Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems

Mingyi Hong*, Zhi-Quan Luo† and Meisam Razaviyayn‡

October 7, 2014

Abstract

The alternating direction method of multipliers (ADMM) is widely used to solve large-scale linearly constrained optimization problems, convex or nonconvex, in many engineering fields. However there is a general lack of theoretical understanding of the algorithm when the objective function is nonconvex. In this paper we analyze the convergence of the ADMM for solving certain nonconvex *consensus* and *sharing* problems, and show that the classical ADMM converges to the set of stationary solutions, provided that the penalty parameter in the augmented Lagrangian is chosen to be sufficiently large. For the sharing problems, we show that the ADMM is convergent regardless of the number of variable blocks. Our analysis does not impose any assumptions on the iterates generated by the algorithm, and is broadly applicable to many ADMM variants involving proximal update rules and various flexible block selection rules.

KEY WORDS: ADMM, nonconvex optimization, linear constraints

AMS(MOS) Subject Classifications: 49, 90.

*Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011, USA. Email: `mingyi@iastate.edu`

†The Chinese University of Hong Kong, Shenzhen, China and Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: `luozq@cuhk.edu.cn` and `luozq@umn.edu`

‡Department of Electrical Engineering, Stanford University. Email: `meisamr@stanford.edu`

1 Introduction

Consider the following linearly constrained (possibly nonsmooth or/and nonconvex) problem with K blocks of variables $\{x_k\}_{k=1}^K$:

$$\begin{aligned} \min \quad & f(x) := \sum_{k=1}^{K} g_k(x_k) + \ell(x_1, \cdots, x_K) \\ \text{s.t.} \quad & \sum_{k=1}^{K} A_k x_k = q, \ x_k \in X_k, \ \forall\, k = 1, \cdots, K \end{aligned} \tag{1.1}$$

where $A_k \in \mathbb{R}^{M \times N_k}$ and $q \in \mathbb{R}^M$; $X_k \in \mathbb{R}^{N_k}$ is a closed convex set; $\ell(\cdot)$ is a smooth (possibly nonconvex) function; each $g_k(\cdot)$ can be either a smooth function, or a convex nonsmooth function. Let us define $A := [A_1, \cdots, A_k]$. The augmented Lagrangian for problem (1.1) is given by

$$L(x; y) = \sum_{k=1}^{K} g_k(x_k) + \ell(x_1, \cdots, x_K) + \langle y, q - Ax \rangle + \frac{\rho}{2} \|q - Ax\|^2, \tag{1.2}$$

where $\rho > 0$ is a constant representing the primal stepsize.

To solve problem (1.1), let us consider a popular algorithm called the alternating direction method of multipliers (ADMM), whose steps are given below:

Algorithm 0. ADMM for Problem (1.1)

At each iteration $t+1$, update the primal variables:

$$x_k^{t+1} = \arg\min_{x_k \in X_k} L(x_1^{t+1}, \cdots, x_{k-1}^{t+1}, x_k, x_{k+1}^{t}, \cdots, x_K^{t}; y^t), \ \forall\, k = 1, \cdots, K. \tag{1.3}$$

Update the dual variable:

$$y^{t+1} = y^t + \rho(q - Ax^{t+1}). \tag{1.4}$$

The ADMM algorithm was originally introduced in early 1970s [1,2], and has since been studied extensively [3–6]. Recently it has become widely popular in modern big data related problems arising in machine learning, computer vision, signal processing, networking and so on; see [7–14] and the references therein. In practice, the algorithm often exhibits faster convergence than traditional primal-dual type algorithms such as the dual ascent algorithm [15–17] or the method of multipliers [18]. It is also particularly suitable for parallel implementation [7].

There is a vast literature that applies the ADMM to various problems in the form of (1.1). Unfortunately, theoretical understanding of the algorithm is still fairly limited. For example, most of its convergence analysis is done for certain special form of problem (1.1) — the *two-block convex* *separable* problems, where $K = 2$, $\ell = 0$ and g_1, g_2 are both convex. In this case, ADMM is known to converge under very mild conditions; see [6] and [7]. Under the same conditions, several recent works [19, 20] have shown that the ADMM converges with the sublinear rate of $\mathcal{O}(\frac{1}{t})$ (or $\mathcal{O}(\frac{1}{t^2})$ for the accelerated version [21, 22] under more stringent conditions). Reference [23] has shown that the ADMM converges linearly when the objective function as well as the constraints satisfy certain additional assumptions. For the *multi-block* separable convex problems where $K \geq 3$, it is known that the original ADMM can diverge for certain pathological problems [24]. Therefore, most research effort in this direction has been focused on either analyzing problems with additional conditions, or showing convergence for variants of the ADMM; see for example [24–32]. It is worth mentioning that when the objective function is not separable across the variables (e.g., the coupling function $\ell(\cdot)$ appears in the objective), the convergence of the ADMM is still open, even in the case where $K = 2$ and $f(\cdot)$ is convex. Recent works of [27, 33] have shown that when problem (1.1) is convex but not necessarily separable, and when certain error bound condition is satisfied, then the ADMM iteration converges to the set of primal-dual optimal solutions, provided that the dual stepsize decreases in time.

Unlike the convex case, for which the behavior of ADMM has been investigated quite extensively, when the objective becomes nonconvex, the convergence issue of ADMM remains largely open. Nevertheless, it has been observed by many researchers that the ADMM works extremely well for various applications involving nonconvex objectives, such as the nonnegative matrix factorization [34, 35], phase retrieval [36], distributed matrix factorization [37], distributed clustering [38], sparse zero variance discriminant analysis [39], polynomial optimization [40], tensor decomposition [41], matrix separation [42], matrix completion [43], asset allocation [44] and so on. However, to the best of our knowledge, existing convergence analysis of ADMM for nonconvex problems is very limited — all known global convergence analysis needs to impose uncheckable conditions on the sequence generated by the algorithm. For example, references [40, 42–44] show global convergence of the ADMM to the set of stationary solutions for their respective nonconvex problems, by making the key assumptions that the limit points do exist, and that the successive differences of the iterates (both primal and dual) converge to zero. However such assumption is nonstandard and overly restrictive. It is not clear whether the same convergence result can be claimed without making assumptions on the iterates. Reference [45] analyzes a family of splitting algorithms (which includes the ADMM as a special case) for certain nonconvex quadratic optimization problem, and shows that they converge to the stationary solution when certain condition on the dual stepsize is met.

The aim of this paper is to provide some theoretical justification on the good performance of the ADMM for nonconvex problems. Specifically, we establish the convergence of ADMM for certain types of nonconvex problems including the consensus and sharing problems without making any assumptions on the iterates. Our analysis shows that, as long as the objective functions g_k's and ℓ satisfy certain regularity conditions, and the stepsize ρ is chosen large enough (with computable bounds), then the iterates generated by the ADMM is guaranteed to converge to the set of stationary solutions. It should be noted that our analysis covers many variants of the ADMM including per-block proximal update and flexible block selection. An interesting consequence of our analysis is that for a particular reformulation of the sharing problem, the *multi-block* ADMM algorithm converges, regardless of the convexity of the objective function. Finally, to facilitate possible applications to other nonconvex problems, we highlight the main proof steps in our analysis framework that can guarantee the global convergence of the ADMM iterates (1.3)–(1.4) to the set of stationary solutions.

2 The Nonconvex Consensus Problem

2.1 The Basic Problem

Consider the following nonconvex global consensus problem with regularization

$$\begin{aligned} \min \quad & f(x) := \sum_{k=1}^{K} g_k(x) + h(x) \\ \text{s.t.} \quad & x \in X \end{aligned} \tag{2.1}$$

where g_k's are a set of smooth, possibly nonconvex functions, while $h(x)$ is a convex nonsmooth regularization term. This problem is related to the convex global consensus problem discussed heavily in [7, Section 7], but with the important difference that g_k's can be nonconvex.

In many practical applications, g_k's need to be handled by a single agent, such as a thread or a processor. This motivates the following consensus formulation. Let us introduce a set of new variables $\{x_k\}_{k=1}^K$, and transform problem (2.1) equivalently to the following linearly constrained problem

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} g_k(x_k) + h(x) \\ \text{s.t.} \quad & x_k = x, \ \forall\, k = 1, \cdots, K, \quad x \in X. \end{aligned} \tag{2.2}$$

The augmented Lagrangian function is given by

$$L(\{x_k\}, x; y) = \sum_{k=1}^{K} g_k(x_k) + h(x) + \sum_{k=1}^{K} \langle y_k, x_k - x \rangle + \sum_{k=1}^{K} \frac{\rho_k}{2} \|x_k - x\|^2. \tag{2.3}$$

Note that this augmented Lagrangian is slightly different from the one expressed in (1.2), as we have used a set of different penalization parameters $\{\rho_k\}$, one for each equality constraint $x_k = x$. We note that there can be many other variants of the basic consensus problem, such as the *general* *form consensus optimization*, the *sharing* problem and so on. We will discuss some of those variants in the later sections.

2.2 The ADMM Algorithm for Nonconvex Consensus

The problem (2.2) can be solved distributedly by applying the classical ADMM. The details are given in the table below.

Algorithm 1. The Classical ADMM for the Consensus Problem (2.2)

At each iteration $t+1$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x \in X} L(\{x_k^t\}, x; y^t) = \operatorname{prox}_{\iota(X)+h}\left[\frac{\sum_{k=1}^K \rho_k x_k^t + \sum_{k=1}^K y_k^t}{\sum_{k=1}^K \rho_k}\right]. \tag{2.4}$$

Each node k computes x_k by solving:

$$x_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k - x^{t+1}\|^2. \tag{2.5}$$

Each node k updates the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k\left(x_k^{t+1} - x^{t+1}\right). \tag{2.6}$$

In the above description, we have used the proximity operator, defined below. Let $h : \operatorname{dom}(h) \mapsto \mathbb{R}$ be a (possibly nonsmooth) convex function. For every $x \in \operatorname{dom}(h)$, the *proximity operator* of h is defined as [46, Section 31]

$$\operatorname{prox}_h(x) = \operatorname*{argmin}_u \; h(u) + \frac{1}{2}\|x - u\|^2.$$

When $h = \iota(X)$, the indicator function of a convex set X, the resulting proximity operator reduces to the usual projection operator proj_X. In the x update step, if the nonsmooth penalization $h(\cdot)$ does not appear in the objective, then this step can be simplified as

$$x^{t+1} = \operatorname*{argmin}_{x \in X} L(\{x_k^t\}, x; y^t) = \operatorname{proj}_X\left[\frac{\sum_{k=1}^K \rho_k x_k^t + \sum_{k=1}^K y_k^t}{\sum_{k=1}^K \rho_k}\right]. \tag{2.7}$$

Note that the above algorithm has the exact form as the classical ADMM described in [7], where the variable x is taken as the first block of primal variable, and the collection $\{x_k\}_{k=1}^K$ as the second block. The two primal blocks are updated in a sequential (i.e., Gauss-Seidel) manner, followed by an inexact dual ascent step.

In what follows, we consider a more general version of ADMM which includes Algorithm 1 as a special case. In particular, we propose a *flexible* ADMM algorithm in which there is a greater flexibility in choosing the order of the update of both the primal and the dual variables. Specifically, we consider the following two types of variable block update order rules: let $k = 1, 2, ..., K$ be the indices for the primal variable blocks $x_1, x_2, ..., x_K$ and $k = 0$ be the index for primal variable block x, and let $\mathcal{C}^t \subseteq \{0, 1, \cdots, K\}$ denote the set of variables updated in iteration t, then

1. *Randomized update rule*: At each iteration $t+1$, a variable block k is chosen randomly with probability p_k^{t+1},

$$\Pr\left(k \in \mathcal{C}^{t+1} \mid x^t, y^t, \{x_k^t\}\right) = p_k^{t+1} \geq p_{\min} > 0. \tag{2.8}$$

2. *Essentially cyclic update rule*: There exists a given period $T \geq 1$ during which each index is updated at least once. More specifically, at iteration t, update all the variables in an index set $\mathcal{C}^t$ whereby

$$\bigcup_{i=1}^{T} \mathcal{C}^{t+i} = \{0, 1, \cdots, K\}, \ \forall\, t. \tag{2.9}$$

We call this update rule a *period-T* essentially cyclic update rule.

Algorithm 2. The Flexible ADMM for the Consensus Problem (2.2)

At each iteration $t+1$, pick an index set $\mathcal{C}^{t+1} \subseteq \{0, \cdots, K\}$.

If $0 \in \mathcal{C}^{t+1}$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x \in X} L(\{x_k^t\}, x; y^t) = \operatorname{prox}_{\iota(X)+h}\left[\frac{\sum_{k=1}^K \rho_k x_k^t + \sum_{k=1}^K y_k^t}{\sum_{k=1}^K \rho_k}\right]. \tag{2.10}$$

Else $x^{t+1} = x^t$.

If $k \neq 0$ and $k \in \mathcal{C}^{t+1}$, node k computes x_k by solving:

$$x_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k - x^{t+1}\|^2. \tag{2.11}$$

Update the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k\left(x_k^{t+1} - x^{t+1}\right). \tag{2.12}$$

Else $x_k^{t+1} = x_k^t$, $y_k^{t+1} = y_k^t$.

We note that the randomized version of Algorithm 2 is similar to that of the convex consensus algorithms studied in [47] and [48]. It is also related to the *randomized BSUM-M* algorithm studied in [27]. The difference with the latter is that in the randomized BSUM-M, the dual variable is viewed as an additional block that can be randomly picked (independent of the way that the primal blocks are picked), whereas in Algorithm 2, the dual variable y_k is always updated whenever the corresponding primal variable x_k is updated. To the best of our knowledge, the period-T essentially cyclic update rule is a new variant of the ADMM.

Notice that Algorithm 1 is simply the period-1 essentially cyclic rule, which is a special case of Algorithm 2. Therefore we will focus on analyzing Algorithm 2. To this end, we make the following assumption.

Assumption A.

A1. There exists a positive constant $L_k > 0$ such that

$$\|\nabla_k g_k(x_k) - \nabla_k g_k(z_k)\| \le L_k \|x_k - z_k\|, \ \forall\ x_k, z_k,\ k = 1, \cdots, K.$$

Moreover, h is convex (possible nonsmooth); X is a closed convex set.

A2. For all k, the stepsize ρ_k is chosen large enough such that:

1. For all k, the x_k subproblem (2.11) is strongly convex with modulus $\gamma_k(\rho_k)$;
2. For all k, $\rho_k \gamma_k(\rho_k) > 2L_k^2$ and $\rho_k \ge L_k$.

A3. $f(x)$ is bounded from below over X.

We have the following remarks regarding to the assumptions made above.

- As ρ_k inceases, the subproblem (2.11) will be eventually strongly convex with respect to x_k. The corresponding strong convexity modulus $\gamma_k(\rho_k)$ is a monotonic increasing function of ρ_k.
- Whenever $g_k(\cdot)$ is nonconvex (therefore $\rho_k > \gamma_k(\rho_k)$), the condition $\rho_k \gamma_k(\rho_k) \ge 2L_k^2$ implies $\rho_k \ge L_k$.
- By construction, $L(\{x_k\}, x; y)$ is also strongly convex with respect to x, with a modulus $\gamma := \sum_{k=1}^{K} \rho_k$.
- Assumption A makes no assumption on the *iterates* generated by the algorithm. This is in contrast to the existing analysis of the nonconvex ADMM algorithms [34, 40, 43].

Now we begin to analyze Algorithm 2. We first make several definitions. Let $t(k)$ (resp. $t(0)$) denote the latest iteration index that x_k (resp. x) is updated before iteration $t+1$, i.e.,

$$\begin{aligned} t(k) &= \arg\max_{r \le t} \{x_k^{r-1} \mid x_k^{r-1} \ne x_k^t\},\ k = 1, \cdots, K, \\ t(0) &= \arg\max_{r \le t} \{x^{r-1} \mid x^{r-1} \ne x^t\}. \end{aligned} \tag{2.13}$$

This definition implies that $x_k^t = x_k^{t(k)}$ for all k, and that $x^t = x^{t(0)}$.

Also define new vectors $\hat{x}^{t+1}$, $\{\hat{x}_k^{t+1}\}$ and $\hat{y}^{t+1}$ by

$$\hat{x}^{t+1} = \operatorname*{argmin}_{x \in X} L(\{x_k^t\}, x; y^t), \tag{2.14}$$

$$\hat{x}_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - \hat{x}^{t+1} \rangle + \frac{\rho_k}{2} \|x_k - \hat{x}^{t+1}\|^2, \ \forall\ k \tag{2.15}$$

$$\hat{y}_k^{t+1} = y_k^t + \rho_k \left(\hat{x}_k^{t+1} - \hat{x}^{t+1} \right). \tag{2.16}$$

In words, $(\hat{x}^{t+1}, \{\hat{x}_k^{t+1}\}, \hat{y}^{t+1})$ is a "virtual" iterate assuming that all variables are updated at iteration $t+1$.

We first show that the size of the successive difference of the dual variables can be bounded above by that of the primal variables.

Lemma 2.1 *Suppose Assumption A holds. Then for Algorithm 2 with either randomized or essentially cyclic update rule, the following are true*

$$L_k^2 \|x_k^{t+1} - x_k^t\|^2 \geq \|y_k^{t+1} - y_k^t\|^2, \ \forall\ k = 1, \cdots, K, \tag{2.17}$$

$$L_k^2 \|x_k^{t+1} - x_k^{t(k)}\|^2 \geq \|y_k^{t+1} - y_k^{t(k)}\|^2, \ \forall\ k = 1, \cdots, K, \tag{2.18}$$

$$L_k^2 \|\hat{x}_k^{t+1} - x_k^t\|^2 \geq \|\hat{y}_k^{t+1} - y_k^t\|^2, \ \forall\ k = 1, \cdots, K. \tag{2.19}$$

Proof. We will show the first inequality. The rest of the inequalities follow a similar line of argument.

To prove (2.17), first note that the case for $k \notin \mathcal{C}^{t+1}$ is trivial, as both sides of (2.17) evaluate to zero. Suppose $k \in \mathcal{C}^{t+1}$. From the x_k update step (2.11) we have the following optimality condition

$$\nabla g_k(x_k^{t+1}) + y_k^t + \rho_k(x_k^{t+1} - x^{t+1}) = 0, \ \forall\ k \in \mathcal{C}^{t+1}. \tag{2.20}$$

Combined with the dual variable update step (2.12) we obtain

$$\nabla g_k(x_k^{t+1}) = -y_k^{t+1}, \ \forall\ k \in \mathcal{C}^{t+1}. \tag{2.21}$$

Combining this with Assumption A1, and noting that for any given k, y_k and x_k are always updated in the same iteration, we obtain for all $k \in \mathcal{C}^{t+1}$

$$\|y_k^{t+1} - y_k^t\| = \|y_k^{t+1} - y_k^{t(k)}\| \leq \|\nabla g_k(x_k^{t+1}) - \nabla g_k(x_k^{t(k)})\| \leq L_k \|x_k^{t+1} - x_k^{t(k)}\| = L_k \|x_k^{t+1} - x_k^t\|.$$

The desired result follows. **Q.E.D.**

Next, we use (2.17) to bound the difference of the augmented Lagrangian.

Lemma 2.2 *For Algorithm 2 with either randomized or period-T essentially cyclic update rule, we have the following*

$$\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \\
&\leq \sum_{k\neq 0, k\in \mathcal{C}^{t+1}} \left(\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2}\right) \|x_k^{t+1} - x_k^t\|^2 - \iota\{0 \in \mathcal{C}^{t+1}\}\frac{\gamma}{2}\|x^{t+1} - x^t\|^2.
\end{aligned} \tag{2.22}$$

where $\iota\{0 \in \mathcal{C}^{t+1}\}$ *is the indicator function that takes the value* 1 *if* $0 \in \mathcal{C}^{t+1}$ *is true, and takes value* 0 *otherwise.*

Proof. We first split the successive difference of the augmented Lagrangian by

$$\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \\
&= \big(L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^{t+1}\}, x^{t+1}; y^t)\big) + \big(L(\{x_k^{t+1}\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t)\big)
\end{aligned} \tag{2.23}$$

The first term in (2.23) can be bounded by

$$\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^{t+1}\}, x^{t+1}; y^t) \\
&= \sum_{k=1}^{K} \langle y_k^{t+1} - y_k^t, x_k^{t+1} - x^{t+1}\rangle \\
&\overset{\text{(a)}}{=} \sum_{k\neq 0, k\in \mathcal{C}^{t+1}} \rho_k \|x_k^{t+1} - x^{t+1}\|^2 \\
&= \sum_{k\neq 0, k\in \mathcal{C}^{t+1}} \frac{1}{\rho_k}\|y_k^{t+1} - y_k^t\|^2
\end{aligned} \tag{2.24}$$

where in (a) we have use the fact that $y_k^{t+1} - y_k^t = 0$ for all variable block x_k that has not been updated (i.e., $k \neq 0, k \notin \mathcal{C}^{t+1}$). The second term in (2.23) can be bounded by

$$
\begin{aligned}
& L(\{x_k^{t+1}\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& = L(\{x_k^{t+1}\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^{t+1}; y^t) + L(\{x_k^t\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& \overset{\text{(a)}}{\leq} \sum_{k=1}^{K} \left(\nabla_{x_k} \langle L(\{x_k^{t+1}\}, x^{t+1}; y^t), x_k^{t+1} - x_k^t \rangle - \frac{\gamma_k(\rho_k)}{2} \|x_k^{t+1} - x_k^t\|^2 \right) \\
& \qquad + \partial_x \langle L(\{x_k^t\}, x^{t+1}; y^t), x^{t+1} - x^t \rangle - \frac{\gamma}{2} \|x^{t+1} - x^t\|^2 \\
& \overset{\text{(b)}}{=} \sum_{k \neq 0, k \in \mathcal{C}^{t+1}} \left(\nabla_{x_k} \langle L(\{x_k^{t+1}\}, x^{t+1}; y^t), x_k^{t+1} - x_k^t \rangle - \frac{\gamma_k(\rho_k)}{2} \|x_k^{t+1} - x_k^t\|^2 \right) \\
& \qquad + \iota\{0 \in \mathcal{C}^{t+1}\} \left(\partial_x \langle L(\{x_k^t\}, x^{t+1}; y^t), x^{t+1} - x^t \rangle - \frac{\gamma}{2} \|x^{t+1} - x^t\|^2 \right) \\
& \overset{\text{(c)}}{\leq} - \sum_{k \neq 0, k \in \mathcal{C}^{t+1}} \frac{\gamma_k(\rho_k)}{2} \|x_k^{t+1} - x_k^t\|^2 - \iota\{0 \in \mathcal{C}^{t+1}\} \frac{\gamma}{2} \|x^{t+1} - x^t\|^2,
\end{aligned} \tag{2.25}
$$

where in (a) we have used the fact that $L(\{x_k\}, x; y)$ is strongly convex w.r.t. each x_k and x, with modulus $\gamma_k(\rho_k)$ and γ, respectively; in (b) we have used the fact that when $k \notin \mathcal{C}^{t+1}$ (resp. $0 \notin \mathcal{C}^{t+1}$), $x_k^{t+1} = x_k^t$ (resp. $x^{t+1} = x^t$); in (c) we have used the optimality of each subproblem (2.11) and (2.10). Combining the above two inequalities (2.24) and (2.25), we obtain

$$
\begin{aligned}
& L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \\
& \leq - \sum_{k \neq 0, k \in \mathcal{C}^{t+1}} \frac{\gamma_k(\rho_k)}{2} \|x_k^{t+1} - x_k^t\|^2 + \sum_{k \neq 0, k \in \mathcal{C}^{t+1}} \frac{1}{\rho_k} \|y_k^{t+1} - y_k^t\|^2 - \iota\{0 \in \mathcal{C}^{t+1}\} \frac{\gamma}{2} \|x^{t+1} - x^t\|^2 \\
& \leq \sum_{k \neq 0, k \in \mathcal{C}^{t+1}} \left(\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2} \right) \|x_k^{t+1} - x_k^t\|^2 - \iota\{0 \in \mathcal{C}^{t+1}\} \frac{\gamma}{2} \|x^{t+1} - x^t\|^2
\end{aligned}
$$

where the last inequality is due to (2.17). **Q.E.D.**

The above result implies that if the following condition is satisfied:

$$
\rho_k \gamma_k(\rho_k) \geq 2L_k^2, \ \forall\ k = 1, \cdots, K, \tag{2.26}
$$

then the value of the augmented Lagrangian function will always decrease. Note that as long as $\gamma_k(\rho_k) \neq 0$, one can always find a ρ_k large enough such that the above condition is satisfied, as the left hand side (lhs) of (2.26) is monotonically increasing w.r.t. ρ_k, while the right hand side (rhs) is a constant.

Next we show that $L(\{x_k^t\}, x^t; y^t)$ is in fact convergent.

Lemma 2.3 *Suppose Assumption A is true. Let $\{\{x_k^t\}, x^t, y^t\}$ be generated by Algorithm 2 with either the essentially cyclic rule or the randomized rule. Then we have*

$$\lim_{t\to\infty} L(\{x_k^t\}, x^t, y^t) = L^* > -\infty \tag{2.27}$$

for some constant L^*.

Proof. Notice that the augmented Lagrangian function can be expressed as

$$\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) \\
&= h(x^{t+1}) + \sum_{k=1}^{K} \Big(g_k(x_k^{t+1}) + \langle y_k^{t+1}, x_k^{t+1} - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k^{t+1} - x^{t+1}\|^2 \Big) \\
&\overset{\text{(a)}}{=} h(x^{t+1}) + \sum_{k=1}^{K} \Big(g_k(x_k^{t+1}) + \langle \nabla g_k(x_k^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{\rho_k}{2}\|x_k^{t+1} - x^{t+1}\|^2 \Big) \\
&\overset{\text{(b)}}{\geq} h(x^{t+1}) + \sum_{k=1}^{K} g_k(x^{t+1}) = f(x^{t+1})
\end{aligned} \tag{2.28}$$

where (b) comes from the Lipschitz continuity of the gradient of g_k's (Assumption A1), and the fact that $\rho_k \geq L_k$ for all $k = 1, \cdots, K$ (Assumption A2). To see why (a) is true, we first observe that due to (2.21), we have for all $k \neq 0$ and $k \in \mathcal{C}^{t+1}$

$$\langle y_k^{t+1}, x_k^{t+1} - x^{t+1}\rangle = \langle \nabla g_k(x_k^{t+1}), x^{t+1} - x_k^{t+1}\rangle.$$

For all $k \neq 0$ and $k \notin \mathcal{C}^{t+1}$, it follows from $x_k^{t+1} = x_k^t = x_k^{t(k)} = x_k^{t(k)+1}$ and $y_k^{t+1} = y_k^t = y_k^{t(k)} = y_k^{t(k)+1}$ that

$$\langle y_k^{t+1}, x_k^{t+1} - x^{t+1}\rangle = \langle y_k^{t(k)+1}, x_k^{t(k)+1} - x^{t+1}\rangle = \langle \nabla g_k(x_k^{t(k)+1}), x^{t+1} - x_k^{t(k)+1}\rangle = \langle \nabla g_k(x_k^{t+1}), x^{t+1} - x_k^{t+1}\rangle.$$

Combining these two cases shows that (a) is true.

Clearly, (2.28) and Assumption A3 together imply that $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})$ is lower bounded. This combined with (2.22) says that whenever the stepsize ρ_k's are chosen sufficiently large (as per Assumption A2), $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})$ is monotonically decreasing and is convergent. This completes the proof. **Q.E.D.**

We are now ready to prove our first main result, which asserts that the sequence of iterates generated by Algorithm 2 converges to the set of stationary solution of problem (2.2).

Theorem 2.1 *Assume that Assumption A is satisfied. Then we have the following*

1. *We have* $\lim_{t\to\infty}\|x_k^{t+1}-x^{t+1}\|=0,\ k=1,\cdots,K$, *deterministically for the essentially cyclic update rule and almost surely for the randomized update rule.*

2. *Let* $(\{x_k^*\},x^*,y^*)$ *denote any limit point of the sequence* $\{\{x_k^{t+1}\},x^{t+1},y^{t+1}\}$ *generated by Algorithm 2. Then the following statement is true (deterministically for the essentially cyclic update rule and almost surely for the randomized update rule)*

$$
\begin{aligned}
&0=\nabla g_k(x_k^*)+y_k^*,\quad k=1,\cdots,K.\\
&x^*\in\arg\min_{x\in X}\ h(x)+\sum_{k=1}^K\langle y_k^*,x_k^*-x\rangle\\
&x_k^*=x^*,\quad k=1,\cdots,K.
\end{aligned}
$$

   That is, any limit point of Algorithm 2 is a stationary solution of problem (2.2)*.*

3. *If* X *is a compact set, then the sequence of iterates generated by Algorithm 2 converges to the set of stationary solutions of problem* (2.2)*.*

Proof. We first show part (1) of the theorem. For the essentially cyclic update rule, Lemma 2.2 implies that

$$
\begin{aligned}
&L(\{x_k^{t+T}\},x^{t+T};y^{t+T})-L(\{x_k^t\},x^t;y^t)\\
&\le\sum_{i=1}^T\sum_{k\neq 0,k\in\,\mathcal{C}^{t+i}}\left(\frac{L_k^2}{\rho_k}-\frac{\gamma_k(\rho_k)}{2}\right)\|x_k^{t+i}-x_k^{t+i-1}\|^2-\iota\{0\in\ \mathcal{C}^{t+i}\}\frac{\gamma}{2}\|x^{t+i-1}-x^{t+i}\|^2\\
&=\sum_{i=1}^T\sum_{k=1}^K\left(\frac{L_k^2}{\rho_k}-\frac{\gamma_k(\rho_k)}{2}\right)\|x_k^{t+i}-x_k^{t+i-1}\|^2-\frac{\gamma}{2}\|x^{t+i-1}-x^{t+i}\|^2,
\end{aligned}
$$

where the last equality follows from the fact $x^{t+i-1}=x^{t+i}$ if $0\notin\ \mathcal{C}^{t+i}$ and $x_k^{t+i}=x_k^{t+i-1}$ if $k\notin\ \mathcal{C}^{t+i}$ and $k\neq 0$. Using the fact that each index in $\{0,\cdots,K\}$ will be updated at least once during $[t,\ t+T]$, as well as Lemma 2.3 and the bounds for ρ_k's in Assumption A2, we have

$$
\|x^{t+1}-x^{t(0)}\|\to 0,\quad \|x_k^{t+1}-x_k^{t(k)}\|\to 0,\ \forall\ k=1,\cdots,K. \tag{2.29}
$$

By Lemma 2.1, we further obtain $\|y_k^{t+1}-y_k^{t(k)}\|\to 0$ for all $k=1,2,...,K$. In light of the dual update step of Algorithm 2, the fact that $\|y_k^{t+1}-y_k^{t(k)}\|\to 0$ implies that $\|x_k^{t+1}-x^{t+1}\|\to 0$.

For the randomized update rule, we can take the conditional expectation (over the choice of the blocks) on both sides of (2.22) and obtain

$$
\begin{aligned}
&\mathbb{E}\left[L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \mid \{x_k^t\}, x^t; y^t\right] \\
&\le \mathbb{E}\left[\sum_{k\neq 0, k\in\ \mathcal{C}^{t+1}} \left(\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2}\right)\|x_k^{t+1} - x_k^t\|^2 - \iota\{0\in\ \mathcal{C}^{t+1}\}\frac{\gamma}{2}\|x^{t+1} - x^t\|^2 \,\middle|\, \{x_k^t\}, x^t; y^t\right] \\
&\le \sum_{k=1}^{K} p_k p_0 \left(\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2}\right)\|\hat{x}_k^{t+1} - x_k^t\|^2 - p_0\frac{\gamma}{2}\|\hat{x}^{t+1} - x^t\|^2 \\
&\le p_{\min}^2 \sum_{k=1}^{K} \left(\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2}\right)\|\hat{x}_k^{t+1} - x_k^t\|^2 - p_{\min}\frac{\gamma}{2}\|\hat{x}^{t+1} - x^t\|^2
\end{aligned}
$$

where in the last two inequalities, we have used the fact that ρ_k's satisfy Assumption A2, hence $\frac{L_k^2}{\rho_k} - \frac{\gamma_k(\rho_k)}{2} < 0$ for all k; the last inequality is from the fact that $p_k \ge p_{\min}$ for all $k = 0, \cdots, K$. Applying the Supermartigale Convergence Theorem [49, Proposition 4.2], we have that $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})$ is convergent almost surely (a.s.), and that

$$
\|\hat{x}^{t+1} - x^t\| \to 0, \quad \|\hat{x}_k^{t+1} - x_k^t\| \to 0,\ \forall\ k = 1, \cdots, K, \quad \text{a.s.} \tag{2.30}
$$

By Lemma 2.1, we further obtain $\|\hat{y}_k^{t+1} - y_k^t\| \to 0,$ a.s. and for all $k = 1, 2, ..., K$. Finally, from the definition of $\hat{y}^{t+1}$, we see that $\|\hat{y}_k^{t+1} - y_k^t\| \to 0$ a.s. implies that $\|\hat{x}^{t+1} - \hat{x}_k^{t+1}\| \to 0$ a.s. for all $k = 1, 2, ..., K$.

Next we show part (2) of the theorem. For simplicity, we consider only the essentially cyclic rule as the proof for the randomized rule is similar. We begin by examining the optimality condition for the x_k and x subproblems at iteration $t+1$. Suppose $k \neq 0,\ k \in\ \mathcal{C}^{t+1}$, then we have

$$
\nabla g_k(x_k^{t+1}) + y_k^t + \rho_k(x_k^{t+1} - x^{t+1}) = 0. \tag{2.31}
$$

Similarly, suppose $0 \in\ \mathcal{C}^{t+1}$, we have

$$
\left\langle x - x^{t+1}, \partial h(x^{t+1}) - \sum_{k=1}^{K}\left(y_k^t - \rho_k(x^{t+1} - x^t)\right)\right\rangle \ge 0,\ \forall\ x \in X.
$$

These inequalities imply that

$$
\begin{aligned}
&\nabla g_k(x_k^{t+1}) + y_k^t + \rho_k(x_k^{t+1} - x^{t+1}) = 0,\ k \neq 0,\ k \in\ \mathcal{C}^{t+1} \\
&h(x) - h(x^{t+1}) + \left\langle x - x^{t+1}, \sum_{k=1}^{K}\left(-y_k^t + \rho_k(x^{t+1} - x^t)\right)\right\rangle \ge 0,\ \forall\ x \in X,\ \text{if } 0 \in\ \mathcal{C}^{t+1}.
\end{aligned} \tag{2.32}
$$

Using the definition of essentially cyclic update rule, we have that for all t

$$\nabla g_k(x_k^{r(k)}) + y_k^{r(k)} = 0,\ \forall\ k,\ \text{for some } r(k) \in [t,\ t+T],$$
$$h(x) - h(x^{r(0)}) + \left\langle x - x^{r(0)}, \sum_{k=1}^{K} \left(-y_k^{r(0)-1} + \rho_k(x^{r(0)} - x_k^{r(0)-1})\right) \right\rangle \geq 0,$$
$$\forall\ x \in X,\ \text{for some } r(0) \in [t,\ t+T].$$

Note that T is finite, and that $\|x_k^{t+1} - x_k^t\| \to 0$, $\|x^{t+1} - x^t\| \to 0$ and $\|y_k^{t+1} - y_k^t\| \to 0$, we have

$$\|x_k^{r(k)} - x_k^{t+1}\| \to 0,\ \forall\ k, \quad \|x^{r(0)} - x^{t+1}\| \to 0,$$
$$\|y_k^{t+1} - y_k^{r(k)}\| \to 0, \quad \|y_k^{t+1} - y_k^{r(0)-1}\| \to 0,\ \forall\ k. \tag{2.33}$$

Using this result, taking limit for (2.33), and using the fact that $\|x_k^{t+1} - x_k^t\| \to 0$, $x^{t+1} \to x^*$, $x_k^{t+1} \to x_k^*$, $y_k^{t+1} \to y_k^*$ for all k, we have

$$\nabla g_k(x_k^*) + y_k^* = 0,\ k = 1, \cdots, K$$
$$h(x) - h(x^*) + \langle x - x^*, -\sum_{k=1}^{K} y_k^* \rangle \geq 0,\ \forall x \in X. \tag{2.34}$$

Due to the fact that $\|y_k^{t+1} - y_k^t\| \to 0$ for all k, we have that primal feasibility is achieved in the limit, i.e.,

$$x_k^* = x^*,\ \forall\ k = 1, \cdots, K. \tag{2.35}$$

This set of equalities together with (2.34) imply

$$h(x) + \langle x_k^* - x, \sum_{k=1}^{K} y_k^* \rangle - \left(h(x^*) + \langle x_k^* - x^*, \sum_{k=1}^{K} y_k^* \rangle \right) \geq 0,\ \forall x \in X. \tag{2.36}$$

This concludes the proof of part (2).

To prove part (3), we only need to show that there exists a limit point for each of the sequences $\{x_k^t\}$, $\{x^t\}$ and $\{y^t\}$. Let us consider only the essentially cyclic rule. Due to the compactness assumption of X, it is obvious that $\{x^t\}$ must have a limit point. Also by a similar argument leading to (2.29), we see that $\|x_k^t - x^t\| \to 0$, thus for each k, x^t must also lie in a compact set thus have a limit point. Note that the Lipschitz continuity of ∇g_k combined with the compactness of the set X implies that the set $\{\nabla g_k(x) \mid x \in X\}$ is bounded, therefore $\{\nabla g_k(x^t)\}$ is a bounded sequence. Using (2.21), we conclude that that $\{y_k^t\}$ is also a bounded sequence, therefore must have at least one limit point. **Q.E.D.**

The analysis presented above is different from the conventional analysis of the ADMM algorithm where the main effort is to bound the distance between the current iterate and the optimal solution set. The above analysis is partly motivated by our previous analysis of the convergence of ADMM for multi-block convex problems, where the progress of the algorithm is measured by the combined decrease of certain primal and dual gaps; see [25, Theorem 3.1]. Nevertheless, the nonconvexity of the problem makes it difficult to estimate either the primal or the dual optimality gaps. Therefore we choose to use the decrease of the augmented Lagrangian as a measure of the progress of the algorithm.

2.3 The Proximal ADMM

One potential limitation of Algorithms 1 and 2 is the requirement that each subproblem (2.11) needs to be solved *exactly*, while in certain practical applications cheap iterations are preferred. In this section, we consider an important extension of Algorithm 1–2 in which the above restriction is removed. The main idea is to take a proximal step instead of minimizing the augmented Lagrangian function exactly with respect to each variable block. Like in the previous section, we will analyze a generalized version, termed *flexible* proximal ADMM, where there is more freedom in choosing the update schedules.

Algorithm 3. A Flexible Proximal ADMM for the Consensus Problem (2.2)

At each iteration $t+1$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x\in X} \; L(\{x_k^t\}, x; y^t) = \operatorname{prox}_{\iota(X)+h}\left[\frac{\sum_{k=1}^K \rho_k x_k^t + \sum_{k=1}^K y_k^t}{\sum_{k=1}^K \rho_k}\right]. \tag{2.37}$$

Pick a set $\mathcal{C}^{t+1} \subseteq \{1, \cdots, K\}$.

If $k \in \mathcal{C}^{t+1}$, update x_k by solving:

$$x_k^{t+1} = \operatorname*{argmin}_{x_k} \langle \nabla g_k(x^{t+1}), x_k - x^{t+1}\rangle + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k + L_k}{2}\|x_k - x^{t+1}\|^2. \tag{2.38}$$

Update the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k\left(x_k^{t+1} - x^{t+1}\right). \tag{2.39}$$

Else let $x_k^{t+1} = x_k^t$, $y_k^{t+1} = y_k^t$.

Notice that the x_k update step is different from the conventional proximal update (e.g., [7]). In particular, the linearization is done with respect to x^{t+1} instead of x_k^t computed in the previous iteration. This modification is instrumental in the convergence analysis of Algorithm 3.

Here we use the *period-T essentially cyclic rule* to decide the set $\mathcal{C}^{t+1}$ at each iteration. We have the following remarks regarding to Algorithm 3. Note that there is a slight difference between

Algorithm 3 and Algorithm 2. In Algorithm 3, the block variable x is updated *in every iteration* while in Algorithm 2 the update of x is also governed by block selection rules.

Now we begin analyzing Algorithm 3. We make the following assumptions in this section.

Assumption B. For all k, the stepsize ρ_k is chosen large enough such that:

$$\beta_k := \rho_k - 3L_k - \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 2L_k^2 > 0 \tag{2.40}$$

$$\alpha_k := \frac{\rho_k}{2} - T^2 \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 8L_k^2 > 0 \tag{2.41}$$

$$\rho_k \geq 5L_k,\ k = 1, \cdots, K. \tag{2.42}$$

Again let $t(k)$ denote the last iteration that x_k is updated before $t+1$, i.e.,

$$t(k) = \arg\max_{r \leq t} \{x_k^{r-1} \mid x_k^{r-1} \neq x_k^t\},\ k = 1, \cdots, K. \tag{2.43}$$

Note that we do not need $t(0)$ anymore because the variable x is updated in every iteration. Clearly, we have $x_k^t = x_k^{t(k)}$ and as a result, $y_k^t = y_k^{t(k)}$. We have the following result.

Lemma 2.4 *Suppose Assumption B and Assumptions A1, A3 are satisfied. Then for Algorithm 3, the following is true for the essentially cyclic block selection rule*

$$2L_k^2(4\|x^{t+1} - x^{t(k)}\|^2 + \|x_k^{t+1} - x_k^t\|^2) \geq \|y_k^{t+1} - y_k^t\|^2,\ k = 1, \cdots, K, \tag{2.44}$$

$$2L_k^2(4\|x^{t+1} - x^{t(k)}\|^2 + \|x_k^{t+1} - x_k^{t(k)}\|^2) \geq \|y_k^{t+1} - y_k^{t(k)}\|^2,\ k = 1, \cdots, K. \tag{2.45}$$

Proof. Here we focused on showing the first inequality. The second one uses a similar derivation.

Suppose $k \notin \mathcal{C}^{t+1}$, then the inequality is trivially true, as $y_k^{t+1} = y_k^t$.

For any $k \in \mathcal{C}^{t+1}$, we observe from the update of x_k step (2.38) that the following is true

$$\nabla g_k(x^{t+1}) + y_k^t + (\rho_k + L_k)(x_k^{t+1} - x^{t+1}) = 0,\ k \in \mathcal{C}^{t+1}, \tag{2.46}$$

or equivalently

$$\nabla g_k(x^{t+1}) + L_k(x_k^{t+1} - x^{t+1}) = -y_k^{t+1},\ k \in \mathcal{C}^{t+1}. \tag{2.47}$$

Therefore we have, for all $k \in \mathcal{C}^{t+1}$

$$\begin{aligned}\|y_k^{t+1} - y_k^t\| = \|y_k^{t+1} - y_k^{t(k)}\| &= \|\nabla g_k(x^{t+1}) - \nabla g_k(x^{t(k)}) + L_k(x_k^{t+1} - x^{t+1}) - L_k(x_k^{t(k)} - x^{t(k)})\| \\ &= \|\nabla g_k(x^{t+1}) - \nabla g_k(x^{t(k)}) + L_k(x_k^{t+1} - x^{t+1}) - L_k(x_k^t - x^{t(k)})\| \\ &\leq L_k(2\|x^{t+1} - x^{t(k)}\| + \|x_k^{t+1} - x_k^t\|)\end{aligned}$$

where the last step follows from triangular inequality and the fact $x_k^t = x_k^{t(k)}$ (cf. the definition of $t(k)$). The above result further implies that

$$2L_k^2(4\|x^{t+1} - x^{t(k)}\|^2 + \|x_k^{t+1} - x_k^t\|^2) \ge \|y_k^{t+1} - y_k^t\|^2, \; k = 1, \cdots, K \tag{2.48}$$

which is the desired result. **Q.E.D.**

Next, we upper bound the successive difference of the augmented Lagrangian. To this end, let us define the following functions

$$\ell_k(x_k; x^{t+1}, y^t) = g_k(x_k) + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k - x^{t+1}\|^2$$
$$u_k(x_k; x^{t+1}, y^t) = g_k(x^{t+1}) + \langle \nabla g_k(x^{t+1}), x_k - x^{t+1}\rangle + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k + L_k}{2}\|x_k - x^{t+1}\|^2.$$

Using these short-hand definitions, we have

$$L(\{x_k^{t+1}\}, x^{t+1}; y^t) = \sum_{k=1}^{K} \ell_k(x_k^{t+1}; x^{t+1}, y^t) \tag{2.49}$$
$$x_k^{t+1} = \arg\min_{x_k} u_k(x_k; x^{t+1}, y^t), \; \forall\, k \in \mathcal{C}^{t+1}. \tag{2.50}$$

The lemma below bounds the difference between $\ell_k(x_k^{t+1}; x^{t+1}, y^t)$ and $\ell_k(x_k^t; x^{t+1}, y^t)$.

Lemma 2.5 *Suppose Assumption A1 is satisfied. Let $\{x_k^t, x^t, y^t\}$ be generated by Algorithm 3 with essential cyclic block update rule. Then we have the following*

$$\begin{aligned}&\ell_k(x_k^{t+1}; x^{t+1}, y^t) - \ell_k(x_k^t; x^{t+1}, y^t)\\ &\le -(\rho_k - 3L_k)\|x_k^{t+1} - x_k^t\|^2 + \frac{4L_k}{\rho_k^2}\|y_k^{t+1} - y_k^t\|^2, \quad k = 1, \cdots, K.\end{aligned} \tag{2.51}$$

Proof. When $k \notin \mathcal{C}^{t+1}$, the inequality is trivially true. We focus on the case $k \in \mathcal{C}^{t+1}$. From the definition of $\ell_k(\cdot)$ and $u_k(\cdot)$ we have the following

$$\ell_k(x_k; x^{t+1}, y^t) \le u_k(x_k; x^{t+1}, y^t), \; \forall\, x_k, \; k = 1, \cdots, K. \tag{2.52}$$

Observe that when $k \in \mathcal{C}^{t+1}$, x_k^{t+1} is generated according to (2.50). Due to the strong convexity of $u_k(x_k; x^{t+1}, y^t)$ with respect to x_k, we have

$$u_k(x_k^{t+1}; x^{t+1}, y^t) - u_k(x_k^t; x^{t+1}, y^t) \le -(\rho_k + L_k)\|x_k^t - x_k^{t+1}\|^2, \; \forall\, k \in \mathcal{C}^{t+1}. \tag{2.53}$$

Further, we have the following series of inequalities

$$
\begin{aligned}
& u_k(x_k^t; x^{t+1}, y^t) - \ell_k(x_k^t; x^{t+1}, y^t) \\
& = g_k(x^{t+1}) + \langle \nabla g_k(x^{t+1}), x_k^t - x^{t+1} \rangle + \langle y_k^t, x_k^t - x^{t+1} \rangle + \frac{\rho_k + L_k}{2} \|x_k^t - x^{t+1}\|^2 \\
& \quad - \Big(g_k(x_k^t) + \langle y_k^t, x_k^t - x^{t+1} \rangle + \frac{\rho_k}{2} \|x_k^t - x^{t+1}\|^2 \Big) \\
& = g_k(x^{t+1}) - g_k(x_k^t) + \langle \nabla g_k(x^{t+1}), x_k^t - x^{t+1} \rangle + \frac{L_k}{2} \|x_k^t - x^{t+1}\|^2 \\
& \le \langle \nabla g_k(x^{t+1}) - \nabla g_k(x_k^t), x_k^t - x^{t+1} \rangle + L_k \|x_k^t - x^{t+1}\|^2 \\
& \le 2L_k \|x_k^t - x^{t+1}\|^2 \le 4L_k \left(\|x_k^t - x_k^{t+1}\|^2 + \|x_k^{t+1} - x^{t+1}\|^2 \right),
\end{aligned} \tag{2.54}
$$

where both inequalities follow from Assumption A1. Combining (2.52) – (2.54) we obtain

$$
\begin{aligned}
& \ell_k(x_k^{t+1}; x^{t+1}, y^t) - \ell_k(x_k^t; x^{t+1}, y^t) \\
& \le u_k(x_k^{t+1}; x^{t+1}, y^t) - u_k(x_k^t; x^{t+1}, y^t) + u_k(x_k^t; x^{t+1}, y^t) - \ell_k(x_k^t; x^{t+1}, y^t) \\
& \le -(\rho_k - 3L_k) \|x_k^t - x_k^{t+1}\|^2 + 4L_k \|x_k^{t+1} - x^{t+1}\|^2 \\
& = -(\rho_k - 3L_k) \|x_k^t - x_k^{t+1}\|^2 + \frac{4L_k}{\rho_k^2} \|y_k^{t+1} - y_k^t\|^2, \ \forall\, k \in \mathcal{C}^{t+1}.
\end{aligned}
$$

The desired result then follows. **Q.E.D.**

Next, we bound the difference of the augmented Lagrangian function values.

Lemma 2.6 *Assume the same set up as in Lemma 2.5. Then we have*

$$
\begin{aligned}
& L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^1\}, x^1; y^1) \\
& \le -\sum_{i=1}^{t} \sum_{k=1}^{K} \beta_k \|x_k^{i+1} - x_k^i\|^2 - \sum_{i=1}^{t} \sum_{k=1}^{K} \alpha_k \|x^{i+1} - x^i\|^2
\end{aligned} \tag{2.55}
$$

where we β_k and α_k are the positive constants defined in (2.40) *and* (2.41).

Proof. We first bound the successive difference $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t)$. Again we decompose it as in (2.23), and bound the resulting two differences separately.

The first term in (2.23) can be again expressed as

$$
L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^{t+1}\}, x^{t+1}; y^t) = \sum_{k=1}^{K} \frac{1}{\rho_k} \|y_k^{t+1} - y_k^t\|^2.
$$

To bound the second term in (2.23), we use Lemma 2.5. We use an argument similar to the proof of (2.25) to obtain

$$
\begin{aligned}
& L(\{x_k^{t+1}\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& = L(\{x_k^{t+1}\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^{t+1}; y^t) + L(\{x_k^t\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& = \sum_{k=1}^K \left(\ell_k(x_k^{t+1}; x^{t+1}, y^t) - \ell_k(x_k^t; x^{t+1}, y^t)\right) + L(\{x_k^t\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& \le -\sum_{k=1}^K \left((\rho_k - 3L_k)\|x_k^{t+1} - x_k^t\|^2 - \frac{4L_k}{\rho_k^2}\|y_k^{t+1} - y_k^t\|^2\right) - \frac{\gamma}{2}\|x^{t+1} - x^t\|^2
\end{aligned} \tag{2.56}
$$

where the last inequality follows from Lemma 2.5 and the strong convexity of $L(\{x_k^t\}, x; y^t)$ with respect to the variable x (with modulus $\gamma = \sum_{k=1}^K \rho_k$) at $x = x^{t+1}$.

Combining the above two inequalities, we obtain

$$
\begin{aligned}
& L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \\
& \le \sum_{k=1}^K \left(-(\rho_k - 3L_k)\|x_k^{t+1} - x_k^t\|^2 + \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right)\|y_k^{t+1} - y_k^t\|^2\right) - \frac{\gamma}{2}\|x^{t+1} - x^t\|^2 \\
& \overset{\text{(a)}}{\le} \sum_{k=1}^K \left(-(\rho_k - 3L_k)\|x_k^{t+1} - x_k^t\|^2 + \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 2L_k^2 (4\|x^{t+1} - x^{t(k)}\|^2 + \|x_k^{t+1} - x_k^t\|^2)\right) - \frac{\gamma}{2}\|x^{t+1} - x^t\|^2 \\
& \overset{\text{(b)}}{=} -\sum_{k=1}^K \left(\rho_k - 3L_k - \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 2L_k^2\right)\|x_k^{t+1} - x_k^t\|^2 - \sum_{k=1}^K \left(\frac{\rho_k}{2}\right)\|x^{t+1} - x^t\|^2 \\
& \qquad + \sum_{k=1}^K \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 8L_k^2 \|x^{t(k)} - x^{t+1}\|^2 \\
& \le -\sum_{k=1}^K \left(\rho_k - 3L_k - \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 2L_k^2\right)\|x_k^{t+1} - x_k^t\|^2 - \sum_{k=1}^K \left(\frac{\rho_k}{2}\right)\|x^{t+1} - x^t\|^2 \\
& \qquad + \sum_{k=1}^K T\left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right) 8L_k^2 \sum_{i=0}^{\min\{T-1, t-1\}} \|x^{t-i+1} - x^{t-i}\|^2
\end{aligned} \tag{2.57}
$$

where in (a) we have used (2.44); in (b) we have used the fact that $\gamma = \sum_{k=1}^K \rho_k$; in the last inequality we have used the definition of the period-T essentially cyclic update rule which implies that

$$
\|x^{t+1} - x^{t(k)}\| \le \sum_{i=0}^{\min\{T-1, t-1\}} \|x^{t-i+1} - x^{t-i}\| \Longrightarrow \|x^{t+1} - x^{t(k)}\|^2 \le T \sum_{i=0}^{\min\{T-1, t-1\}} \|x^{t-i+1} - x^{t-i}\|^2.
$$

Then for any given t, the difference $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^1\}, x^1; y^1)$ is obtained by summing (2.57) over all iterations. Specifically, we obtain

$$
\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^1\}, x^1; y^1) \\
&\le -\sum_{i=1}^{t}\sum_{k=1}^{K}\left(\rho_k - 3L_k - \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right)2L_k^2\right)\|x_k^{i+1} - x_k^i\|^2 \\
&\qquad - \sum_{i=1}^{t}\sum_{k=1}^{K}\left(\frac{\rho_k}{2} - T^2\left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right)8L_k^2\right)\|x^{i+1} - x^i\|^2 \\
&= -\sum_{i=1}^{t}\sum_{k=1}^{K}\alpha_k\|x_k^{i+1} - x_k^i\|^2 - \sum_{i=1}^{t}\sum_{k=1}^{K}\beta_k\|x^{i+1} - x^i\|^2.
\end{aligned}
$$

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

We conclude that to make the augmented Lagrangian decrease at each iteration, it is sufficient to require that $\alpha_k > 0$ and $\beta_k > 0$ for all k, or more specifically:

$$
\begin{aligned}
&\rho_k - 3L_k - \left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right)2L_k^2 > 0,\ k = 1, \cdots, K, \\
&\frac{\rho_k}{2} - T^2\left(\frac{4L_k}{\rho_k^2} + \frac{1}{\rho_k}\right)8L_k^2 > 0,\ k = 1, \cdots, K.
\end{aligned}
\tag{2.58}
$$

Note that one can always find a set of ρ_k's large enough such that the above condition is satisfied.

Next we show that $L(\{x_k^t\}, x^t; y^t)$ is convergent.

Lemma 2.7 *Suppose Assumption B is satisfied. Then Algorithm 3 with period-T essentially cyclic update rule generates a sequence of augmented Lagrangian that satisfies*

$$
\lim_{t\to\infty} L(\{x_k^t\}, x^t, y^t) = L^* > -\infty \tag{2.59}
$$

for some constant L^*.

Proof. Observe that the augmented Lagrangian can be expressed as

$$
\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})\\
&= h(x^{t+1}) + \sum_{k=1}^{K}\left(g_k(x_k^{t+1}) + \langle y_k^{t+1}, x_k^{t+1} - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\right)\\
&\overset{\text{(a)}}{=} h(x^{t+1}) + \sum_{k=1}^{K}\left(g_k(x_k^{t+1}) + \langle \nabla g_k(x^{t+1}) + L_k(x_k^{t+1} - x^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{\rho_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\right)\\
&= h(x^{t+1}) + \sum_{k=1}^{K}\left(g_k(x_k^{t+1}) + \langle \nabla g_k(x^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{\rho_k - 2L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\right)\\
&\overset{\text{(b)}}{\geq} h(x^{t+1}) + \sum_{k=1}^{K}\left(g_k(x^{t+1}) + \frac{\rho_k - 5L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\right)\\
&= f(x^{t+1}) + \sum_{k=1}^{K}\frac{\rho_k - 5L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2 \qquad (2.60)
\end{aligned}
$$

where (a) is from (2.47); (b) is due to the following inequalities

$$
\begin{aligned}
g_k(x^{t+1}) &\leq g_k(x_k^{t+1}) + \langle \nabla g_k(x_k^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\\
&= g_k(x_k^{t+1}) + \langle \nabla g_k(x_k^{t+1}) - \nabla g_k(x^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \langle \nabla g_k(x^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2\\
&\leq g_k(x_k^{t+1}) + \langle \nabla g_k(x^{t+1}), x^{t+1} - x_k^{t+1}\rangle + \frac{3L_k}{2}\|x_k^{t+1} - x^{t+1}\|^2.
\end{aligned}
$$

Clearly, combining the inequality (2.60) with Assumptions B and A3 yields that $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})$ is lower bounded. It follows from Lemma 2.6 that whenever the stepsize ρ_k's are chosen sufficiently large (as per Assumption B), $L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1})$ will monotonically decrease and is convergent. This completes the proof. **Q.E.D.**

Using Lemmas 2.4–2.7, we arrive at the following convergence result. The proof is similar to Theorem 2.1, and is thus omitted.

Theorem 2.2 *Suppose that Assumptions A1, A3 and B hold. Then the following is true for Algorithm 3.*

1. *We have* $\lim_{t\to\infty}\|x^{t+1} - x_k^{t+1}\| = 0$, $k = 1, \cdots, K$.
2. *Let* $(\{x_k^*\}, x^*, y^*)$ *denote any limit point of the sequence* $\{\{x_k^{t+1}\}, x^{t+1}, y^{t+1}\}$ *generated by Algorithm 3 with period-*T *essentially cyclic block update rule. Then* $(\{x_k^*\}, x^*, y^*)$ *is a stationary solution of problem* (2.2).
3. *If* X *is a compact set, then Algorithm 3 with period-*T *essentially cyclic block update rule converges to the set of stationary solutions of problem* (2.2).

3 The Nonconvex Sharing Problem

Consider the following well-known sharing problem (see, e.g., [7, Section 7.3] for motivation)

$$\begin{aligned} \min \quad & f(x_1, \cdots, x_K) := \sum_{k=1}^{K} g_k(x_k) + \ell\left(\sum_{k=1}^{K} A_k x_k\right) \\ \text{s.t.} \quad & x_k \in X_k,\ k = 1, \cdots, K, \end{aligned} \tag{3.1}$$

where $x_k \in \mathbb{R}^{N_k}$ is the variable associated with a given agent k, and $A_k \in \mathbb{R}^{M \times N_k}$ is some data matrix. The variables are coupled through the function $\ell(\cdot)$.

To facilitate distributed computation, this problem can be equivalently formulated into a linearly constrained problem by introducing an additional variable $x \in \mathbb{R}^M$:

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} g_k(x_k) + \ell\left(x\right) \\ \text{s.t.} \quad & \sum_{k=1}^{K} A_k x_k = x, \quad x_k \in X_k,\ k = 1, \cdots, K. \end{aligned} \tag{3.2}$$

The augmented Lagrangian for this problem is given by

$$L(\{x_k\}, x; y) = \sum_{k=1}^{K} g_k(x_k) + \ell(x) + \left\langle x - \sum_{k=1}^{K} A_k x_k, y \right\rangle + \frac{\rho}{2}\left\| x - \sum_{k=1}^{K} A_k x_k \right\|^2. \tag{3.3}$$

Note that we have chosen a special reformulation in (3.2): a *single* variable x is introduced which leads to a problem with a *single* linear constraint. Applying the classical ADMM to this reformulation leads to a *multi-block* ADMM algorithm in which $K+1$ block variables $(\{x_k\}_{k=1}^K, x)$ are updated sequentially. As mentioned in the introduction, even in the case where the objective is convex, it is not known whether the multi-block ADMM converges in this case. Variants of the multi-block ADMM has been proposed in the literature to solve this type of multi-block problems; see recent developments in [24–28] and the references therein.

In this section, we show that the classical ADMM, together with several of its extensions using different block selection rules, converge even when the objective function is nonconvex. The main assumptions for convergence are that the stepsize ρ is large enough, and that the coupling function $\ell(x)$ should be smooth (more detailed conditions will be given shortly). Similarly as in the previous sections, we consider a generalized version of ADMM with two types of block update rules: the period-T essentially cyclic rule and the randomized rule. The detailed algorithm is given in the table below.

Algorithm 4. The Flexible ADMM for the Sharing Problem (3.2)

At each iteration $t \geq 1$, pick an index set $\mathcal{C}^{t+1} \in \{0, \cdots, K\}$.

For $k = 1, \cdots, K$

If $k \in \mathcal{C}^{t+1}$, then agent k updates x_k by:

$$x_k^{t+1} = \arg\min_{x_k \in X_k} g_k(x_k) - \langle y^t, A_k x_k \rangle + \frac{\rho}{2} \left\| x^t - \sum_{j<k} A_j x_j^{t+1} - \sum_{j>k} A_j x_j^t - A_k x_k \right\|^2 \quad (3.4)$$

Else $x_k^{t+1} = x_k^t$.

If $0 \in \mathcal{C}^{t+1}$, update the variable x by:

$$x^{t+1} = \arg\min_x \ell(x) + \langle y^t, x \rangle + \frac{\rho}{2} \left\| x - \sum_{k=1}^{K} A_k x_k^{t+1} \right\|^2. \quad (3.5)$$

Update the dual variable:

$$y^{t+1} = y^t + \rho \left(x^{t+1} - \sum_{k=1}^{K} A_k x_k^{t+1} \right). \quad (3.6)$$

Else $x^{t+1} = x^t$, $y^{t+1} = y^t$.

The analysis of Algorithm 4 follows similar argument as that of Algorithm 3. Therefore we will only provide an outline for it.

First, we make the following assumptions in this section.

Assumption C.

C1. There exists a positive constant $L > 0$ such that

$$\|\nabla \ell(x) - \nabla \ell(z)\| \leq L \|x - z\|, \ \forall\, x, z.$$

Moreover, X_k's are closed convex sets; each A_k is full column rank, with $\rho_{\min}(A_k^T A_k) > 0$.

C2. The stepsize ρ is chosen large enough such that:

(1) Each x_k subproblem (3.4) as well as the x subproblem (3.5) is strongly convex, with modulus $\{\gamma_k(\rho)\}_{k=1}^K$ and $\gamma(\rho)$, respectively.

(2) $\rho\gamma(\rho) > 2L^2$, and that $\rho \geq L$.

C3. $f(x_1, \cdots, x_K)$ is lower bounded over $\prod_{k=1}^K X_k$.

C4. g_k is either smooth nonconvex or convex (possibly nonsmooth). For the former case, there exists $L_k > 0$ such that $\|\nabla g_k(x_k) - \nabla g_k(z_k)\| \leq L_k \|x_k - z_k\|$, $\forall\, x_k, z_k \in X_k$.

Note that compared with Assumptions A and B, in this case we no longer require that each g_k to be smooth. Define an index set $\mathcal{K} \subseteq \{1, \cdots, K\}$, such that g_k is convex if $k \in \mathcal{K}$, and nonconvex smooth otherwise. Further, the requirement that A_k is full column rank is needed to make the x_k subproblem (3.4) strongly convex.

Our convergence analysis consists of a series of lemmas whose proofs, for the most part, are omitted since they are similar to that of Lemma 2.1–Lemma 2.3.

Lemma 3.1 *Suppose Assumption C is satisfied. Then for Algorithm 4 with either essentially cyclic rule or the randomized rule, the following is true*

$$\nabla \ell(x^{t+1}) = -y^{t+1}, \text{ if } 0 \in \mathcal{C}^{t+1}$$
$$L^2\|x^{t+1} - x^t\|^2 \geq \|y^{t+1} - y^t\|^2,\ L^2\|x^{t+1} - x^{t(k)}\|^2 \geq \|y^{t+1} - y^{t(k)}\|^2,\ L^2\|\hat{x}^{t+1} - x^t\|^2 \geq \|\hat{y}^{t+1} - y^t\|^2.$$

Lemma 3.2 *Suppose Assumption C is satisfied. Then for Algorithm 4 with either essentially cyclic rule or the randomized rule, the following is true*

$$\begin{aligned}
&L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_k^t\}, x^t; y^t) \\
&\leq \sum_{k\neq 0, k\in \mathcal{C}^{t+1}} -\frac{\gamma_k(\rho)}{2}\|x_k^{t+1} - x_k^t\|^2 - \iota\{0 \in \mathcal{C}^{t+1}\}\left(\frac{\gamma(\rho)}{2} - \frac{L^2}{\rho}\right)\|x^{t+1} - x^t\|^2.
\end{aligned} \tag{3.7}$$

Lemma 3.3 *Assume the same set up as in Lemma 3.2. Then the following is true*

$$\lim_{t\to\infty} L(\{x_k^{t+1}\}, x^{t+1}; y^{t+1}) = L^* > -\infty \tag{3.8}$$

for some constant L^*.

Proof. We have the following series of inequalities

$$\begin{aligned}
&L(x^{r+1}, \{x^{r+1}; y^{r+1}\}) \\
&= \sum_{k=1}^K g_k(x_k^{t+1}) + \ell(x^{t+1}) + \left\langle x^{t+1} - \sum_{k=1}^K A_k x_k^{t+1}, y^{t+1}\right\rangle + \frac{\rho}{2}\left\|x^{t+1} - \sum_{k=1}^K A_k x_k^{t+1}\right\|^2 \\
&= \sum_{k=1}^K g_k(x_k^{t+1}) + \ell(x^{t+1}) + \left\langle \sum_{k=1}^K A_k x_k^{t+1} - x^{t+1}, \nabla \ell(x^{t+1})\right\rangle + \frac{\rho}{2}\left\|x^{t+1} - \sum_{k=1}^K A_k x_k^{t+1}\right\|^2 \\
&\geq \sum_{k=1}^K g_k(x_k^{t+1}) + \ell\left(\sum_{k=1}^K A_k x_k^{t+1}\right) + \frac{\rho - L}{2}\left\|x^{t+1} - \sum_{k=1}^K A_k x_k^{t+1}\right\|^2.
\end{aligned}$$

The last inequality comes from the fact that

$$\ell\left(\sum_{k=1}^{K} A_k x_k^{t+1}\right) \le \ell(x^{t+1}) + \left\langle \sum_{k=1}^{K} A_k x_k^{t+1} - x^{t+1}, \nabla \ell(x^{t+1}) \right\rangle + \frac{L}{2}\left\| x^{t+1} - \sum_{k=1}^{K} A_k x_k^{t+1} \right\|^2.$$

Using assumptions C2.– C3. leads to the desired result. **Q.E.D.**

Theorem 3.1 *Suppose that Assumption C holds. Then the following is true for Algorithm 4, either deterministically for the essentially cyclic update rule or almost surely for the randomized update rule.*

1. *We have* $\lim_{t\to\infty} \|x_k^{t+1} - x^{t+1}\| = 0$, $k = 1, \cdots, K$.
2. *Let* $(\{x_k^*\}, x^*, y^*)$ *denote any limit point of the sequence* $\{\{x_k^{t+1}\}, x^{t+1}, y^{t+1}\}$ *generated by Algorithm 4. Then* $(\{x_k^*\}, x^*, y^*)$ *is a stationary solution of problem* (3.2) *in the sense that*

$$\begin{aligned}
& x_k^* \in \arg\min_{x_k \in X_k} \; g_k(x_k) + \langle y^*, -A_k x_k \rangle, \; k \in \mathcal{K}, \\
& \left\langle x_k - x_k^*, \nabla g_k(x_k^*) - A_k^T y^* \right\rangle \ge 0, \; \forall\, x_k \in X_k, \; k \notin \mathcal{K}, \\
& \nabla \ell(x^*) + y^* = 0, \\
& \sum_{k=1}^{K} A_k x_k^* = x^*.
\end{aligned}$$

3. *If* X_k *is a compact set for all* k*, then Algorithm 4 converges to the set of stationary solutions of problem* (3.2)*.*

The following corollary specializes the previous convergence result to the case where all g_k's as well as ℓ are convex (but not necessarily strongly convex). We emphasize that this is still a nontrivial result, since unlike [25, 27, 29, 32], we do not require the dual stepsize to be small or the g_k's and ℓ to be strongly convex. Therefore it is not known whether the classical ADMM converges for the multi-block problem (3.2), even for the convex case.

Corollary 3.1 *Suppose that Assumptions C1 and C3 hold, and that* g_k *and* ℓ *are convex. Further, suppose that Assumption C2 is weakened with the following assumption*

1. *The stepsize* ρ *is chosen large enough such that* $\rho > \sqrt{2}L$.

Then the flexible ADMM algorithm (i.e., Algorithm 4), converges to the set primal dual optimal solution $(\{x_k^*\}, x^*, y^*)$ *of problem* (2.2)*, either deterministically for the essentially cyclic update rule or almost surely for the randomized update rule.*

Similar to the consensus problem, one can extend Algorithm 4 to its proximal version. Here the benefit offered by the proximal-type algorithms is twofold: *i)* one can remove the strong convexity requirement posed in Assumption C2-(1) ; *ii)* one can allow inexact and simple update for each block variable. However, the analysis is a bit more involved, as the stepsize ρ as well as the proximal coefficient for each subproblem needs to be carefully bounded. Due to the fact that the analysis follows almost identical steps as those in Section 2.3, we will not present them here.

4 Extensions

In this paper, we analyze the behavior of the ADMM method in the absence of convexity. We show that when the stepsize is chosen sufficiently large, the ADMM and several of its variants converge to the set of stationary solutions for certain consensus and sharing problems.

Our analysis is based on using the augmented Lagrangian as a potential function to guide the iterate convergence. This approach may be extended to other nonconvex problems. In particular, if the following set of sufficient conditions (see Assumption D below) are satisfied, then the convergence of the ADMM is guaranteed for the nonconvex problem (1.1). In practice these conditions should be verified case by case for different applications, just like what we have done for the consensus and sharing problems.

Assumption D

D1. The iterations are well defined, meaning the function $L(x^t; y^t)$ is lower bounded for all t.

D2. There exists a constant $\sigma > 0$ such that $\|y^{t+1} - y^t\|^2 \le \sigma \|x^{t+1} - x^t\|^2$, for all t.

D3. $g_k(\cdot)$ is either smooth nonconvex or nonsmooth convex. The coupling function $\ell(\cdot)$ is smooth with Lipschitz continuous gradient L. Moreover, $\ell(\cdot)$ is convex with respect to each block variable x_k, but is not necessarily jointly convex with x. X_k is a closed convex set. Problem (1.1) is feasible, that is, $\{x \mid Ax = q\} \bigcap_{k=1}^{K} \mathrm{relint} X_k \neq \emptyset$.

D4. The stepsize ρ is chosen large enough such that each subproblem is strongly convex with modulus $\gamma_k(\rho)$, which is a nondecreasing function of ρ. Further, $\rho\gamma_k(\rho) > 2\sigma$ for all k.

Following a similar argument leading to Theorem 2.1, we can show that as long as Assumption D is satisfied, then the primal feasibility $\|q - \sum_{k=1}^{K} A_k x_k^{t+1}\|$ will be reached in the limit, and that every limit point of the sequence $\{\{x_k^{t+1}\}, x^{t+1}, y^{t+1}\}$ is a stationary solution of problem (1.1). The main drawback of Assumption D is that it is made on the iterates rather than on the problem. For different linearly constrained optimization problem, one still needs to verify that these conditions are indeed valid, as we have done for the consensus and the sharing problem considered in this paper.

References

[1] R. Glowinski and A. Marroco, "Sur l'approximation, par elements finis d'ordre un,et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares," *Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle*, vol. 9, pp. 41–76, 1975.

[2] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, pp. 17–40, 1976.

[3] J. Eckstein, "Splitting methods for monotone operators with applications to parallel optimization," 1989, Ph.D Thesis, Operations Research Center, MIT.

[4] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[5] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer-Verlag, New York, 1984.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods, 2nd ed*, Athena Scientific, Belmont, MA, 1997.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, 2011.

[8] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for l1-minimization with applications to compressed sensing," *SIAM Journal on Imgaging Science*, vol. 1, no. 1, pp. 143–168, Mar. 2008.

[9] J. Yang, Y. Zhang, and W. Yin, "An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise," *SIAM Journal on Scientific Computing*, vol. 31, no. 4, pp. 2842–2865, 2009.

[10] X. Zhang, M. Burger, and S. Osher, "A unified primal-dual algorithm framework based on Bregman iteration," *Journal of Scientific Computing*, vol. 46, no. 1, pp. 20–46, 2011.

[11] K. Scheinberg, S. Ma, and D. Goldfarb, "Sparse inverse covariance selection via alternating linearization methods," in *Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[12] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc wsns with noisy links - part i: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350 – 364, 2008.

[13] C. Feng, H. Xu, and B. Li, "An alternating direction method approach to cloud traffic management," 2014, preprint.

[14] W.-C. Liao, M. Hong, Hamid Farmanbar, Xu Li, Z.-Q. Luo, and Hang Zhang, "Min flow rate maximization for software defined radio access networks," *IEEE Journal on Selected Areas in Communication*, vol. 32, no. 6, pp. 1282–1294, 2014.

[15] D. P. Bertsekas, *Nonlinear Programming, 2nd ed*, Athena Scientific, Belmont, MA, 1999.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[17] A. Nedic and A. Ozdaglar, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2009.

[18] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Method*, Academic Press, 1982.

[19] B. He and X. Yuan, "On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method," *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[20] R. Monteiro and B. Svaiter, "Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers," *SIAM Journal on Optimization*, vol. 23, no. 1, pp. 475–507, 2013.

[21] T. Goldstein, B. O'Donoghue, and S. Setzer, "Fast alternating direction optimization methods," *UCLA CAM technical report*, 2012.

[22] D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions," *Mathematical Programming*, vol. 141, no. 1-2, pp. 349–382, 2012.

[23] W. Deng and W. Yin, "On the global linear convergence of alternating direction methods," 2012, preprint.

[24] C. Chen, B. He, X. Yuan, and Y. Ye, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," 2013, preprint.

[25] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *arXiv preprint arXiv:1208.3922*, 2012.

[26] B. He, M. Tao, and X. Yuan, "Alternating direction method with Gaussian back substitution for separable convex programming," *SIAM Journal on Optimization*, vol. 22, pp. 313–340, 2012.

[27] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo, "A block successive upper bound minimization method of multipliers for linearly constrained convex optimization," 2013, Preprint, available online arXiv:1401.7079.

[28] X. Wang, M. Hong, S. Ma, and Z.-Q. Luo, "Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers," 2013, preprint.

[29] D. Han and X. Yuan, "A note on the alternating direction method of multipliers," *Journal of Optimization Theory and Applications*, vol. 155, no. 1, pp. 227–238, 2012.

[30] W. Deng, M. Lai, Z. Peng, and W. Yin, "Parallel multi-block ADMM with o(1/k) convergence," Preprint, available online at arXiv: 1312.3040., 2014.

[31] B. He, H. Xu, and X. Yuan, "On the proximal jacobian decomposition of alm for multipleblock separable convex minimization problems and its relationship to ADMM," 2013, Preprint, available on Optimization-Online.

[32] T. Lin, S. Ma, and S. Zhang, "On the global linear convergence of the admm with multi-block variables," 2014, Preprint.

[33] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo, "A block coordinate descent method of multipliers: Convergence analysis and applications," in *International Conference on Acoustics, Speech and Signal Processing*, 2014.

[34] Y. Zhang, "An alternating direction algorithm for nonnegative matrix factorization," 2010, Preprint.

[35] D. L. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[36] Z. Wen, C. Yang, X. Liu, and S. Marchesini, "Alternating direction methods for classical and ptychographic phase retrieval," *Inverse Problems*, vol. 28, no. 11, pp. 1–18, 2012.

[37] R. Zhang and J. T. Kwok, "Asynchronous distributed admm for consensus optimization," in *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[38] P.A. Forero, A. Cano, and G.B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 707–724, Aug 2011.

[39] B. Ames and M. Hong, "Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis," 2014, Preprint.

[40] B. Jiang, S. Ma, and S. Zhang, "Alternating direction method of multipliers for real and complex polynomial optimization models," 2013, Preprint.

[41] A. P. Liavas and N. D. Sidiropoulos, "Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers," 2014, Preprint, available at arXiv:1409.2383v1.

[42] Y. Shen, Z. Wen, and Y. Zhang, "Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization," *Optimization Methods Software*, vol. 29, no. 2, pp. 239–263, Mar. 2014.

[43] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Journal of Frontiers of Mathematics in China, Special Issues on Computational Mathematics*, pp. 365–384, 2011.

[44] Z. Wen, X. Peng, X. Liu, X. Bai, and X. Sun, "Asset allocation under the basel accord risk measures," 2013, Preprint.

[45] Y. Zhang, "Convergence of a class of stationary iterative methods for saddle point problems," 2010, Preprint.

[46] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[47] E. Wei and A. Ozdaglar, "On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers," 2013, Preprint, available at arXiv:1307.8254.

[48] T.-H. Chang, "A proximal dual consensus admm method for multi-agent constrained optimization," 2014, Preprint, available at arXiv:1409.3307.

[49] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.