \left(\frac{\sigma_{i}}{2} + \frac{1}{2\rho} \right) \| \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*} \|_{2}^{2}$$

$$+ \frac{1}{2\rho} \| \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{\hat{k}_{i}+1} \|_{2}^{2} - \frac{1}{2\rho} \| \mathbf{x}_{i}^{\hat{k}_{i}+1} - \mathbf{x}_{i}^{*} \|_{2}^{2}$$

$$+ (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda})^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) + (2\boldsymbol{\lambda}^{\hat{k}_{i}+1} - \boldsymbol{\lambda}^{\hat{k}_{i}} - \boldsymbol{\lambda}^{k+1})^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) \leq 0.$$

$$(67)$$

At each iteration $k \geq 0$, for any $i \in \mathcal{A}_k^{\complement}$, applying Lemma 3.1 with $\hat{\mathbf{z}} = \mathbf{x}_i^{k+1}$, $\bar{\mathbf{z}} = \mathbf{x}_i^{\bar{k}_i+1}$, and $\mathbf{z} = \mathbf{x}_i^*$, we

$$f_{i}(\mathbf{x}_{i}^{k+1}) - f_{i}(\mathbf{x}_{i}^{*}) + \boldsymbol{\lambda}^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) + \left(\frac{\sigma_{i}}{2} + \frac{1}{2\rho} \right) \| \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*} \|_{2}^{2}$$

$$+ \frac{1}{2\rho} \| \mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{\bar{k}_{i}+1} \|_{2}^{2} - \frac{1}{2\rho} \| \mathbf{x}_{i}^{\bar{k}_{i}+1} - \mathbf{x}_{i}^{*} \|_{2}^{2}$$

$$+ (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda})^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) + (2\boldsymbol{\lambda}^{\bar{k}_{i}+1} - \boldsymbol{\lambda}^{\bar{k}_{i}} - \boldsymbol{\lambda}^{k+1})^{T} \left(A_{i} \mathbf{x}_{i}^{k+1} - A_{i} \mathbf{x}_{i}^{*} \right) \leq 0.$$
 (68)

Summing (67) over all $i \in \mathcal{A}_k$ and (68) over all $i \in \mathcal{A}_k^\complement$ yields

$$\begin{split} &\sum_{i=1}^{N} f_i(\mathbf{x}_i^{k+1}) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*) + \underbrace{\boldsymbol{\lambda}^T \sum_{i=1}^{N} \left(A_i \mathbf{x}_i^{k+1} - A_i \mathbf{x}_i^*\right)}_{(\mathbf{a})} + \frac{\sigma_{min}}{2} \sum_{i=1}^{N} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|_2^2 \\ &+ \underbrace{\frac{1}{2\rho} \sum_{i=1}^{N} \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|_2^2}_{(\mathbf{b})} \\ &+ \underbrace{\frac{1}{2\rho} \left[\sum_{i \in \mathcal{A}_k} \left(\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^{\hat{k}_i + 1}\|_2^2 - \|\mathbf{x}_i^{\hat{k}_i + 1} - \mathbf{x}_i^*\|_2^2 \right) + \sum_{i \in \mathcal{A}_k^0} \left(\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^{\bar{k}_i + 1}\|_2^2 - \|\mathbf{x}_i^{\bar{k}_i + 1} - \mathbf{x}_i^*\|_2^2 \right) \right]}_{(\mathbf{c})} \\ &+ \underbrace{\left(\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda} \right)^T \sum_{i=1}^{N} \left(A_i \mathbf{x}_i^{k+1} - A_i \mathbf{x}_i^* \right)}_{(\mathbf{d})} \\ &+ \sum_{i \in \mathcal{A}_k} \left(2 \boldsymbol{\lambda}^{\hat{k}_i + 1} - \boldsymbol{\lambda}^{\hat{k}_i} - \boldsymbol{\lambda}^{k+1} \right)^T \left(A_i \mathbf{x}_i^{k+1} - A_i \mathbf{x}_i^* \right) \\ &+ \sum_{i \in \mathcal{A}_k} \left(2 \boldsymbol{\lambda}^{\bar{k}_i + 1} - \boldsymbol{\lambda}^{\bar{k}_i} - \boldsymbol{\lambda}^{k+1} \right)^T \left(A_i \mathbf{x}_i^{k+1} - A_i \mathbf{x}_i^* \right) \end{split}$$

$$+\sum_{i\in\mathcal{A}_{k}^{\mathbf{G}}} (2\boldsymbol{\lambda}^{\bar{k}_{i}+1} - \boldsymbol{\lambda}^{\bar{k}_{i}} - \boldsymbol{\lambda}^{k+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*})$$

$$\leq 0.$$
(69)

The term (a) can be rewritten as:

$$(a) = \boldsymbol{\lambda}^T \left(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \sum_{i=1}^N A_i \mathbf{x}_i^* \right) = \boldsymbol{\lambda}^T \left(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \mathbf{b} \right).$$
 (70)

The term (b) + (c) can be rewritten as:

$$(b) + (c) = \frac{1}{2\rho} \sum_{i \in \mathcal{A}_k} \left(\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|_2^2 + \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^{\hat{k}_i + 1}\|_2^2 - \|\mathbf{x}_i^{\hat{k}_i + 1} - \mathbf{x}_i^*\|_2^2 \right)$$

$$+ \frac{1}{2\rho} \sum_{i \in \mathcal{A}_k^0} \left(\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|_2^2 + \|\mathbf{x}_i^{k+1} - \mathbf{x}_i^{\bar{k}_i + 1}\|_2^2 - \|\mathbf{x}_i^{\bar{k}_i + 1} - \mathbf{x}_i^*\|_2^2 \right)$$

$$\geq 0.$$

$$(71)$$

The term (d) can be rewritten as:

$$(d) = (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda})^T \left(\sum_{i=1}^N A_i \mathbf{x}_i^{k+1} - \sum_{i=1}^N A_i \mathbf{x}_i^* \right) = \frac{1}{\rho} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda})^T (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k).$$
(72)

We substitute (70), (71) and (72) into (69), and sum it over $k = 0 \dots K - 1$. Taking the average yields

$$\frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} f_i(\mathbf{x}_i^{k+1}) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*) + \frac{1}{K} \lambda^T \left(\sum_{k=0}^{K-1} \sum_{i=1}^{N} A_i \mathbf{x}_i^{k+1} - \mathbf{b} \right)$$

$$\leq -\frac{\sigma_{min}}{2K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2} \\
-\frac{1}{\rho K} \underbrace{\sum_{k=0}^{K-1} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda})^{T} (\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k})}_{\text{(e)}} \\
+\frac{1}{K} \underbrace{\sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} (\boldsymbol{\lambda}^{\hat{k}_{i}} + \boldsymbol{\lambda}^{k+1} - 2\boldsymbol{\lambda}^{\hat{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*})}_{\text{(f)}} \\
+\frac{1}{K} \underbrace{\sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} (\boldsymbol{\lambda}^{\bar{k}_{i}} + \boldsymbol{\lambda}^{k+1} - 2\boldsymbol{\lambda}^{\bar{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*})}_{\text{(g)}}.$$
(73)

The term (e) in (73) can be rewritten as:

$$(e) = \frac{1}{2} \sum_{k=0}^{K-1} (\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}\|_{2}^{2} - \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{k}\|_{2}^{2} + \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2})$$

$$= \frac{1}{2} \|\boldsymbol{\lambda}^{K} - \boldsymbol{\lambda}\|_{2}^{2} - \frac{1}{2} \|\boldsymbol{\lambda}^{0} - \boldsymbol{\lambda}\|_{2}^{2} + \frac{1}{2} \sum_{k=0}^{K-1} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2}.$$

$$(74)$$

The term (f) in (73) can be bounded as:

$$\sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} (\lambda^{\hat{k}_{i}} + \lambda^{k+1} - 2\lambda^{\hat{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\
= \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} (\lambda^{\hat{k}_{i}} - \lambda^{\hat{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) + \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} (\lambda^{k+1} - \lambda^{\hat{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\
= \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} \sum_{l=\hat{k}_{i}} (\lambda^{l} - \lambda^{l+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\
+ \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}} \sum_{l=\hat{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\
\leq \sum_{i \in \mathcal{A}_{k}} \sum_{k=0}^{K-1} \sum_{l=\hat{k}_{i}} (\frac{1}{2\delta^{2}} \|\lambda^{l} - \lambda^{l+1}\|_{2}^{2} + \frac{\delta^{2} \|A_{i}\|_{2}^{2}}{2} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}) \\
+ \sum_{i \in \mathcal{A}_{k}} \sum_{k=0}^{K-1} \sum_{l=\hat{k}_{i}+1} (\frac{1}{2\delta^{2}} \|\lambda^{l+1} - \lambda^{l}\|_{2}^{2} + \frac{\delta^{2} \|A_{i}\|_{2}^{2}}{2} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}) \\
\leq \sum_{i=1}^{N} \sum_{k=0}^{K-1} (\tau - 1) (\frac{1}{2\delta^{2}} \|\lambda^{k} - \lambda^{k+1}\|_{2}^{2} + \frac{\delta^{2} \|A_{i}\|_{2}^{2}}{2} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}) \\
+ \sum_{i=1}^{N} \sum_{k=0}^{K-1} (\tau - 1) (\frac{1}{2\delta^{2}} \|\lambda^{k+1} - \lambda^{k}\|_{2}^{2} + \frac{\delta^{2} \|A_{i}\|_{2}^{2}}{2} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}) \\
\leq \frac{(\tau - 1)N}{\delta^{2}} \sum_{k=0}^{K-1} \|\lambda^{k+1} - \lambda^{k}\|_{2}^{2} + (\tau - 1)\delta^{2}A_{\max}^{2} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}, \tag{75}$$

where the first inequality is obtained by (40), and the second inequality is due to the fact that the term $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|_2^2$ does not appear more than $\tau - 1$ times for each iteration k.

Similarly, the term (g) in (73) can be bounded as:

$$\begin{split} &\sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} (\lambda^{\bar{k}_{i}} + \lambda^{k+1} - 2\lambda^{\bar{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &= \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} (\lambda^{\bar{k}_{i}} - \lambda^{\bar{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) + \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} (\lambda^{\bar{k}_{i}+1} - \lambda^{\bar{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{i \in \mathcal{A}_{k}^{0}} (\lambda^{k+1} - \lambda^{\bar{k}_{i}+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &= \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{i \in \bar{\lambda}_{k}^{0}} (\lambda^{l} - \lambda^{l+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{k=0}^{K-1} \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l+1} - \lambda^{l})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l} - \lambda^{l+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l} - \lambda^{l+1})^{T} (A_{i}\mathbf{x}_{i}^{k+1} - A_{i}\mathbf{x}_{i}^{*}) \\ &+ \sum_{i \in \mathcal{A}_{k}^{0}} \sum_{k=0}^{K-1} \sum_{i = \bar{k}_{i}+1} (\lambda^{l} - \lambda^{l$$

By substituting (74), (75) and (76) into (73) and denoting

$$\bar{\mathbf{x}}_i^K = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{x}_i^{k+1},$$

for all $i = 1 \dots N$, we have:

$$\sum_{i=1}^{N} f_{i}(\bar{\mathbf{x}}_{i}^{K}) - \sum_{i=1}^{N} f_{i}(\mathbf{x}_{i}^{*}) + \boldsymbol{\lambda}^{T} \left(\sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}_{i}^{K} - \mathbf{b} \right)$$

$$\leq \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} f_{i}(\mathbf{x}_{i}^{k+1}) - \sum_{i=1}^{N} f_{i}(\mathbf{x}_{i}^{*}) + \frac{1}{K} \boldsymbol{\lambda}^{T} \sum_{k=0}^{K-1} \left(\sum_{i=1}^{N} A_{i} \mathbf{x}_{i}^{k+1} - \mathbf{b} \right)$$

$$\leq \frac{1}{2\rho K} \|\boldsymbol{\lambda}^{0} - \boldsymbol{\lambda}\|_{2}^{2} - \frac{1}{2\rho K} \|\boldsymbol{\lambda}^{K} - \boldsymbol{\lambda}\|_{2}^{2} \\
+ \left(-\frac{1}{2\rho K} + \frac{5(\tau - 1)N}{2\delta^{2}K}\right) \sum_{k=0}^{K-1} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^{k}\|_{2}^{2} \\
+ \left(-\frac{\sigma_{min}}{2K} + \frac{5(\tau - 1)\delta^{2}A_{\max}^{2}}{2K}\right) \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\mathbf{x}_{i}^{k+1} - \mathbf{x}_{i}^{*}\|_{2}^{2}, \tag{77}$$

where the first inequality is due to the convexity of f_i for all $i=1\dots N$. By choosing $\delta^2 \leq \frac{\sigma_{min}}{5(\tau-1)A_{\max}^2}$ and $\rho \leq \frac{\delta^2}{5(\tau-1)N}$, which implies

$$\rho \le \frac{\sigma_{min}}{25N(\tau - 1)^2 A_{\max}^2},$$

we derive:

$$\sum_{i=1}^{N} f_i(\bar{\mathbf{x}}_i^K) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*) + \boldsymbol{\lambda}^T \left(\sum_{i=1}^{N} A_i \bar{\mathbf{x}}_i^K - \mathbf{b} \right) \le \frac{1}{2\rho K} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}\|_2^2.$$
 (78)

Let $\lambda = \lambda^* + \frac{\sum_{i=1}^N A_i \bar{\mathbf{x}}^K - \mathbf{b}}{\|\sum_{i=1}^N A_i \bar{\mathbf{x}}^K - \mathbf{b}\|_2}$, and note that by the duality theory, we have:

$$\sum_{i=1}^{N} f_i(\bar{\mathbf{x}}_i^K) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*) + (\boldsymbol{\lambda}^*)^T \left(\sum_{i=1}^{N} A_i \bar{\mathbf{x}}_i^K - \mathbf{b}\right) \ge 0.$$
 (79)

Then, we further derive:

$$\left\| \sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}^{K} - \mathbf{b} \right\|_{2}$$

$$\leq \sum_{i=1}^{N} f_{i} (\bar{\mathbf{x}}_{i}^{K}) - \sum_{i=1}^{N} f_{i} (\mathbf{x}_{i}^{*}) + (\boldsymbol{\lambda}^{*})^{T} \left(\sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}_{i}^{K} - \mathbf{b} \right) + \left\| \sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}^{K} - \mathbf{b} \right\|_{2}$$

$$\leq \frac{1}{2\rho K} \left\| \boldsymbol{\lambda}^{0} - \left(\boldsymbol{\lambda}^{*} + \frac{\sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}^{K} - \mathbf{b}}{\left\| \sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}^{K} - \mathbf{b} \right\|_{2}} \right) \right\|_{2}^{2}, \tag{80}$$

which implies

$$\left\| \sum_{i=1}^N A_i \bar{\mathbf{x}}^K - \mathbf{b} \right\|_2 \le \frac{1}{K} \left[\frac{1}{2\rho} \max_{\|\boldsymbol{\gamma}\|_2 \le 1} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^* - \boldsymbol{\gamma}\|_2^2 \right] \stackrel{\Delta}{=} \frac{C_1}{K}.$$

On the other hand, let $\lambda = \lambda^*$, and note that

$$\sum_{i=1}^{N} f_{i}(\bar{\mathbf{x}}_{i}^{K}) - \sum_{i=1}^{N} f_{i}(\mathbf{x}_{i}^{*}) + (\boldsymbol{\lambda}^{*})^{T} \left(\sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}_{i}^{K} - \mathbf{b} \right)$$

$$\geq \left| \sum_{i=1}^{N} f_{i}(\bar{\mathbf{x}}_{i}^{K}) - \sum_{i=1}^{N} f_{i}(\mathbf{x}_{i}^{*}) \right| - \|\boldsymbol{\lambda}^{*}\|_{2} \cdot \|\sum_{i=1}^{N} A_{i} \bar{\mathbf{x}}_{i}^{K} - \mathbf{b}\|_{2}.$$
(81)

Then, we have:

$$\left|\sum_{i=1}^{N} f_i(\bar{\mathbf{x}}_i^K) - \sum_{i=1}^{N} f_i(\mathbf{x}_i^*)\right| \le \|\boldsymbol{\lambda}^*\|_2 \cdot \|\sum_{i=1}^{N} A_i \bar{\mathbf{x}}_i^K - \mathbf{b}\|_2 + \frac{1}{K} (\frac{1}{2\rho} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|_2^2) \stackrel{\Delta}{=} \frac{\delta_{\boldsymbol{\lambda}} C_1 + C_2}{K}, \quad (82)$$

where $\delta_{\lambda} = \|\lambda^*\|_2$ and $C_2 = \frac{1}{2a} \|\lambda^0 - \lambda^*\|_2^2$.