Structured Nonconvex and Nonsmooth Optimization: Algorithms and Iteration Complexity Analysis

Bo Jiang * Tianyi Lin † Shiqian Ma ‡ Shuzhong Zhang §

May 06, 2016

Abstract

Nonconvex and nonsmooth optimization problems are frequently encountered in much of statistics, business, science and engineering, but they are not yet widely recognized as a *technology* in the sense of scalability. A reason for this relatively low degree of popularity is the lack of a well developed system of theory and algorithms to support the applications, as is the case for its convex counterpart. This paper aims to take one step in the direction of *disciplined nonconvex and nonsmooth optimization*. In particular, we consider in this paper some constrained nonconvex optimization models in block decision variables, with or without coupled affine constraints. In the case of no coupled constraints, we show a sublinear rate of convergence to an ϵ-stationary solution in the form of variational inequality for a generalized conditional gradient method, where the convergence rate is shown to be dependent on the Hölderian continuity of the gradient of the smooth part of the objective. For the model with coupled affine constraints, we introduce corresponding ϵ-stationarity conditions, and propose two proximal-type variants of the ADMM to solve such a model, assuming the proximal ADMM updates can be implemented for all the block variables except for the last block, for which either a gradient step or a majorization-minimization step is implemented. We show an iteration complexity bound of $O(1/\epsilon^2)$ to reach an ϵ-stationary solution for both algorithms. Moreover, we show that the same iteration complexity of a proximal BCD method follows immediately. Numerical results are provided to illustrate the efficacy of the proposed algorithms for tensor robust PCA.

Keywords: Structured Nonconvex Optimization, ϵ-Stationary Point, Iteration Complexity, Conditional Gradient Method, Alternating Direction Method of Multipliers, Block Coordinate Descent Method

Mathematics Subject Classification: 90C26, 90C06, 90C60

*Research Center for Management Science and Data Analytics, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China. Research of this author was supported in part by National Natural Science Foundation of China (Grant 11401364).

†Research Center for Management Science and Data Analytics, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. Research of this author was supported in part by the Hong Kong Research Grants Council General Research Fund (Grant 14205314).

§Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455. Research of this author was supported in part by the National Science Foundation (Grant CMMI-1462408).

1 Introduction

In this paper, we consider the following nonconvex and nonsmooth optimization problem with multiple block variables:

$$\begin{array}{ll} \min & f(x_1, x_2, \cdots, x_N) + \sum\limits_{i=1}^{N-1} r_i(x_i) \\ \text{s.t.} & \sum_{i=1}^N A_i x_i = b, \ x_i \in \mathcal{X}_i, \ i = 1, \ldots, N-1, \end{array} \tag{1.1}$$

where f is differentiable and possibly nonconvex, and each r_i is possibly nonsmooth and nonconvex, $i = 1, \ldots, N-1$; $A_i \in \mathbb{R}^{m \times n_i}$, $b \in \mathbb{R}^m$, $x_i \in \mathbb{R}^{n_i}$; and $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ are convex sets, $i = 1, 2, \ldots, N-1$. A special case of (1.1) is when the affine constraints are absent, and there is no block structure of the variables, which leads to the following more compact form

$$\min \ \Phi(x) := f(x) + r(x), \ \text{s.t.} \ x \in S \subset \mathbb{R}^n, \tag{1.2}$$

where S is a convex and compact set. In this paper, we propose several first-order algorithms for computing an ϵ-stationary point (to be defined later) of (1.1) and (1.2), and analyze their iteration complexities. Throughout this paper, we assume that the sets of the stationary points to (1.1) and (1.2) are non-empty.

Problem (1.1) arises from a variety of interesting applications. For example, one of the nonconvex models for matrix robust PCA can be casted as follows (see, e.g., [45]), which seeks to decompose a given matrix $M \in \mathbb{R}^{m \times n}$ to a superposition of a low-rank matrix Z, a sparse matrix E and a noise matrix B:

$$\begin{array}{ll} \min_{X,Y,Z,E,B} & \|Z - XY^\top\|_F^2 + \alpha \mathcal{R}(E) \\ \text{s.t.} & M = Z + E + B \\ & \|B\|_F \leq \eta, \end{array} \tag{1.3}$$

where $X \in \mathbb{R}^{m \times r}$, $Y \in \mathbb{R}^{n \times r}$, with $r < \min(m, n)$ being the estimated rank of Z; $\eta > 0$ is the noise level, $\alpha > 0$ is a weighting parameter; $\mathcal{R}(E)$ is a regularization function that can improve the sparsity of E. One of the widely used regularization functions is the ℓ_1 norm, which is convex and nonsmooth. However, there are also many nonconvex regularization functions that are widely used in statistical learning and information theory, such as smoothly clipped absolute deviation (SCAD) [21], log-sum penalty (LSP) [15], minimax concave penalty (MCP) [50], and capped-ℓ_1 penalty [51, 52], and they are nonsmooth at point 0 if composed with the absolute value function, which is usually the case in statistical learning. Clearly (1.3) is in the form of (1.1). Another example of the form (1.1) is the following nonconvex tensor robust PCA model (see, e.g., [48]), which seeks to decompose a given tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ into a superposition of a low-rank tensor $\mathcal{Z}$, a sparse tensor $\mathcal{E}$ and a noise tensor $\mathcal{B}$:

$$\begin{array}{ll} \min_{X_i,\mathcal{C},\mathcal{Z},\mathcal{E},\mathcal{B}} & \|\mathcal{Z} - \mathcal{C} \times_1 X_1 \times_2 X_2 \times_3 \cdots \times_d X_d\|_F^2 + \alpha \mathcal{R}(\mathcal{E}) \\ \text{s.t.} & \mathcal{T} = \mathcal{Z} + \mathcal{E} + \mathcal{B} \\ & \|\mathcal{B}\|_F \leq \eta, \end{array} \tag{1.4}$$

where $\mathcal{C}$ is the core tensor that has smaller size than $\mathcal{Z}$, and X_i are matrices with appropriate sizes, $i = 1, \ldots, d$. In fact, the "low-rank" tensor in the above model corresponds to the tensor with a small core; however a recent work [32] demonstrates that the CP-rank of the core regardless of its size could be as large as the original tensor. Therefore, if one wants to find the low CP-rank decomposition, then the following model is preferred:

$$\begin{array}{ll} \min_{X_i,\mathcal{Z},\mathcal{E},\mathcal{B}} & \|\mathcal{Z} - [\![X_1, X_2, \cdots, X_d]\!]\|^2 + \alpha\, \mathcal{R}(\mathcal{E}) + \|\mathcal{B}\|^2 \\ \text{s.t.} & \mathcal{T} = \mathcal{Z} + \mathcal{E} + \mathcal{B}, \end{array} \tag{1.5}$$

for $X_i = [a^{i,1}, a^{i,2}, \cdots, a^{i,R}] \in \mathbb{R}^{n_i \times R}$, $1 \le i \le d$ and

$$[\![X_1, X_2, \cdots, X_d]\!] := \sum_{r=1}^{R} a^{1,r} \otimes a^{2,r} \otimes \cdots \otimes a^{d,r}, \tag{1.6}$$

where"$\otimes$" denotes the outer product of vectors, and R is an estimation of the CP-rank. In addition, the so-called sparse tensor PCA problem [1], which seeks the best sparse rank-one approximation for a given d-th order tensor $\mathcal{T}$, can also be formulated in the form of (1.1):

$$\begin{array}{ll} \min & -\mathcal{T}(x_1, x_2, \cdots, x_d) + \alpha \sum_{i=1}^{d} \mathcal{R}(x_i) \\ \text{s.t.} & x_i \in S_i = \{x \mid \|x\|_2^2 \le 1\},\ i = 1, 2, \ldots, d, \end{array} \tag{1.7}$$

where $\mathcal{T}(x_1, x_2, \cdots, x_d) = \sum_{i_1,\ldots,i_d} \mathcal{T}_{i_1,\ldots,i_d} (x_1)_{i_1} \cdots (x_d)_{i_d}$.

The convergence and iteration complexity for various nonconvex and nonsmooth optimization problems have recently attracted considerable research attention; see e.g. [3, 6–8, 10, 11, 19, 24, 25, 38]. In this paper, we propose several solution methods that use only the first-order information of the objective function, including generalized conditional gradient method, variants of alternating direction method of multipliers, and proximal block coordinate descent method, for solving (1.1) and (1.2). Specifically, we propose a generalized conditional gradient (GCG) method for solving (1.2). We prove that GCG can find an ϵ-stationary point for (1.2) in $O(\epsilon^{-q})$ iterations under certain mild conditions, where q is a parameter in the Hölder condition that characterizes the degree of smoothness of f. In other words, the speed of the algorithm's convergence depends on the degree of "smoothness" of the objective function. It should be noted that a similar iteration bound that depends on the parameter q was only recently reported in the context of convex optimization [13]. Furthermore, we show that if f is concave, then GCG finds an ϵ-stationary point for (1.2) in $O(1/\epsilon)$ iterations. For the affinely constrained problem (1.1), we propose two algorithms (called proximal ADMM-g and proximal ADMM-m in this paper) that can both be viewed as variants of the alternating direction method of multipliers (ADMM). Recently, there have been some emerging interests on the ADMM for nonconvex problems (see, e.g., [2, 29, 30, 35, 46, 47, 49]). The results in [35, 46, 47, 49] only proved the convergence of ADMM to a stationary point, and no iteration complexity analysis was provided. Moreover, the objective function is required to satisfy the so-called Kurdyka-Łojasiewicz (KL) property [9, 33, 39, 40] to ensure those convergence results. In [30], Hong, Luo and Razaviyayn analyzed the convergence of ADMM for solving nonconvex consensus and sharing problems. Note that they also analyzed the iteration complexity of ADMM for the consensus problem. However, they require the nonconvex part of the objective function to be smooth, and nonsmooth part to be convex. In contrast, r_i in our model (1.1) can be nonconvex and nonsmooth at the same time. Moreover, we allow general set constraints $x_i \in \mathcal{X}_i, i = 1, \ldots, N-1$, while the consensus problem in [30] only allows the set constraint for one block variable. The very recent work by Hong [29] discusses the iteration complexity of an augmented Lagrangian method for finding an ϵ-stationary point for the following problem:

$$\min\ f(x),\ \text{s.t. } Ax = b, x \in \mathbb{R}^n, \tag{1.8}$$

under the assumption that f is differentiable. We will compare our results with [29] in more details in Section 3.

Throughout this paper, we make the following assumption.

Assumption 1.1 *All subproblems in our algorithms, though possibly nonconvex, can be solved to global optimality.*

We shall show later that the solvability of our subproblems usually boils down to the computability of the proximal mapping with respect to the nonsmooth part of the objective function.

Besides, the proximal mappings of the aforementioned nonsmooth regularization functions, including the ℓ_1 norm, SCAD, LSP, MCP and Capped-ℓ_1 penalty, all admit closed-form solutions, and the explicit formulae can be found in [26].

Before proceeding, let us first summarize:

Our contributions.

(i) We provide a systematic study on how to define an ϵ-stationary point of (1.1) and (1.2). For (1.1), our definition of ϵ-stationary point covers the cases when each r_i is convex, or r_i is Lipschitz continuous (possibly nonconvex), or r_i is lower semi-continuous (possibly nonconvex).

(ii) We propose a generalized conditional gradient method for solving (1.2) and analyze its iteration complexity for obtaining an ϵ-stationary point of (1.2).

(iii) We propose two ADMM variants (proximal ADMM-g and proximal ADMM-m) for solving (1.1), under certain conditions on A_N. We also analyze their iteration complexities for obtaining an ϵ-stationary point of (1.1).

(iv) As an extension, we also show how to use proximal ADMM-g and proximal ADMM-m to find an ϵ-stationary point of (1.1) without assuming any assumption on A_N.

(v) As a by-product, we also propose a proximal block coordinate descent (BCD) method with cyclic order for solving (1.1) when the affine constraints are absent, and show that its iteration complexity can be obtained directly from that of proximal ADMM-g and proximal ADMM-m.

Notation. We use $\|x\|_2$ to denote the Euclidean norm of vector x, and $\|x\|_H^2$ to denote $x^\top Hx$ for some positive definite matrix H. For set S and scalar $p > 1$, we denote

$$\mathrm{diam}_p(S) := \max_{x,y \in S} \|x - y\|_p, \tag{1.9}$$

where $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Without specification, we denote $\|x\| = \|x\|_2$ and $\mathrm{diam}(S) = \mathrm{diam}_2(S)$ for short. We use $\mathrm{dist}(x, S)$ to denote the Euclidean distance of vector x to set S. Given a matrix A, its spectral norm, largest singular value and smallest singular value are denoted by $\|A\|_2$, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ respectively. We use $\lceil a \rceil$ to denote the largest integer that is less than or equal to scalar a.

Organization. The rest of this paper is organized as follows. In Section 2 we give the definition of ϵ-stationary point of (1.2) and propose a generalized conditional gradient method that solves (1.2) and analyze its iteration complexity for obtaining such an ϵ-stationary point of (1.2). In Section 3 we give three definitions of ϵ-stationarity for (1.1) under different settings and propose two ADMM variants that solve (1.1) and analyze their iteration complexities to reach an ϵ-stationary point of (1.1). In Section 4 we provide some extensions of the results in Section 3. In particular, we first show how to remove some of the conditions that we assume in Section 3, and then we propose a proximal BCD method to solve (1.1) without affine constraints and provide an iteration complexity analysis. In Section 5, we provide numerical results to illustrate the practical efficiency of the proposed algorithms.

2 A generalized conditional gradient method

In this section, we propose a GCG method for solving (1.2) and analyze its iteration complexity. The conditional gradient (CG) method, also known as the Frank-Wolfe method, was originally proposed in [22], and regained a lot of popularity recently due to its capability of solving large-scale problems (see, [4, 5, 23, 27, 31, 34, 42]). However, these works focus on solving convex problems. Bredies et al. [14] considered a generalized conditional gradient method for solving nonconvex problems in Hilbert space, which is similar to our algorithm, but no iteration complexity was provided.

Throughout this section, we make the following assumption regarding to problem (1.2).

Assumption 2.1 *In (1.2), function $r(x)$ is convex and nonsmooth, and the constraint set S is convex and compact. Moreover, f is differentiable and there exist some $p > 1$ and $\rho > 0$ such that*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\rho}{2}\|y - x\|_p^p, \quad \forall x, y \in S. \tag{2.1}$$

The inequality (2.1) is the so-called Hölder condition and was also used in other papers that discuss first-order algorithms (e.g., [20]). It can be shown that (2.1) holds for a variety of functions. In fact, we have the following results.

Proposition 2.2

(i) If f is concave, then (2.1) holds for any $p > 0$ and $\rho > 0$.

(ii) If the gradient of f satisfies

$$\|\nabla f(x) - \nabla f(y)\|_q^q \leq M\|x - y\|_p^p, \quad \forall\, x, y \in S, \tag{2.2}$$

for some $M > 0$, $1 < p \leq 2$ and $\frac{1}{p} + \frac{1}{q} = 1$, then (2.1) holds.

(iii) (2.1) holds for the p-norm function

$$f(x) = \sum_{i=1}^{n} x_i^p, \qquad \text{where } 1 < p \leq 2 \text{ and } x_i > 0, i = 1, \ldots, n. \tag{2.3}$$

Proof. Part (i) is obvious. For (ii), let $z = y - x$ and $g(\alpha) = f(x + \alpha z)$, it follows that

$$\begin{aligned} f(y) - f(x) &= \int_0^1 g'(\alpha) z \, d\alpha = \int_0^1 \nabla f(x + \alpha z)^\top z \, d\alpha \\ &\leq \nabla f(x)^\top z + \left| \int_0^1 (\nabla f(x + \alpha z) - \nabla f(x))^\top z \, d\alpha \right| \\ &\leq \nabla f(x)^\top z + \int_0^1 \|\nabla f(x + \alpha z) - \nabla f(x)\|_q \|z\|_p \, d\alpha \\ &\leq \nabla f(x)^\top z + M^{1/q} \|z\|_p^{1 + \frac{p}{q}} \int_0^1 \alpha^{\frac{p}{q}} \, d\alpha \\ &= \nabla f(x)^\top z + \frac{M^{1/q}}{p} \|z\|_p^p, \end{aligned}$$

where the last equality is due to $\frac{1}{p} + \frac{1}{q} = 1$. Thus, the function with Lipschitz continuous gradient automatically satisfies inequality (2.1) for $p = q = 2$. In fact, condition (2.2) reflects the degree of the Hölderian continuity of ∇f, which was also considered in [20] to construct a so-called inexact first-order oracle. For Part (iii), we observe that the function is separable with respect to all x_i, so it suffices to show that there exists some ρ such that:

$$v^p \leq u^p + (u^p)'(v - u) + \frac{\rho}{2}|v - u|^p = u^p + p\, u^{p-1}(v - u) + \frac{\rho}{2}|v - u|^p, \tag{2.4}$$

when $1 < p \leq 2$. If $u = 0$, then the inequality trivially holds for any $\rho \geq 2$; otherwise we can divide both sides by $|u|^p$ and aim to prove an equivalent formulation:

$$k^p \leq 1 + p(k-1) + \frac{\rho}{2}|k-1|^p, \tag{2.5}$$

where $k = v/u$. To this end, define

$$g(k) := \begin{cases} 0 & \text{if } k = 1 \\ \frac{k^p - 1 - p(k-1)}{|k-1|^p} & \text{otherwise.} \end{cases}$$

Observe that $\lim_{k \to +\infty} g(k) = 1$ and by the L'Hospital rule

$$\lim_{k \to 1} g(k) = \begin{cases} 0 & \text{if } 1 < p < 2 \\ 1 & \text{if } p = 2, \end{cases}$$

so $g(k)$ is upper bounded on $\mathbb{R}$ and there exits some $\hat{\rho}$ such that (2.5) holds. Finally by letting $\rho = \max\{2, \hat{\rho}\}$, the inequality (2.4) follows. □

2.1 An ϵ-stationary point for problem (1.2)

Our definition of an ϵ-stationary point of (1.2) is given as follows.

Definition 2.3 *We call x to be an ϵ-stationary point ($\epsilon \geq 0$) of (1.2) if the following mixed variational inequality conditions is satisfied:*

$$\psi_S(x) := \nabla f(x)^\top (y - x) + r(y) - r(x) \geq -\epsilon, \quad \forall y \in S. \tag{2.6}$$

If $\epsilon = 0$, we call x to be a stationary point of (1.2).

When $\epsilon = 0$, the condition (2.6) is stronger than the commonly used KKT condition in the sense that it is a necessary condition for local minimum of (1.2). To see this, suppose that there exists some $y \in S$ such that $\nabla f(x)^\top (y - x) + r(y) - r(x) < 0$. Denote $d = y - x$. Then the directional derivative along direction d at point x satisfies

$$\begin{aligned} (f + r)'(x; d) &= \lim_{\alpha \searrow 0} \frac{f(x + \alpha d) + r(x + \alpha d) - f(x) - r(x)}{\alpha} \\ &\leq \lim_{\alpha \searrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha} + \lim_{\alpha \searrow 0} \frac{(1-\alpha) r(x) + \alpha r(x + d) - r(x)}{\alpha} \\ &= \nabla f(x)^\top (y - x) + r(y) - r(x) < 0, \end{aligned}$$

where the first inequality is due to the convexity of r. As a result, x cannot be a local minimizer of problem (1.2).

We now compare our Definition 2.3 with some existing definitions of ϵ-stationary point in the literature. For the smooth unconstrained problem $\{\min \ f(x)\}$, it is natural to define the ϵ-stationary point using the criterion $\|\nabla f(x)\|_2 \leq \epsilon$. Nesterov [43] and Cartis et al. [17] showed that the gradient descent type methods with properly chosen step size need $O(1/\epsilon^2)$ iterations to find such a point. Moreover, Cartis et al. [16] constructed an example showing that the $O(1/\epsilon^2)$ iteration complexity is tight for the steepest descent type algorithm. However, the case for the constrained nonconvex optimization is more complicated. When $r \equiv 0$ in (1.2), i.e., the objective function is differentiable, Cartis et al. [18] proposed the following measure:

$$\chi_S(x) := \left| \min_{x + d \in S, \|d\|_2 \leq 1} \nabla f(x)^\top d \right| \leq \epsilon. \tag{2.7}$$

They showed that it requires no more than $O(1/\epsilon^2)$ iterations for the adaptive cubic regularization algorithm in [18] to find an x satisfying (2.7). Ghadimi et al. [25] gave the following definition of an ϵ-stationary point of (1.2). Define

$$P_S(x,\gamma) := \frac{1}{\gamma}(x - x^+), \quad \text{where } x^+ = \arg\min_{y\in S} \nabla f(x)^\top y + \frac{1}{\gamma} V(y,x) + r(y), \tag{2.8}$$

where $\gamma > 0$ and V is a prox-function. Ghadimi et al. [25] proposed a projected gradient algorithm to solve (1.2) and proved that it takes no more than $O(1/\epsilon^2)$ iterations to find an x satisfying

$$\|P_S(x,\gamma)\|_2^2 \le \epsilon. \tag{2.9}$$

We have the following proposition regarding the relationships among the three definitions of ϵ-stationary point defined in (2.6), (2.7) and (2.9).

Proposition 2.4 *Consider* (1.2)*.*

(i) When $r(x) \equiv 0$, if $\psi_S(x) \ge -\epsilon$, then $\chi_S(x) \le \epsilon$;

(ii) Suppose that the prox-function $V(y,x) = \|y-x\|_2^2/2$, then $\psi_S(x) \ge -\epsilon$ implies $\|P_S(x,\gamma)\|_2^2 \le \epsilon/\gamma$. Conversely, if we further assume that the gradient function $\nabla f(x)$ is continuous, then $\|P_S(x,\gamma)\|_2^2 \le \epsilon$ implies

$$\psi_S(x) \ge -(\gamma\tau + \gamma\varsigma + \mathrm{diam}(S))\sqrt{\epsilon}, \tag{2.10}$$

where $\tau = \max_{x\in S} \|\nabla f(x)\|_2$, $\varsigma = \max_{x\in S}\min_{z\in\partial r(x)} \|z\|_2$. Note that ς is finite, because S is compact.

Proof. Part (i) follows from (2.6) and (2.7) by the following relationships:

$$\begin{aligned} \psi_S(x) \ge -\epsilon &\Longrightarrow \nabla f(x)^\top (y-x) \ge -\epsilon, \forall\ \|y-x\|_2 \le 1, y\in S \\ &\Longrightarrow 0 \ge \min_{x+d\in S, \|d\|_2\le 1} \nabla f(x)^\top d \ge -\epsilon \Longrightarrow \chi_S(x) \le \epsilon. \end{aligned}$$

We now prove part (ii). Since $V(y,x) = \|y-x\|_2^2/2$, (2.8) implies that

$$\left(\nabla f(x) + \frac{1}{\gamma}(x^+ - x) + z\right)^\top (y - x^+) \ge 0, \quad \forall\ y \in S, \tag{2.11}$$

where $z \in \partial r(x^+)$. By choosing $y = x$ in (2.11) one can get

$$\nabla f(x)^\top (x - x^+) + r(x) - r(x^+) \ge (\nabla f(x) + z)^\top (x - x^+) \ge \frac{1}{\gamma}\|x^+ - x\|_2^2. \tag{2.12}$$

Therefore, if $\psi_S(x) \ge -\epsilon$, then $\|P_S(x,\gamma)\|_2^2 \le \frac{\epsilon}{\gamma}$ holds. To show the other direction, note that for $x, x^+ \in S$, one can choose $w \in \partial r(x)$ such that

$$\varsigma\, \|x - x^+\|_2 \ge w^\top (x - x^+) \ge r(x) - r(x^+),$$

which together with (2.12) implies that

$$\begin{aligned} & \nabla f(x)^\top (y-x) + r(y) - r(x) + (\|\nabla f(x)\|_2 + \varsigma)\|x - x^+\|_2 \\ \ge\ & \nabla f(x)^\top (y-x) + r(y) - r(x) + \nabla f(x)^\top (x - x^+) + r(x) - r(x^+) \\ =\ & \nabla f(x)^\top (y - x^+) + r(y) - r(x^+) \\ \ge\ & (\nabla f(x) + z)^\top (y - x^+) \ge -\frac{1}{\gamma}(x^+ - x)^\top (y - x^+) \ge -\frac{1}{\gamma}\, \|y - x^+\|_2 \|x^+ - x\|_2, \quad \forall y \in S, \end{aligned}$$

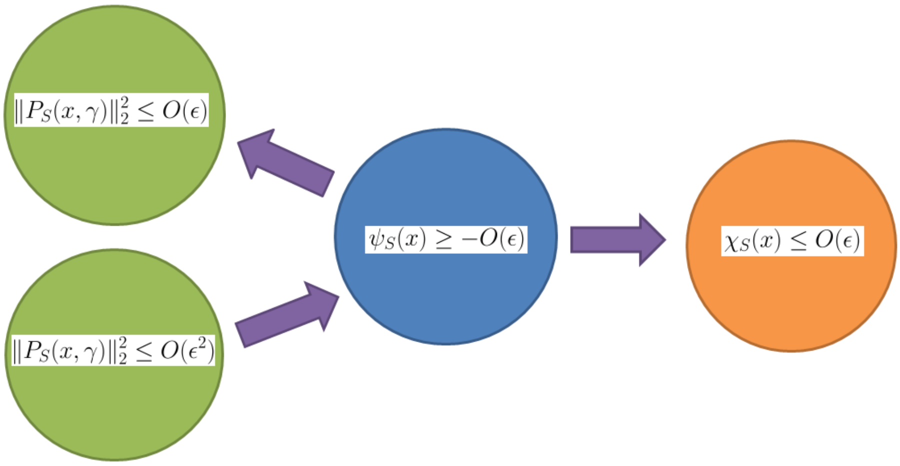

Figure 1: Relationship of the three definitions of ϵ-stationary point of (1.2)

where $z \in \partial r(x^+)$, the second inequality follows from the convexity of $r(x)$ and the third inequality is due to the optimality condition of (2.8). (2.10) follows by rearranging the terms in the above inequality. □

Under the conditions in Proposition 2.4, the relationship of these three definitions of ϵ-stationary point of (1.2) is depicted in Figure 1, which shows that our definition (2.6) is to some extent more general than (2.7) and (2.9).

2.2 GCG method and its iteration complexity

For given point z, we define a linearization of the objective function of (1.2) as:

$$\ell(x; z) := f(z) + \nabla f(z)^\top (x - z) + r(x), \tag{2.13}$$

which is obtained by linearizing the smooth part (function f) of Φ in (1.2). Our GCG method for solving (1.2) is described in Algorithm 1.

Algorithm 1 Generalized Conditional Gradient Algorithm (GCG) for solving (1.2)

Require: Given $x^0 \in S$
for $k = 0, 1, \ldots$ **do**
[Step 1] $y^k = \arg\min_{y \in S} \ell(y; x^k)$, and let $d^k = y^k - x^k$;
[Step 2] $\alpha_k = \arg\min_{\alpha\in[0,1]} \alpha \nabla f(x^k)^\top d^k + \alpha^p \frac{\rho}{2}\|d^k\|_p^p + (1-\alpha) r(x^k) + \alpha r(y^k)$;
[Step 3] Set $x^{k+1} = (1-\alpha_k)x^k + \alpha_k y^k$.
end for

Note that when the nonsmooth function r is absent, GCG differs from the classical CG method only by the choice of the step size α_k.

Remark 2.5 *According to Assumption 1.1, we assume that the problem in Step 1 of Algorithm 1 can be solved to global optimality. See [41] for problems arising from sparse PCA that satisfy this assumption.*

Remark 2.6 *It is easy to see that the sequence $\{\Phi(x^k)\}$ generated by GCG is monotonically nonincreasing, which implies that any cluster point of $\{x^k\}$ cannot be a strict local maximizer.*

Before we proceed to the main result on iteration complexity of GCG, we need the following lemma that gives a sufficient condition for ϵ-stationary point of (1.2). This lemma is inspired by [24], and it indicates that if the progress gained by solving (2.14) is small, then z is close to a stationary point of (1.2).

Lemma 2.7 *Define*

$$z_\ell := \operatorname*{argmin}_{x \in S} \; \ell(x; z). \tag{2.14}$$

The improvement of the linearization at point z is defined as

$$\triangle \ell_z := \ell(z; z) - \ell(z_\ell; z) = -\nabla f(z)^\top (z_\ell - z) + r(z) - r(z_\ell).$$

Given $\epsilon \geq 0$, for any $z \in S$, if $\triangle \ell_z \leq \epsilon$, then z is an ϵ-stationary point of (1.2) as defined in Definition 2.3.

Proof. Since z_ℓ is optimal to (2.14), we have

$$\ell(y; z) - \ell(z_\ell; z) = \nabla f(z)^\top (y - z_\ell) + r(y) - r(z_\ell) \geq 0, \forall y \in S,$$

which implies that

$$\begin{aligned}
& \nabla f(z)^\top (y - z) + r(y) - r(z) \\
= \; & \nabla f(z)^\top (y - z_\ell) + r(y) - r(z_\ell) + \nabla f(z)^\top (z_\ell - z) + r(z_\ell) - r(z) \\
\geq \; & \nabla f(z)^\top (z_\ell - z) + r(z_\ell) - r(z), \forall y \in S.
\end{aligned}$$

It then follows immediately that if $\triangle \ell_z \leq \epsilon$, then $\nabla f(z)^\top (y - z) + r(y) - r(z) \geq -\triangle \ell_z \geq -\epsilon$. □

We are now ready to give the main result of the iteration complexity of GCG (Algorithm 1) for obtaining an ϵ-stationary point of (1.2).

Theorem 2.8 *For any $\epsilon \in (0, \operatorname{diam}_p^p(S)\rho)$, GCG finds an ϵ-stationary point of (1.2) within $\left\lceil \frac{2(\Phi(x^0) - \Phi^*)(\operatorname{diam}_p^p(S)\rho)^{q-1}}{\epsilon^q} \right\rceil$ iterations, where $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof. For ease of presentation, we denote $D := \operatorname{diam}_p(S)$ and $\triangle \ell^k := \triangle \ell_{x^k}$. By Assumption 2.1, using the fact that $\frac{\epsilon}{D^p \rho} < 1$, and by the definition of α_k in Algorithm 1, we have

$$\begin{aligned}
& \left(\frac{\epsilon}{D^p \rho}\right)^{\frac{1}{p-1}} \triangle \ell^k - \frac{1}{2\rho^{1/(p-1)}} \left(\frac{\epsilon}{D}\right)^{\frac{p}{p-1}} \\
\leq \; & -\left(\frac{\epsilon}{D^p \rho}\right)^{\frac{1}{p-1}} \left(\nabla f(x^k)^\top (y^k - x^k) + r(y^k) - r(x^k)\right) - \frac{\rho}{2} \left(\frac{\epsilon}{D^p \rho}\right)^{\frac{p}{p-1}} \|y^k - x^k\|_p^p \\
\leq \; & -\alpha_k \left(\nabla f(x^k)^\top (y^k - x^k) + r(y^k) - r(x^k)\right) - \frac{\rho \alpha_k^p}{2} \|y^k - x^k\|_p^p \\
\leq \; & -\nabla f(x^k)^\top (x^{k+1} - x^k) + r(x^k) - r(x^{k+1}) - \frac{\rho}{2} \|x^{k+1} - x^k\|_p^p \\
\leq \; & f(x^k) - f(x^{k+1}) + r(x^k) - r(x^{k+1}) = \Phi(x^k) - \Phi(x^{k+1}),
\end{aligned} \tag{2.15}$$

where the third inequality is due to the convexity of function r and the fact that $x^{k+1} - x^k = \alpha_k (y^k - x^k)$, and the last inequality is due to (2.1). Furthermore, (2.15) immediately yields

$$\triangle \ell^k \leq \left(\frac{\epsilon}{D^p \rho}\right)^{-\frac{1}{p-1}} \left(\Phi(x^k) - \Phi(x^{k+1})\right) + \frac{\epsilon}{2}. \tag{2.16}$$

For any integer $K > 0$, summing (2.16) over $k = 0, 1, \ldots, K$, yields

$$K \min_{k\in\{0,1,\ldots,K\}} \triangle\ell^k \le \sum_{k=1}^{K} \triangle\ell^k \le \left(\frac{\epsilon}{D^p\rho}\right)^{-\frac{1}{p-1}} \left(\Phi(x^0) - \Phi(x^{K+1})\right) + \frac{\epsilon}{2}K$$
$$\le \left(\frac{\epsilon}{D^p\rho}\right)^{-\frac{1}{p-1}} \left(\Phi(x^0) - \Phi^*\right) + \frac{\epsilon}{2}K,$$

where Φ^* is the optimal value of (1.2). It is easy to see that by setting $K = \left\lceil \frac{2(\Phi(x^0)-\Phi^*)(D^p\rho)^{q-1}}{\epsilon^q} \right\rceil$, the above inequality implies $\triangle\ell_{x^{k^*}} \le \epsilon$, where $k^* := \mathrm{argmin}_{k\in\{1,\ldots,K\}} \triangle\ell^k$. According to Lemma 2.7, x^{k^*} is an ϵ-stationary point of (1.2) as defined in Definition 2.3. □

We have the following immediate corollary when f is a concave function.

Corollary 2.9 *When f is a concave function, if we set $\alpha_k = 1$ for all k in GCG (Algorithm 1), then it returns an ϵ-stationary point of (1.2) within $\left\lceil \frac{\Phi(x^0)-\Phi^*}{\epsilon} \right\rceil$ iterations.*

Proof. By setting $\alpha_k = 1$ in Algorithm 1 we know that $x^{k+1} = y^k$ for all k. Since f is concave, it holds that

$$\triangle\ell^k = -\nabla f(x^k)^\top (x^{k+1} - x^k) + r(x^k) - r(x^{k+1}) \le \Phi(x^k) - \Phi(x^{k+1}).$$

Summing this inequality over $k = 0, 1, \ldots, K$ yields

$$K \min_{k\in\{0,1,\ldots,K\}} \triangle\ell^k \le \Phi(x^0) - \Phi^*,$$

which leads to the desired result immediately. □

3 Variants of ADMM for solving nonconvex problems with affine constraints

In this section, we propose two variants of the ADMM (Alternating Direction Method of Multipliers) for solving the general problem (1.1), and prove their iteration complexities for obtaining an ϵ-stationary point (to be defined later) under certain conditions. Throughout this section, we assume the following two assumptions regarding problem (1.1).

Assumption 3.1 *The partial gradient of the function f with respect to x_N is Lipschitz continuous with Lipschitz constant $L > 0$, i.e., for any $(x_1^1, \cdots, x_N^1)$ and $(x_1^2, \cdots, x_N^2) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1} \times \mathbb{R}^{n_N}$, it holds that*

$$\left\|\nabla_N f(x_1^1, x_2^1, \cdots, x_N^1) - \nabla_N f(x_1^2, x_2^2, \cdots, x_N^2)\right\| \le L \left\|\left(x_1^1 - x_1^2, x_2^1 - x_2^2, \cdots, x_N^1 - x_N^2\right)\right\|, \tag{3.1}$$

which implies that for any $(x_1, \cdots, x_{N-1}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1}$ and $x_N, \hat{x}_N \in \mathbb{R}^{n_N}$, it holds that

$$f(x_1, \cdots, x_{N-1}, x_N) \le f(x_1, \cdots, x_{N-1}, \hat{x}_N) + (x_N - \hat{x}_N)^\top \nabla_N f(x_1, \cdots, x_{N-1}, \hat{x}_N) + \frac{L}{2}\|x_N - \hat{x}_N\|^2. \tag{3.2}$$

Assumption 3.2 *f and $r_i, i = 1, \ldots, N-1$ are all lower bounded, and we denote*

$$f^* = \min_{x_i\in\mathcal{X}_i, i=1,\ldots,N-1; x_N\in\mathbb{R}^{n_N}} \{f(x_1, x_2, \cdots, x_N)\}$$

and $r_i^* = \min_{x_i\in\mathcal{X}_i} \{r_i(x_i)\}$ for $i = 1, 2, \ldots, N-1$.

3.1 Preliminaries

To characterize the optimality conditions of (1.1) when r_i is nonsmooth and nonconvex, we need to recall the definition of generalized gradient (see, e.g., [44]).

Definition 3.3 *Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous function. Suppose $h(\bar{x})$ is finite for a given $\bar{x}$. For $v \in \mathbb{R}^n$, we say that*

(i) v is a regular subgradient (also called Frechét subdifferential) of h at $\bar{x}$, written $v \in \hat{\partial} h(\bar{x})$, if

$$\lim_{x \neq \bar{x}} \inf_{x \to \bar{x}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0;$$

(ii) v is a general subgradient of h at $\bar{x}$, written $v \in \partial h(\bar{x})$, if there exist sequences $\{x^k\}$ and $\{v^k\}$ such that $x^k \to \bar{x}$ with $h(x^k) \to h(\bar{x})$, and $v^k \in \hat{\partial} h(x^k)$ with $v^k \to v$ when $k \to \infty$.

The following proposition lists some well known facts on semi-continuous functions that will be used in our analysis later.

Proposition 3.4 *Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper lower semi-continuous functions. Then it holds that:*

(i) (Theorem 10.1 in [44]) The Fermat's rule remains true: if $\bar{x}$ is a local minimum of h, then $0 \in \partial h(\bar{x})$.

(ii) If h is continuously differentiable function, then $\partial(h + g)(x) = \nabla h(x) + \partial g(x)$.

(iii) (Exercise 10.10 in [44]) If h is locally Lipschitz continuous at $\bar{x}$, then $\partial(h + g)(x) \subset \partial h(x) + \partial g(x)$.

As a direct consequence of Proposition 3.4, we have the following optimality conditions based on variational inequality (VI) for a general constrained optimization problem.

Lemma 3.5 *Suppose $h(x)$ is locally Lipschitz continuous, X is a closed and convex set, and $\bar{x}$ is a local minimum of $\min_{x \in X} h(x)$. Then there exists $v \in \partial h(\bar{x})$ such that $(x - \bar{x})^\top v \geq 0, \forall x \in X$.*

Proof. Note that $\min_{x \in X} h(x)$ is equivalent to $\min_x h(x) + \delta_X(x)$, where $\delta_X(x) = 0$ if $x \in X$ and $\delta_X(x) = 0$ otherwise. According to Proposition 3.4, it holds that

$$0 \in \partial(h + \delta_X)(\bar{x}) \subset \partial h(\bar{x}) + \partial \delta_X(\bar{x}) = \partial h(\bar{x}) + N_X(\bar{x}),$$

where $N_X(\bar{x})$ is the normal cone of X at point $\bar{x}$. That is, there exits $v \in \partial h(\bar{x})$ such that $-v \in N_X(\bar{x})$. By invoking the definition of normal cone of a convex set, we obtain the desired result. □

In our analysis, we frequently use the following identity that holds for any vectors a, b, c, d,

$$(a - b)^\top (c - d) = \frac{1}{2} \left(\|a - d\|_2^2 - \|a - c\|_2^2 + \|b - c\|_2^2 - \|b - d\|_2^2 \right), \tag{3.3}$$

and the following inequality that holds for any vectors a, b, and scalar $\xi > 0$

$$a^\top b \leq \frac{1}{4\xi} \|a\|_2^2 + \xi \|b\|_2^2. \tag{3.4}$$

3.2 An ϵ-stationary point for problem (1.1)

We now introduce notions of ϵ-stationarity for (1.1) in the following three settings: (i) **Setting 1**: r_i is a convex function, and $\mathcal{X}_i$ is a compact set, for $i = 1, \ldots, N-1$; (ii) **Setting 2**: r_i is Lipschitz continuous, and $\mathcal{X}_i$ is a compact set, for $i = 1, \ldots, N-1$; (iii) **Setting 3**: r_i is lower semi-continuous, and $\mathcal{X}_i = \mathbb{R}^{n_i}$, for $i = 1, \ldots, N-1$.

Definition 3.6 (ϵ-stationary point of (1.1) in Setting 1) *Under the conditions in* **Setting 1**, *for $\epsilon \geq 0$, we call $(x_1^*, \cdots, x_N^*, \lambda^*)$ to be an ϵ-stationary point of (1.1) if for any $(x_1, \cdots, x_N, \lambda) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1} \times \mathbb{R}^{n_N} \times \mathbb{R}^m$, it holds that*

$$r_i(x_i) - r_i(x_i^*) + (x_i - x_i^*)^\top \left[\nabla_i f(x_1^*, \cdots, x_N^*) - A_i^\top \lambda^*\right] \geq -\epsilon, \; i = 1, \ldots, N-1, \quad (3.5)$$

$$\left\|\nabla_N f(x_1^*, \ldots, x_{N-1}^*, x_N^*) - A_N^\top \lambda^*\right\| \leq \epsilon, \quad (3.6)$$

$$\left\|\sum_{i=1}^N A_i x_i^* - b\right\| \leq \epsilon. \quad (3.7)$$

If $\epsilon = 0$, we call $(x_1^*, \cdots, x_N^*, \lambda^*)$ to be a stationary point of (1.1).

Definition 3.7 (ϵ-stationary point of (1.1) in Setting 2) *Under the conditions in* **Setting 2**, *for $\epsilon \geq 0$, we call $(x_1^*, \cdots, x_N^*, \lambda^*)$ to be an ϵ-stationary point of (1.1) if for any $(x_1, \cdots, x_N, \lambda) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1} \times \mathbb{R}^{n_N} \times \mathbb{R}^m$, (3.6), (3.7) and the following hold:*

$$(x_i - x_i^*)^\top \left[g_i^* + \nabla_i f(x_1^*, \cdots, x_N^*) - A_i^\top \lambda^*\right] \geq -\epsilon, \; i = 1, \ldots, N-1, \quad (3.8)$$

where $g_i^* \in \partial r_i(\hat{x}_i)$ is a general subgradient of r_i at $\hat{x}_i$ as defined in Definition 3.3, $i = 1, \ldots, N-1$. If $\epsilon = 0$, we call $(x_1^*, \cdots, x_N^*, \lambda^*)$ to be a stationary point of (1.1).

If $\mathcal{X}_i = \mathbb{R}^{n_i}$ for $i = 1, \ldots, N-1$, then the VI kind conditions in Definition 3.7 reduce to the following one.

Definition 3.8 (ϵ-stationary point of (1.1) in Setting 3) *Under the conditions in* **Setting 3**, *for $\epsilon \geq 0$, we call $(x_1^*, \ldots, x_N^*, \lambda^*)$ to be an ϵ-stationary point of (1.1) if (3.6), (3.7) and the following hold:*

$$\mathrm{dist}\left(-\nabla_i f(x_1^*, \cdots, x_N^*) + A_i^\top \lambda^*, \partial r_i(x_i^*)\right) \leq \epsilon, \; i = 1, \ldots, N-1, \quad (3.9)$$

where $\partial r_i(x_i^*)$ is the general subgradient of r_i at x_i^*, $i = 1, 2, \ldots, N-1$. If $\epsilon = 0$, we call $(x_1^*, \cdots, x_N^*, \lambda^*)$ to be a stationary point of (1.1).

The three settings of problem (1.1) considered in this section and their corresponding definitions of ϵ-stationary point, are summarized in Table 1.

Table 1: ϵ-stationary point of (1.1) in three settings

	r_i, $i = 1, \ldots, N-1$	$\mathcal{X}_i$, $i = 1, \ldots, N-1$	ϵ-stationary point
Setting 1	convex	$\mathcal{X}_i \subset \mathbb{R}^{n_i}$ compact	Definition 3.6
Setting 2	Lipschitz continuous	$\mathcal{X}_i \subset \mathbb{R}^{n_i}$ compact	Definition 3.7
Setting 3	lower semi-continuous	$\mathcal{X}_i = \mathbb{R}^{n_i}$	Definition 3.8

A very recent work of Hong [29] proposes a definition of ϵ-stationary point of problem (1.8), and analyzes the iteration complexity of a proximal augmented Lagrangian method for obtaining such a solution. Specifically, (x^*, λ^*) is called an ϵ-stationary point of (1.8) in [29] if $Q(x^*, \lambda^*) \leq \epsilon$, where

$$Q(x, \lambda) := \|\nabla_x \mathcal{L}_\beta(x, \lambda)\|^2 + \|Ax - b\|^2,$$

and $\mathcal{L}_\beta(x, \lambda) := f(x) - \lambda^\top (Ax - b) + \frac{\beta}{2} \|Ax - b\|^2$ is the augmented Lagrangian function of (1.8). Note that [29] assumes that f is differentiable and has bounded gradient in (1.8). The following result reveals that an ϵ-stationary point in [29] is equivalent to an $O(\sqrt{\epsilon})$-stationary point of (1.1) as defined in Definition 3.8 with $r_i = 0$ and f being differentiable. Note that there is no set constraint in (1.8), so the definition of ϵ-stationary point in [29] does not apply to Definitions 3.6 and 3.7.

Proposition 3.9 *Consider the ϵ-stationary point in Definition 3.8 applied to problem (1.8), i.e., one block variable and $r_i(x) = 0$. Then (x^*, λ^*) is a $\gamma_1\sqrt{\epsilon}$-stationary point in Definition 3.8, with $\gamma_1 = 1/(\sqrt{2\beta^2\|A\|_2^2 + 3})$, implies $Q(x^*, \lambda^*) \leq \epsilon$. On the contrary, if $Q(x^*, \lambda^*) \leq \epsilon$, then (x^*, λ^*) is a $\gamma_2\sqrt{\epsilon}$-stationary condition in Definition 3.8, where $\gamma_2 = \sqrt{2(1 + \beta^2\|A\|_2^2)}$.*

Proof. Suppose (x^*, λ^*) is a $\gamma_1\sqrt{\epsilon}$-stationary point as defined in Definition 3.8. Then we have

$$\|\nabla f(x^*) - A^\top\lambda^*\| \leq \gamma_1\sqrt{\epsilon} \quad \text{and} \quad \|Ax^* - b\| \leq \gamma_1\sqrt{\epsilon},$$

which implies that

$$\begin{aligned} Q(x^*, \lambda^*) &= \|\nabla f(x^*) - A^\top\lambda^* + \beta A^\top(Ax^* - b)\|^2 + \|Ax^* - b\|^2 \\ &\leq 2\|\nabla f(x^*) - A^\top\lambda^*\|^2 + 2\beta^2\|A\|_2^2\|Ax^* - b\|^2 + \|Ax^* - b\|^2 \\ &\leq 2\gamma_1^2\epsilon + (2\beta^2\|A\|_2^2 + 1)\gamma_1^2\epsilon = \epsilon. \end{aligned}$$

On the other hand, if $Q(x^*, \lambda^*) \leq \epsilon$, then we have $\|\nabla f(x^*) - A^\top\lambda^* + \beta A^\top(Ax^* - b)\|^2 \leq \epsilon$ and $\|Ax^* - b\|^2 \leq \epsilon$. Therefore,

$$\begin{aligned} \|\nabla f(x^*) - A^\top\lambda^*\|^2 &\leq 2\|\nabla f(x^*) - A^\top\lambda^* + \beta A^\top(Ax^* - b)\|^2 + 2\| - \beta A^\top(Ax^* - b)\|^2 \\ &\leq 2\|\nabla f(x^*) - A^\top\lambda^* + \beta A^\top(Ax^* - b)\|^2 + 2\beta^2\|A\|_2^2\|Ax^* - b\|^2 \\ &\leq 2(1 + \beta^2\|A\|_2^2)\,\epsilon. \end{aligned}$$

The desired result then follows immediately. □

In the following, we introduce two variants of ADMM, named proximal ADMM-g and proximal ADMM-m, that solve (1.1) with some further assumptions on A_N. In particular, proximal ADMM-g assumes $A_N = I$, and proximal ADMM-m assume A_N is full row rank.

3.3 Proximal gradient-based ADMM (proximal ADMM-g)

Our proximal ADMM-g solves (1.1) under the condition that $A_N = I$. Note that when $A_N = I$, the problem is usually referred to as the sharing problem in the literature, and it has a variety of applications (see, e.g., [12, 30, 36, 37]). Our proximal ADMM-g for solving (1.1) with $A_N = I$ is described in Algorithm 2. It can be seen from Algorithm 2 that proximal ADMM-g is based on the framework of augmented Lagrangian method, and can be viewed as a variant of the ADMM. The augmented Lagrangian function of (1.1) is defined as

$$\mathcal{L}_\beta(x_1, \cdots, x_N, \lambda) := f(x_1, \cdots, x_N) + \sum_{i=1}^{N-1} r_i(x_i) - \left\langle \lambda, \sum_{i=1}^{N} A_i x_i - b \right\rangle + \frac{\beta}{2}\left\| \sum_{i=1}^{N} A_i x_i - b \right\|_2^2,$$

where λ is the Lagrange multiplier associated with the affine constraint, and $\beta > 0$ is a penalty parameter. In each iteration, proximal ADMM-g minimizes the augmented Lagrangian function plus a proximal term for block variables $x_1, \ldots, x_{N-1}$, with other variables being fixed; and then a gradient descent step was conducted for x_N, and finally the Lagrange multiplier λ is updated.

Algorithm 2 Proximal Gradient-based ADMM (proximal ADMM-g) for solving (1.1) with $A_N = I$

Require: Given $(x_1^0, x_2^0, \cdots, x_N^0) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1} \times \mathbb{R}^{n_N}$, $\lambda^0 \in \mathbb{R}^m$

for $k = 0, 1, \ldots$ **do**

[Step 1] $x_i^{k+1} := \text{argmin}_{x_i \in \mathcal{X}_i} \ \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_N^k, \lambda^k) + \frac{1}{2}\left\|x_i - x_i^k\right\|_{H_i}^2$ for some positive definite matrix H_i, $i = 1, \ldots, N-1$

[Step 2] $x_N^{k+1} := x_N^k - \gamma \nabla_N \mathcal{L}_\beta(x_1^{k+1}, x_2^{k+1}, \cdots, x_N^k, \lambda^k)$

[Step 3] $\lambda^{k+1} := \lambda^k - \beta\left(\sum_{i=1}^N A_i x_i^{k+1} - b\right)$

end for

Remark 3.10 *According to Assumption 1.1, we assume that the subproblems in Step 1 of Algorithm 2 can be solved to global optimality. In fact, this can be achieved by choosing an appropriate H_i such that the associated objective function is strongly convex. In addition, when the coupled objective is absent or can be linearized, after choosing some proper matrix H_i, the solution of the corresponding subproblem is given by the proximal mappings of r_i. As we mentioned earlier, many nonconvex regularization functions such as SCAD, LSP, MCP and Capped-ℓ_1 adopt closed-form proximal mappings.*

Before we present the main result on the iteration complexity of proximal ADMM-g, we need some lemmas.

Lemma 3.11 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ is generated by Algorithm 2. The following inequality holds*

$$\left\|\lambda^{k+1} - \lambda^k\right\|^2 \le 3\left[\beta - \frac{1}{\gamma}\right]^2 \left\|x_N^k - x_N^{k+1}\right\|^2 + 3\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^{k-1} - x_N^k\right\|^2 + 3L^2 \sum_{i=1}^{N-1}\left\|x_i^{k+1} - x_i^k\right\|^2. \tag{3.10}$$

Proof. Note that Steps 2 and 3 of Algorithm 2 yield that

$$\lambda^{k+1} = \left(\beta - \frac{1}{\gamma}\right)(x_N^k - x_N^{k+1}) + \nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k). \tag{3.11}$$

Combining (3.11) and (3.1) yields that

$$\begin{aligned}
& \left\|\lambda^{k+1} - \lambda^k\right\|^2 \\
\le\ & \left\|\left(\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^k, \cdots, x_{N-1}^k, x_N^{k-1})\right) + \left[\beta - \frac{1}{\gamma}\right]\left(x_N^k - x_N^{k+1}\right)\right. \\
& \left. - \left[\beta - \frac{1}{\gamma}\right]\left(x_N^{k-1} - x_N^k\right)\right\|^2 \\
\le\ & 3\left\|\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^k, \cdots, x_{N-1}^k, x_N^{k-1})\right\|^2 + 3\left[\beta - \frac{1}{\gamma}\right]^2\left\|x_N^k - x_N^{k+1}\right\|^2 \\
& + 3\left[\beta - \frac{1}{\gamma}\right]^2\left\|x_N^{k-1} - x_N^k\right\|^2 \\
\le\ & 3\left[\beta - \frac{1}{\gamma}\right]^2\left\|x_N^k - x_N^{k+1}\right\|^2 + 3\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^{k-1} - x_N^k\right\|^2 + 3L^2\sum_{i=1}^{N-1}\left\|x_i^{k+1} - x_i^k\right\|^2.
\end{aligned}$$

□

We now define the following function, which will play a crucial role in our analysis:

$$\Psi_G\left(x_1, x_2, \cdots, x_N, \lambda, \bar{x}\right) = \mathcal{L}_\beta(x_1, x_2, \cdots, x_N, \lambda) + \frac{3}{\beta}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\|x_N - \bar{x}\|^2. \quad (3.12)$$

Lemma 3.12 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ is generated by Algorithm 2. In Algorithm 2, assume*

$$\beta > \max\left(\frac{18\sqrt{3}+6}{13}L, \max_{i=1,2,\ldots,N-1}\frac{6L^2}{\sigma_{\min}(H_i)}\right). \quad (3.13)$$

Then $\Psi_G(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k)$ monotonically decreases over $k \geq 0$, where $\gamma > 0$ is set as

$$\gamma \in \begin{cases} \left(\frac{\sqrt{13\beta^2-12\beta L-72L^2}-2\beta}{9\beta^2-12\beta L-72L^2}, +\infty\right), & \text{if } \beta \in \left(\frac{2+2\sqrt{19}}{3}L, +\infty\right) \\ \left(\frac{2\beta-\sqrt{13\beta^2-12\beta L-72L^2}}{72L^2+12\beta L-9\beta^2}, \frac{2\beta+\sqrt{13\beta^2-12\beta L-72L^2}}{72L^2+12\beta L-9\beta^2}\right), & \text{if } \beta \in \left(\frac{18\sqrt{3}+6}{13}L, \frac{2+2\sqrt{19}}{3}L\right]. \end{cases} \quad (3.14)$$

Proof. From Step 1 of Algorithm 2 it is easy to obtain that

$$\mathcal{L}_\beta\left(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k, \lambda^k\right) \leq \mathcal{L}_\beta\left(x_1^k, \cdots, x_N^k, \lambda^k\right) - \sum_{i=1}^{N-1}\frac{1}{2}\left\|x_i^k - x_i^{k+1}\right\|_{H_i}^2. \quad (3.15)$$

From Step 2 of Algorithm 2 we get that

$$\begin{aligned} 0 &= \left(x_N^k - x_N^{k+1}\right)^\top\left[\nabla f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \lambda^k + \beta\left(\sum_{i=1}^{N-1}A_i x_i^{k+1} + x_N^k - b\right) - \frac{1}{\gamma}\left(x_N^k - x_N^{k+1}\right)\right] \\ &\leq f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - f(x_1^{k+1}, \cdots, x_N^{k+1}) + \frac{L}{2}\left\|x_N^k - x_N^{k+1}\right\|^2 - \left(x_N^k - x_N^{k+1}\right)^\top\lambda^k \\ &\quad + \frac{\beta}{2}\left\|x_N^k - x_N^{k+1}\right\|^2 + \frac{\beta}{2}\left\|\sum_{i=1}^{N-1}A_i x_i^{k+1} + x_N^k - b\right\|^2 - \frac{\beta}{2}\left\|\sum_{i=1}^{N-1}A_i x_i^{k+1} + x_N^{k+1} - b\right\|^2 - \frac{1}{\gamma}\left\|x_N^k - x_N^{k+1}\right\|^2 \\ &= \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k, \lambda^k) - \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^k) + \left(\frac{L+\beta}{2} - \frac{1}{\gamma}\right)\left\|x_N^k - x_N^{k+1}\right\|^2, \end{aligned} \quad (3.16)$$

where the inequality follows from (3.2) and (3.3). Moreover, the following equality holds trivially

$$\mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) = \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^k) + \frac{1}{\beta}\left\|\lambda^k - \lambda^{k+1}\right\|^2. \quad (3.17)$$

Combining (3.15), (3.16), (3.17) and (3.10) yields that

$$\begin{aligned} &\mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) - \mathcal{L}_\beta(x_1^k, \cdots, x_N^k, \lambda^k) \\ \leq\ &\left(\frac{L+\beta}{2} - \frac{1}{\gamma}\right)\left\|x_N^k - x_N^{k+1}\right\|^2 - \sum_{i=1}^{N-1}\frac{1}{2}\left\|x_i^k - x_i^{k+1}\right\|_{H_i}^2 + \frac{1}{\beta}\left\|\lambda^k - \lambda^{k+1}\right\|^2 \\ \leq\ &\left(\frac{L+\beta}{2} - \frac{1}{\gamma} + \frac{3}{\beta}\left[\beta - \frac{1}{\gamma}\right]^2\right)\left\|x_N^k - x_N^{k+1}\right\|^2 + \frac{3}{\beta}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^{k-1} - x_N^k\right\|^2 \\ &+ \sum_{i=1}^{N-1}\left(x_i^k - x_i^{k+1}\right)^\top\left(\frac{3L^2}{\beta}I - \frac{1}{2}H_i\right)\left(x_i^k - x_i^{k+1}\right), \end{aligned}$$

which further implies that

$$\begin{aligned} &\Psi_G(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k) - \Psi_G(x_1^k, \cdots, x_N^k, \lambda^k, x_N^{k-1}) \\ \leq\ &\left(\frac{L+\beta}{2} - \frac{1}{\gamma} + \frac{6}{\beta}\left[\beta - \frac{1}{\gamma}\right]^2 + \frac{3L^2}{\beta}\right)\left\|x_N^k - x_N^{k+1}\right\|^2 - \sum_{i=1}^{N-1}\left\|x_i^k - x_i^{k+1}\right\|_{\frac{1}{2}H_i - \frac{3L^2}{\beta}I}^2. \end{aligned} \quad (3.18)$$

It is easy to verify that when $\beta > \frac{18\sqrt{3}+6}{13}L$, then γ defined as in (3.14) ensures that $\gamma > 0$ and

$$\frac{L+\beta}{2} - \frac{1}{\gamma} + \frac{6}{\beta}\left[\beta - \frac{1}{\gamma}\right]^2 + \frac{3L^2}{\beta} < 0. \tag{3.19}$$

Therefore, choosing $\beta > \max\left(\frac{18\sqrt{3}+6}{13}L, \max_{i=1,2,\ldots,N-1} \frac{6L^2}{\sigma_{\min}(H_i)}\right)$ and γ as in (3.14) guarantees that $\Psi_G(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k)$ monotonically decreases over $k \geq 0$. In fact, (3.19) can be verified as follows. By denoting $z = \beta - \frac{1}{\gamma}$, (3.19) is equivalent to

$$12z^2 + 2\beta z + \left(6L^2 + \beta L - \beta^2\right) < 0,$$

which holds when $\beta > \frac{18\sqrt{3}+6}{13}L$ and

$$-\beta - \sqrt{13\beta^2 - 12\beta L - 72L^2} < z < -\beta + \sqrt{13\beta^2 - 12\beta L - 72L^2},$$

i.e.,

$$-2\beta - \sqrt{13\beta^2 - 12\beta L - 72L^2} < -\frac{1}{\gamma} < -2\beta + \sqrt{13\beta^2 - 12\beta L - 72L^2},$$

which holds when γ is chosen as in (3.14). □

Lemma 3.13 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ is generated by Algorithm 2. Under the same conditions as in Lemma 3.12, for any $k \geq 0$, we have*

$$\Psi_G\left(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k\right) \geq \sum_{i=1}^{N-1} r_i^* + f^*,$$

where r_i^* and f^* are defined in Assumption 3.2.

Proof. Note that from (3.11), we have

$$\begin{aligned}
& \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) \\
= \; & \sum_{i=1}^{N-1} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_N^{k+1}) - \left(\sum_{i=1}^{N-1} A_i x_i^{k+1} + x_N^{k+1} - b\right)^\top \nabla_N f(x_1^{k+1}, \cdots, x_N^{k+1}) \\
& + \frac{\beta}{2}\left\|\sum_{i=1}^{N-1} A_i x_i^{k+1} + x_N^{k+1} - b\right\|^2 - \left(\sum_{i=1}^{N-1} A_i x_i^{k+1} + x_N^{k+1} - b\right)^\top \left[\left(\beta - \frac{1}{\gamma}\right)\left(x_N^k - x_N^{k+1}\right)\right. \\
& \left. + \left(\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^{k+1}, \cdots, x_N^{k+1})\right)\right] \\
\geq \; & \sum_{i=1}^{N-1} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, b - \sum_{i=1}^{N-1} A_i x_i^{k+1}) + \left(\frac{\beta}{2} - \frac{\beta}{6} - \frac{L}{2}\right)\left\|\sum_{i=1}^{N-1} A_i x_i^{k+1} + x_N^{k+1} - b\right\|^2 \\
& - \frac{3}{\beta}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^k - x_N^{k+1}\right\|^2 \\
\geq \; & \sum_{i=1}^{N-1} r_i^* + f^* - \frac{3}{\beta}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^k - x_N^{k+1}\right\|^2,
\end{aligned}$$

where the first inequality follows from (3.2) and (3.4) with $\xi = 3/(2\beta)$, and the second inequality is due to $\beta \geq 3L/2$. The desired result follows from the definition of Ψ_G in (3.12). □

Now we are ready to give the iteration complexity of Algorithm 2 for finding an ϵ-stationary point of (1.1).

Theorem 3.14 *Suppose the sequence* $\{(x_1^k,\cdots,x_N^k,\lambda^k)\}$ *is generated by Algorithm 2. Assume* β *satisfies* (3.13) *and* γ *satisfies* (3.14). *Denote* $\kappa_1 := \frac{3}{\beta^2}\left[\left(\beta-\frac{1}{\gamma}\right)^2+L^2\right]$, $\kappa_2 := \left(|\beta-\frac{1}{\gamma}|+L\right)^2$, $\kappa_3 := \left(L+\beta\sqrt{N}\max_{1\le i\le N}[\|A_i\|_2^2]+\max_{1\le i\le N}\|H_i\|_2\right)^2$, $\kappa_4 := \max_{1\le i\le N-1}(\mathrm{diam}(\mathcal{X}_i))^2$ *and*

$$\tau := \min\left\{-\left(\frac{L+\beta}{2}-\frac{1}{\gamma}+\frac{6}{\beta}\left[\beta-\frac{1}{\gamma}\right]^2+\frac{3L^2}{\beta}\right), \min_{i=1,\ldots,N-1}\left\{-\left(\frac{3L^2}{\beta}-\frac{\sigma_{\min}(H_i)}{2}\right)\right\}\right\} > 0. \tag{3.20}$$

Letting

$$K := \begin{cases} \left\lceil \frac{2\max\{\kappa_1,\kappa_2,\kappa_3\cdot\kappa_4\}}{\tau\,\epsilon^2}\left(\Psi_G(x_1^1,\cdots,x_N^1,\lambda^1,x_N^0)-\sum_{i=1}^{N-1}r_i^*-f^*\right)\right\rceil, & \textit{for } \textbf{Settings 1, 2} \\ \left\lceil \frac{2\max\{\kappa_1,\kappa_2,\kappa_3\}}{\tau\,\epsilon^2}\left(\Psi_G(x_1^1,\cdots,x_N^1,\lambda^1,x_N^0)-\sum_{i=1}^{N-1}r_i^*-f^*\right)\right\rceil, & \textit{for } \textbf{Setting 3} \end{cases} \tag{3.21}$$

and denoting $\hat{k} = \operatorname{argmin}_{2\le k\le K+1}\sum_{i=1}^N\left(\|x_i^k-x_i^{k+1}\|^2+\|x_i^{k-1}-x_i^k\|^2\right)$, *then* $(x_1^{\hat{k}},\cdots,x_N^{\hat{k}},\lambda^{\hat{k}})$ *is an* ϵ*-stationary point of optimization problem* (1.1) *with* $A_N = I$, *according to Definitions 3.6, 3.7 and 3.8 under the conditions of Settings 1,2 and 3 in Table 1, respectively.*

Proof. For ease of presentation, denote

$$\theta_k := \sum_{i=1}^N(\|x_i^k-x_i^{k+1}\|^2+\|x_i^{k-1}-x_i^k\|^2). \tag{3.22}$$

By summing (3.18) over $k=1,\ldots,K$, we obtain that

$$\Psi_G(x_1^{K+1},\cdots,x_N^{K+1},\lambda^{K+1},x_N^K)-\Psi_G(x_1^1,\cdots,x_N^1,\lambda^1,x_N^0) \le -\tau\sum_{k=1}^K\sum_{i=1}^N\left\|x_i^k-x_i^{k+1}\right\|^2, \tag{3.23}$$

where τ is defined in (3.20). By invoking Lemmas 3.12 and 3.13, we get

$$\begin{aligned}\min_{2\le k\le K+1}\theta_k &\le \frac{1}{\tau K}\left[\Psi_G(x_1^1,\cdots,x_N^1,\lambda^1,x_N^0)+\Psi_G(x_1^2,\cdots,x_N^2,\lambda^2,x_N^1)-2\sum_{i=1}^N r_i^*-2f^*\right] \\ &\le \frac{2}{\tau K}\left[\Psi_G(x_1^1,\cdots,x_N^1,\lambda^1,x_N^0)-\sum_{i=1}^N r_i^*-f^*\right].\end{aligned}$$

We now give upper bounds to the terms in (3.6) and (3.7) through θ_k. Note that (3.11) implies that

$$\begin{aligned}&\|\lambda^{k+1}-\nabla_N f(x_1^{k+1},\cdots,x_N^{k+1})\| \\ \le\ & |\beta-\frac{1}{\gamma}|\,\|x_N^k-x_N^{k+1}\|+\|\nabla_N f(x_1^{k+1},\cdots,x_{N-1}^{k+1},x_N^k)-\nabla f(x_1^{k+1},\cdots,x_N^{k+1})\| \\ \le\ & \left[|\beta-\frac{1}{\gamma}|+L\right]\|x_N^k-x_N^{k+1}\|,\end{aligned}$$

which yields

$$\|\lambda^{k+1}-\nabla_N f(x_1^{k+1},\cdots,x_N^{k+1})\|^2 \le \left[|\beta-\frac{1}{\gamma}|+L\right]^2\theta_k. \tag{3.24}$$

From Step 3 of Algorithm 2 and (3.10) it is easy to see that

$$\begin{aligned}
&\left\|\sum_{i=1}^{N-1} A_i x_i^{k+1} + x_N^{k+1} - b\right\|^2 = \frac{1}{\beta^2}\|\lambda^{k+1} - \lambda^k\|^2 \\
\leq\ & \frac{3}{\beta^2}\left[\beta - \frac{1}{\gamma}\right]^2 \left\|x_N^k - x_N^{k+1}\right\|^2 + \frac{3}{\beta^2}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\left\|x_N^{k-1} - x_N^k\right\|^2 + \frac{3L^2}{\beta^2}\sum_{i=1}^{N-1}\left\|x_i^k - x_i^{k+1}\right\|^2 \\
\leq\ & \frac{3}{\beta^2}\left[\left(\beta - \frac{1}{\gamma}\right)^2 + L^2\right]\theta_k. \qquad (3.25)
\end{aligned}$$

We now give upper bounds to the terms in (3.5), (3.8) and (3.9) under the three settings in Table 1, respectively.

Setting 3. Because r_i is lower semi-continuous and $\mathcal{X}_i = \mathbb{R}^{n_i}$, $i = 1, \ldots, N-1$, it follows from Step 1 of Algorithm 2 that there exists a general subgradient $g_i \in \partial r_i(x_i^{k+1})$ such that

$$\begin{aligned}
& \operatorname{dist}\left(-\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) + A_i^\top \lambda^{k+1}, \partial r_i(x_i^{k+1})\right) \qquad (3.26) \\
\leq\ & \left\|g_i + \nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top \lambda^{k+1}\right\| \\
=\ & \Bigg\|\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - \nabla_i f(x_1^{k+1}, \cdots, x_i^{k+1}, x_{i+1}^k, \cdots, x_N^k) \\
& + \beta A_i^\top\left(\sum_{j=i+1}^{N} A_j(x_j^{k+1} - x_j^k)\right) - H_i(x_i^{k+1} - x_i^k)\Bigg\| \\
\leq\ & L\sqrt{\sum_{j=i+1}^{N}\|x_j^k - x_j^{k+1}\|^2} + \beta\,\|A_i\|_2 \sum_{j=i+1}^{N}\|A_j\|_2\|x_j^{k+1} - x_j^k\| + \|H_i\|_2\|x_i^{k+1} - x_i^k\|_2 \\
\leq\ & \left(L + \beta\sqrt{N}\max_{i+1\leq j\leq N}[\|A_j\|_2]\,\|A_i\|_2\right)\sqrt{\sum_{j=i+1}^{N}\|x_j^k - x_j^{k+1}\|^2} + \|H_i\|_2\|x_i^{k+1} - x_i^k\|_2 \\
\leq\ & \left(L + \beta\sqrt{N}\max_{1\leq i\leq N}[\|A_i\|_2^2] + \max_{1\leq i\leq N}\|H_i\|_2\right)\sqrt{\theta_k}.
\end{aligned}$$

By combining (3.26), (3.24) and (3.25) we conclude that Algorithm 2 returns an ϵ-stationary point of (1.1) according to Definition 3.8 under conditions in Setting 3 in Table 1.

Setting 2. Under this setting, we know r_i is Lipschitz continuous and $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ is convex and compact. Because $f(x_1, \cdots, x_N)$ is differentiable, we know $r_i(x_i) + f(x_1, \cdots, x_N)$ is locally Lipschitz continuous with respect to x_i for $i = 1, 2, \ldots, N-1$. Similar to (3.26), for any $x_i \in \mathcal{X}_i$, Step 1 of Algorithm 2 yields that

$$\begin{aligned}
& \left(x_i - x_i^{k+1}\right)^\top\left[g_i + \nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top \lambda^{k+1}\right] \qquad (3.27) \\
\geq\ & \left(x_i - x_i^{k+1}\right)^\top\Bigg[\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - \nabla_i f(x_1^{k+1}, \cdots, x_i^{k+1}, x_{i+1}^k, \cdots, x_N^k) \\
& + \beta A_i^\top\left(\sum_{j=i+1}^{N} A_j(x_j^{k+1} - x_j^k)\right) - H_i(x_i^{k+1} - x_i^k)\Bigg] \\
\geq\ & -L\operatorname{diam}(\mathcal{X}_i)\sqrt{\sum_{j=i+1}^{N}\|x_j^k - x_j^{k+1}\|^2} - \beta\|A_i\|_2\operatorname{diam}(\mathcal{X}_i)\sum_{j=i+1}^{N}\|A_j\|_2\|x_j^{k+1} - x_j^k\| \\
& -\operatorname{diam}(\mathcal{X}_i)\,\|H_i\|_2\|x_i^{k+1} - x_i^k\|_2 \\
\geq\ & -\left(\beta\sqrt{N}\max_{1\leq i\leq N}[\|A_i\|_2^2] + L + \max_{1\leq i\leq N}\|H_i\|_2\right)\max_{1\leq i\leq N-1}[\operatorname{diam}(\mathcal{X}_i)]\sqrt{\theta_k},
\end{aligned}$$

where $g_i \in \partial r_i(x_i^{k+1})$ is a general subgradient of r_i at x_i^{k+1}. By combining (3.27), (3.24) and (3.25) we conclude that Algorithm 2 returns an ϵ-stationary point of (1.1) according to Definition 3.7 under conditions in Setting 2 in Table 1.

Setting 1. Under this setting, r_i is convex, so g_i in (3.27) becomes a subgradient of r_i at x_i^{k+1}. Therefore, for $i = 1, 2, \cdots, N-1$ and any $x_i \in \mathcal{X}_i$ we have that

$$
\begin{aligned}
& r_i(x_i) - r_i(x_i^{k+1}) + \left(x_i - x_i^{k+1}\right)^\top \left[\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top \lambda^{k+1}\right] \\
\geq\ & \left(x_i - x_i^{k+1}\right)^\top \left[g_i + \nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top \lambda^{k+1}\right] \\
\geq\ & -\left(\beta\sqrt{N} \max_{1\leq i\leq N} \left[\|A_i\|_2^2\right] + L + \max_{1\leq i\leq N} \|H_i\|_2\right) \max_{1\leq i\leq N-1} \left[\mathrm{diam}(\mathcal{X}_i)\right] \sqrt{\theta_k}.
\end{aligned}
\tag{3.28}
$$

By combining (3.28), (3.24) and (3.25) we conclude that Algorithm 2 returns an ϵ-stationary point of (1.1) according to Definition 3.6 under the conditions of Setting 1 in Table 1. □

Remark 3.15 *Note that the potential function Ψ_G defined in (3.12) is related to the augmented Lagrangian function. The augmented Lagrangian function has been used as a potential function in analyzing the convergence of nonconvex splitting and ADMM methods in [2, 28–30, 35]. See [29] for a more detailed discussion on this.*

Remark 3.16 *In Step 1 of Algorithm 2, we can also replace the function*

$$f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_N^k)$$

by its linearization

$$f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \cdots, x_N^k) + \left(x_i - x_i^k\right)^\top \nabla_i f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \cdots, x_N),$$

so that the subproblem can be solved by computing the proximal mappings of r_i, with some properly chosen matrix H_i. Under the additional condition that

$$\left\|\nabla_i f(x_1^1, x_2^1, \cdots, x_N^1) - \nabla_i f(x_1^2, x_2^2, \cdots, x_N^2)\right\| \leq L_i \left\|\left(x_1^1 - x_1^2, x_2^1 - x_2^2, \cdots, x_N^1 - x_N^2\right)\right\|.$$

and setting $H_i \succ L_i I$ for $i = 1, \ldots, N-1$, the same iteration bound still holds by slightly modifying the analysis above.

3.4 Proximal majorization ADMM (proximal ADMM-m)

Our proximal ADMM-m solves (1.1) under the condition that A_N is full row rank. In this section, we use σ_N to denote the smallest eigenvalue of $A_N A_N^\top$. Note that $\sigma_N > 0$ because A_N is full row rank. Our proximal ADMM-m can be described as in Algorithm 3, where $U(x_1, \cdots, x_{N-1}, x_N, \lambda, \bar{x})$ is defined as

$$
\begin{aligned}
U(x_1, \cdots, x_{N-1}, x_N, \lambda, \bar{x}) \quad = \quad & f(x_1, \cdots, x_{N-1}, \bar{x}) + (x_N - \bar{x})^\top \nabla_N f(x_1, \cdots, x_{N-1}, \bar{x}) \\
& + \frac{L}{2}\|x_N - \bar{x}\|^2 - \left\langle \lambda, \sum_{i=1}^N A_i x_i - b \right\rangle + \frac{\beta}{2}\left\|\sum_{i=1}^N A_i x_i - b\right\|^2.
\end{aligned}
$$

It is worth noting the proximal ADMM-m and proximal ADMM-g differ only in Step 2: Step 2 of proximal ADMM-g takes a gradient step of the augmented Lagrangian function with respect to x_N, while Step 2 of proximal ADMM-m requires to minimize a quadratic function of x_N.

We provide some lemmas that are useful in analyzing the iteration complexity of proximal ADMM-m for solving (1.1).

Algorithm 3 Proximal majorization ADMM (proximal ADMM-m) for solving (1.1) with A_N being full row rank

Require: Given $\left(x_1^0, x_2^0, \cdots, x_N^0\right) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N-1} \times \mathbb{R}^{n_N}$, $\lambda^0 \in \mathbb{R}^m$

for $k = 0, 1, \ldots$ **do**

[Step 1] $x_i^{k+1} := \mathrm{argmin}_{x_i \in \mathcal{X}_i}\ \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_N^k, \lambda^k) + \frac{1}{2}\left\|x_i - x_i^k\right\|_{H_i}^2$ for some positive definite matrix H_i, $i = 1, \ldots, N-1$

[Step 2] $x_N^{k+1} := \mathrm{argmin}_{x_N}\ U(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N, \lambda^k, x_N^k)$

[Step 3] $\lambda^{k+1} := \lambda^k - \beta\left(\sum_{i=1}^N A_i x_i^{k+1} - b\right)$

end for

Lemma 3.17 *Suppose the sequence* $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ *is generated by Algorithm 3. The following inequality holds*

$$\left\|\lambda^{k+1} - \lambda^k\right\|^2 \le \frac{3L^2}{\sigma_N}\left\|x_N^k - x_N^{k+1}\right\|^2 + \frac{6L^2}{\sigma_N}\left\|x_N^{k-1} - x_N^k\right\|^2 + \frac{3L^2}{\sigma_N}\sum_{i=1}^{N-1}\left\|x_i^k - x_i^{k+1}\right\|^2. \tag{3.29}$$

Proof. From the optimality conditions of Step 2 of Algorithm 3, we have

$$\begin{aligned} 0 &= \nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - A_N^\top \lambda^k + \beta A_N^\top\left(\sum_{i=1}^N A_i x_i^{k+1} - b\right) - L\left(x_N^k - x_N^{k+1}\right) \\ &= \nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - A_N^\top \lambda^{k+1} - L\left(x_N^k - x_N^{k+1}\right), \end{aligned}$$

where the second equality is due to Step 3 of Algorithm 3. Therefore, we have

$$\begin{aligned} & \|\lambda^{k+1} - \lambda^k\|^2 \le \sigma_N^{-1}\|A_N^\top \lambda^{k+1} - A_N^\top \lambda^k\|^2 \\ \le\ & \sigma_N^{-1}\|(\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^k, \cdots, x_{N-1}^k, x_N^{k-1})) - L(x_N^k - x_N^{k+1}) + L(x_N^{k-1} - x_N^k)\|^2 \\ \le\ & \frac{3}{\sigma_N}\|\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^k, \cdots, x_{N-1}^k, x_N^{k-1})\|^2 + \frac{3L^2}{\sigma_N}(\|x_N^k - x_N^{k+1}\|^2 + \|x_N^{k-1} - x_N^k\|^2) \\ \le\ & \frac{3L^2}{\sigma_N}\|x_N^k - x_N^{k+1}\|^2 + \frac{6L^2}{\sigma_N}\|x_N^{k-1} - x_N^k\|^2 + \frac{3L^2}{\sigma_N}\sum_{i=1}^{N-1}\|x_i^k - x_i^{k+1}\|^2. \end{aligned}$$

□

We define the following function that will be used in the analysis of proximal ADMM-m:

$$\Psi_L\left(x_1, \cdots, x_N, \lambda, \bar{x}\right) = \mathcal{L}_\beta(x_1, \cdots, x_N, \lambda) + \frac{6L^2}{\beta\sigma_N}\left\|x_N - \bar{x}\right\|^2.$$

Similar as the function used in proximal ADMM-g, we can prove the monotonicity and boundedness of function Ψ_L.

Lemma 3.18 *Suppose the sequence* $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ *is generated by Algorithm 3. In Algorithm 3, assume*

$$\beta > \max\left\{\frac{18L}{\sigma_N},\ \max_{1\le i\le N-1}\left\{\frac{6L^2}{\sigma_N\sigma_{\min}(H_i)}\right\}\right\}. \tag{3.30}$$

Then $\Psi_L(x^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k)$ *monotonically decreases over* $k > 0$.

Proof. By Step 1 of Algorithm 3 one observes that

$$\mathcal{L}_\beta\left(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k, \lambda^k\right) \le \mathcal{L}_\beta\left(x_1^k, \cdots, x_N^k, \lambda^k\right) - \sum_{i=1}^{N-1}\frac{1}{2}\left\|x_i^k - x_i^{k+1}\right\|_{H_i}^2, \tag{3.31}$$

while by Step 2 of Algorithm 3 we have

$$
\begin{aligned}
0 &= \left(x_N^k - x_N^{k+1}\right)^\top \left[\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - A_N{}^\top \lambda^k + \beta A_N{}^\top \left(\sum_{i=1}^N A_i x_i^{k+1} - b\right) - L\left(x_N^k - x_N^{k+1}\right)\right] \\
&\leq f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - f(x_1^{k+1}, \cdots, x_N^{k+1}) - \frac{L}{2}\left\|x_N^k - x_N^{k+1}\right\|^2 - \left(\sum_{i=1}^{N-1} A_i x_i^{k+1} + A_N x_N^k - b\right)^\top \lambda^k \\
&\quad + \left(\sum_{i=1}^N A_i x_i^{k+1} - b\right)^\top \lambda^k + \frac{\beta}{2}\left\|\sum_{i=1}^{N-1} A_i x_i^{k+1} + A_N x_N^k - b\right\|^2 - \frac{\beta}{2}\left\|\sum_{i=1}^N A_i x_i^{k+1} - b\right\|^2 \\
&\quad - \frac{\beta}{2}\left\|A_N x_N^k - A_N x_N^{k+1}\right\|^2 \\
&\leq \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k, \lambda^k) - \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^k) - \frac{L}{2}\left\|x_N^k - x_N^{k+1}\right\|^2,
\end{aligned} \tag{3.32}
$$

where the first inequality is due to (3.2) and (3.3). Moreover, from (3.29) we have

$$
\begin{aligned}
&\mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) - \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^k) = \frac{1}{\beta}\|\lambda^k - \lambda^{k+1}\|^2 \\
\leq\ & \frac{3L^2}{\beta\sigma_N}\|x_N^k - x_N^{k+1}\|^2 + \frac{6L^2}{\beta\sigma_N}\|x_N^{k-1} - x_N^k\|^2 + \frac{3L^2}{\beta\sigma_N}\sum_{i=1}^{N-1}\|x_i^k - x_i^{k+1}\|^2.
\end{aligned} \tag{3.33}
$$

Combining (3.31), (3.32) and (3.33) yields that

$$
\begin{aligned}
&\mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) - \mathcal{L}_\beta(x_1^k, \cdots, x_N^k, \lambda^k) \\
\leq\ & \left(\frac{3L^2}{\beta\sigma_N} - \frac{L}{2}\right)\|x_N^k - x_N^{k+1}\|^2 + \sum_{i=1}^{N-1}\|x_i^k - x_i^{k+1}\|^2_{\frac{3L^2}{\beta\sigma_N}I - \frac{1}{2}H_i} + \frac{6L^2}{\beta\sigma_N}\|x_N^{k-1} - x_N^k\|^2,
\end{aligned}
$$

which further implies that

$$
\begin{aligned}
&\Psi_L(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k) - \Psi_L(x_1^k, \cdots, x_N^k, \lambda^k, x_N^{k-1}) \\
\leq\ & \left(\frac{9L^2}{\beta\sigma_N} - \frac{L}{2}\right)\left\|x_N^k - x_N^{k+1}\right\|^2 + \sum_{i=1}^{N-1}\left(\frac{3L^2}{\beta\sigma_N} - \frac{\sigma_{\min}(H_i)}{2}\right)\left\|x_i^k - x_i^{k+1}\right\|^2 < 0,
\end{aligned} \tag{3.34}
$$

where the second inequality is due to (3.30). This completes the proof. □

The following lemma shows that the function Ψ_L is lower bounded.

Lemma 3.19 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ is generated by Algorithm 3. Under the same conditions as in Lemma 3.18, the sequence $\{\Psi_L(x^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k)\}$ is bounded from below.*

Proof. From Step 3 of Algorithm 3 we have

$$
\begin{aligned}
&\Psi_L(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k) \geq \mathcal{L}_\beta(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}) \\
=\ & \sum_{i=1}^{N-1} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_N^{k+1}) - \left(\sum_{i=1}^N A_i x_i^{k+1} - b\right)^\top \lambda^{k+1} + \frac{\beta}{2}\left\|\sum_{i=1}^N A_i x_i^{k+1} - b\right\|^2 \qquad (3.35) \\
=\ & \sum_{i=1}^{N-1} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_N^{k+1}) - \frac{1}{\beta}(\lambda^k - \lambda^{k+1})^\top \lambda^{k+1} + \frac{1}{2\beta}\|\lambda^k - \lambda^{k+1}\|^2 \\
=\ & \sum_{i=1}^{N-1} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_N^{k+1}) - \frac{1}{2\beta}\|\lambda^k\|^2 + \frac{1}{2\beta}\|\lambda^{k+1}\|^2 + \frac{1}{\beta}\|\lambda^k - \lambda^{k+1}\|^2 \\
\geq\ & \sum_{i=1}^{N-1} r_i^* + f^* - \frac{1}{2\beta}\|\lambda^k\|^2 + \frac{1}{2\beta}\|\lambda^{k+1}\|^2,
\end{aligned}
$$

where the third equality follows from (3.3). Summing this inequality over $k = 0, 1, \ldots, K-1$ for any integer $K \geq 1$ yields that

$$\frac{1}{K}\sum_{k=0}^{K-1} \Psi_L\left(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k\right) \geq \sum_{i=1}^{N-1} r_i^* + f^* - \frac{1}{2\beta}\left\|\lambda^0\right\|^2.$$

Lemma 3.18 stipulates that $\{\Psi_L(x_1^{k+1}, \ldots, x_N^{k+1}, \lambda^{k+1}, x_N^k)\}$ is a monotonically decreasing sequence; the above inequality thus further implies that the entire sequence is bounded from below. □

We are now ready to give the iteration complexity of proximal ADMM-m, whose proof is similar to that of Theorem 3.14.

Theorem 3.20 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k, \lambda^k)\}$ is generated by proximal ADMM-m (Algorithm 3), and β satisfies (3.30). Denote*

$$\kappa_1 := \frac{6L^2}{\beta^2\sigma_N}, \kappa_2 := 4L^2, \kappa_3 := \left(L + \beta\sqrt{N}\max_{1\leq i\leq N}\left[\|A_i\|_2^2\right] + \max_{1\leq i\leq N}\|H_i\|_2\right)^2, \kappa_4 := \max_{1\leq i\leq N-1}(\mathrm{diam}(\mathcal{X}_i))^2$$

and

$$\tau := \min\left\{-\left(\frac{9L^2}{\beta\sigma_N} - \frac{L}{2}\right), \min_{i=1,\ldots,N-1}\left\{-\left(\frac{3L^2}{\beta\sigma_N} - \frac{\sigma_{\min}(H_i)}{2}\right)\right\}\right\} > 0. \quad (3.36)$$

Letting

$$K := \begin{cases} \left\lceil \frac{2\max\{\kappa_1,\kappa_2,\kappa_3\cdot\kappa_4\}}{\tau\,\epsilon^2}(\Psi_L(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) - \sum_{i=1}^{N-1} r_i^* - f^*)\right\rceil, & \textit{for } \textbf{Settings 1,2} \\ \left\lceil \frac{2\max\{\kappa_1,\kappa_2,\kappa_3\}}{\tau\,\epsilon^2}(\Psi_L(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) - \sum_{i=1}^{N-1} r_i^* - f^*)\right\rceil, & \textit{for } \textbf{Setting 3} \end{cases} \quad (3.37)$$

and denoting $\hat{k} := \min_{2\leq k\leq K+1}\sum_{i=1}^N\left(\|x_i^k - x_i^{k+1}\|^2 + \|x_i^{k-1} - x_i^k\|^2\right)$, then $(x_1^{\hat{k}}, \cdots, x_N^{\hat{k}}, \lambda^{\hat{k}})$ is an ϵ-stationary point of problem (1.1) with A_N being full row rank, according to Definitions 3.6, 3.7 and 3.8 under the conditions of Settings 1,2 and 3 in Table 1, respectively.

Proof. By summing (3.34) over $k = 1, \ldots, K$, we obtain that

$$\Psi_L(x_1^{K+1}, \cdots, x_N^{K+1}, \lambda^{K+1}, x_N^K) - \Psi_L(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) \leq -\tau\sum_{k=1}^K\sum_{i=1}^N\left\|x_i^k - x_i^{k+1}\right\|^2, \quad (3.38)$$

where τ is defined in (3.36). From Lemma 3.19 we know that there exists a constant Ψ_L^* such that $\Psi(x_1^{k+1}, \cdots, x_N^{k+1}, \lambda^{k+1}, x_N^k) \geq \Psi_L^*$ holds for any $k \geq 1$. Therefore,

$$\min_{2\leq k\leq K+1}\theta_k \leq \frac{2}{\tau K}\left[\Psi_L(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) - \Psi_L^*\right], \quad (3.39)$$

where θ_k is defined in (3.22), i.e., for K defined as in (3.37), $\theta_{\hat{k}} = O(\epsilon^2)$.

We now give upper bounds to the terms in (3.6) and (3.7) through θ_k. Note that (3.30) implies that

$$\begin{aligned} & \|A_N^\top\lambda^{k+1} - \nabla_N f(x_1^{k+1}, \cdots, x_N^{k+1})\| \\ \leq\ & L\|x_N^k - x_N^{k+1}\| + \|\nabla_N f(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N^k) - \nabla_N f(x_1^{k+1}, \cdots, x_N^{k+1})\| \\ \leq\ & 2L\|x_N^k - x_N^{k+1}\|, \end{aligned}$$

which implies that

$$\|A_N^\top\lambda^{k+1} - \nabla_N f(x_1^{k+1}, \cdots, x_N^{k+1})\|^2 \leq 4L^2\theta_k. \quad (3.40)$$

By Step 3 of Algorithm 3 and (3.29) we have

$$\left\|\sum_{i=1}^{N} A_i x_i^{k+1} - b\right\|^2 = \frac{1}{\beta^2}\|\lambda^{k+1} - \lambda^k\|^2 \tag{3.41}$$

$$\leq \frac{3L^2}{\beta^2\sigma_N}\left\|x_N^k - x_N^{k+1}\right\|^2 + \frac{6L^2}{\beta^2\sigma_N}\left\|x_N^{k-1} - x_N^k\right\|^2 + \frac{3L^2}{\beta^2\sigma_N}\sum_{i=1}^{N-1}\left\|x_i^k - x_i^{k+1}\right\|^2 \leq \frac{6L^2}{\beta^2\sigma_N}\theta_k.$$

The remaining proof is to give upper bounds to the terms in (3.5), (3.8) and (3.9). Since the proof steps are very similar to the that of Theorem 3.14, we shall only provide the key inequalities below.

Setting 3. Under conditions in Setting 3 in Table 1, the inequality (3.26) becomes

$$\begin{aligned}&\text{dist}\left(-\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) + A_i^\top \lambda^{k+1}, \partial r_i(x_i^{k+1})\right)\\ \leq\ &\left(L + \beta\sqrt{N}\max_{1\leq i\leq N}\left[\|A_i\|_2^2\right] + \max_{1\leq i\leq N}\|H_i\|_2\right)\sqrt{\theta_k}.\end{aligned} \tag{3.42}$$

By combining (3.42), (3.40) and (3.41) we conclude that Algorithm 3 returns an ϵ-stationary point of (1.1) according to Definition 3.8 under conditions in Setting 3 in Table 1.

Setting 2. Under conditions in Setting 2 in Table 1, the inequality (3.27) becomes

$$\begin{aligned}&\left(x_i - x_i^{k+1}\right)^\top\left[g_i + \nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top\lambda^{k+1}\right]\\ \geq\ &-\left(\beta\sqrt{N}\max_{1\leq i\leq N}\left[\|A_i\|_2^2\right] + L + \max_{1\leq i\leq N}\|H_i\|_2\right)\max_{1\leq i\leq N-1}\left[\text{diam}(\mathcal{X}_i)\right]\sqrt{\theta_k}.\end{aligned} \tag{3.43}$$

By combining (3.43), (3.40) and (3.41) we conclude that Algorithm 3 returns an ϵ-stationary point of (1.1) according to Definition 3.7 under conditions in Setting 2 in Table 1.

Setting 1. Under conditions in Setting 1 in Table 1, inequality (3.28) becomes

$$\begin{aligned}&r_i(x_i) - r_i(x_i^{k+1}) + \left(x_i - x_i^{k+1}\right)^\top\left[\nabla_i f(x_1^{k+1}, \cdots, x_N^{k+1}) - A_i^\top\lambda^{k+1}\right]\\ \geq\ &-\left(\beta\sqrt{N}\max_{1\leq i\leq N}\left[\|A_i\|_2^2\right] + L + \max_{1\leq i\leq N}\|H_i\|_2\right)\max_{1\leq i\leq N-1}\left[\text{diam}(\mathcal{X}_i)\right]\sqrt{\theta_k}.\end{aligned} \tag{3.44}$$

By combining (3.44), (3.40) and (3.41) we conclude that Algorithm 3 returns an ϵ-stationary point of (1.1) according to Definition 3.6 under conditions in Setting 1 in Table 1. □

Remark 3.21 *In Step 1 of Algorithm 3, we can replace the function* $f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_N^k)$ *by its linearization*

$$f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \cdots, x_N^k) + \left(x_i - x_i^k\right)^\top \nabla_i f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \cdots, x_N).$$

Under the same conditions as in Remark 3.16, the same iteration bound follows by slightly modifying the analysis above.

4 Extensions

4.1 Relax the assumption on the last block variable x_N

It is noted that in (1.1), we have some restrictions on the last block variable x_N, i.e., $r_N \equiv 0$, $\mathcal{X}_N = \mathbb{R}^{n_N}$ and $A_N = I$ or is full row rank. A more general problem to consider is

$$\begin{array}{ll}\min & f(x_1, x_2, \cdots, x_N) + \sum\limits_{i=1}^{N} r_i(x_i)\\ \text{s.t.} & \sum_{i=1}^{N} A_i x_i = b,\ x_i \in \mathcal{X}_i,\ i = 1, \ldots, N,\end{array} \tag{4.1}$$

with the same assumptions as in (1.1), plus $\mathcal{X}_N \subset \mathbb{R}^{n_N}$ being a compact set. Note that we do not assume $A_N = I$ nor that A_N is full row rank. Moreover, note that if $r_N = 0$ and $\mathcal{X}_N = \mathbb{R}^{n_N}$, then (4.1) reduces to (1.1). In the following, we shall briefly illustrate how to use proximal ADMM-m to find an ϵ-stationary point of (4.1), and proximal ADMM-g can be applied in the same manner. We introduce the following problem that is closely related to (4.1):

$$\begin{array}{ll} \min & f(x_1, x_2, \cdots, x_N) + \sum\limits_{i=1}^{N} r_i(x_i) + \frac{\mu}{2}\|x_{N+1}\|^2 \\ \text{s.t.} & \sum_{i=1}^{N} A_i x_i + x_{N+1} = b, \; x_i \in \mathcal{X}_i, \; i = 1, \ldots, N, \end{array} \tag{4.2}$$

where $\mu > 0$. Now proximal ADMM-m is ready to be used for solving (4.2) because $A_{N+1} = I$ and x_{N+1} is unconstrained. We have the following iteration complexity result for proximal ADMM-m to obtain an ϵ-stationary point of (4.1).

Theorem 4.1 *Suppose the sequence $\{(x_1^k, \cdots, x_{N+1}^k, \lambda^k)\}$ is generated by proximal ADMM-m for solving (4.2) with $\mu = 1/\epsilon^2$, where $\epsilon > 0$ is the given tolerance. Under the same conditions and using the same notation as in Theorem 3.20, $(x_1^{\hat{k}}, \cdots, x_{N+1}^{\hat{k}}, \lambda^{\hat{k}})$ is an ϵ-stationary point of (4.2), and $(x_1^{\hat{k}}, \cdots, x_N^{\hat{k}}, \lambda^{\hat{k}})$ is an ϵ-stationary point of (4.1), according to Definitions 3.6, 3.7 and 3.8 under conditions in Settings 1,2 and 3 in Table 1, respectively.*

Proof. We only show the case for Setting 3 in Table 1, i.e., Definition 3.8, and the other two settings can be proved similarly. Note that if $(x_1^*, \cdots, x_{N+1}^*, \lambda^*)$ satisfies (3.9), (3.6) and (3.7) with N being replaced by $N+1$, then it is an ϵ-stationary point of (4.2). Therefore, according to Theorem 3.20, we have

$$\begin{aligned} \operatorname{dist}\left(-\nabla_i f(x_1^{\hat{k}}, \cdots, x_N^{\hat{k}}) + A_i^\top \lambda^{\hat{k}}, \partial r_i(x_i^{\hat{k}})\right) &\le \epsilon, \quad \text{for } i = 1, \ldots, N, &(4.3)\\ \left\| \sum_{i=1}^{N} A_i x_i^{\hat{k}} + x_{N+1}^{\hat{k}} - b \right\| &\le \epsilon. &(4.4) \end{aligned}$$

From (3.35), it holds that

$$\begin{aligned} & \Psi_L(x_1^{k+1}, \cdots, x_{N+1}^{k+1}, \lambda^{k+1}, x_{N+1}^k) \\ \ge & \sum_{i=1}^{N} r_i(x_i^{k+1}) + f(x_1^{k+1}, \cdots, x_N^{k+1}) + \frac{\|x_{N+1}^{k+1}\|^2}{2\epsilon^2} - \left(\sum_{i=1}^{N+1} A_i x_i^{k+1} - b\right)^\top \lambda^{k+1} + \frac{\beta}{2}\left\|\sum_{i=1}^{N+1} A_i x_i^{k+1} - b\right\|^2 \\ \ge & \sum_{i=1}^{N} r_i^* + f^* + \frac{\|x_{N+1}^{k+1}\|^2}{2\epsilon^2} - \frac{1}{2\beta}\|\lambda^{k+1}\|^2 - \frac{\beta}{2}\left\|\sum_{i=1}^{N+1} A_i x_i^{k+1} - b\right\|^2 + \frac{\beta}{2}\left\|\sum_{i=1}^{N+1} A_i x_i^{k+1} - b\right\|^2. \end{aligned}$$

Moreover, by Step 2 of Algorithm 3, we have that $x_{N+1}^{k+1} = \epsilon^2 \lambda^{k+1}$ in this particular problem. Recall that in Theorem 3.20, $\beta \ge 18L = 18\mu = 18/\epsilon^2$. Therefore, when $\beta > \frac{1}{\epsilon^2}$, the above inequality combined with monotonically decreasing of $\{\Psi_L(x_1^{k+1}, \cdots, x_{N+1}^{k+1}, \lambda^{k+1}, x_{N+1}^k)\}$ implies that

$$\frac{17}{18}\frac{\|x_{N+1}^{k+1}\|^2}{2\epsilon^2} \le (1 - 1/(\beta\epsilon^2))\frac{\|x_{N+1}^{k+1}\|^2}{2\epsilon^2} \le \Psi_L(x_1^1, \cdots, x_{N+1}^1, \lambda^1, x_{N+1}^0) - \sum_{i=1}^{N} r_i^* - f^*$$

As a result, the sequence $\left\{\frac{\|x_{N+1}^{k+1}\|^2}{2\epsilon^2}\right\}$ is bounded. Therefore, from (4.4) we get

$$\left\|\sum_{i=1}^{N} A_i x_i^{\hat{k}} - b\right\| \le \left\|\sum_{i=1}^{N} A_i x_i^{\hat{k}} + x_{N+1}^{\hat{k}} - b\right\| + \left\|x_{N+1}^{\hat{k}}\right\| \le C\epsilon, \tag{4.5}$$

for some constant $C > 0$. Finally, (4.3) and (4.5) imply that $(x_1^{\hat{k}}, \cdots, x_N^{\hat{k}}, \lambda^{\hat{k}})$ is an ϵ-stationary point of (4.1), according to Definition 3.8 under the conditions of Setting 3 in Table 1. □

4.2 Proximal BCD (Block Coordinate Descent)

In this section, we propose a proximal block coordinate descent method for solving the following variant of (1.1) and prove its iteration complexity:

$$\begin{array}{ll} \min & F(x_1, x_2, \cdots, x_N) := f(x_1, x_2, \cdots, x_N) + \sum_{i=1}^{N} r_i(x_i) \\ \text{s.t.} & x_i \in \mathcal{X}_i, \ i = 1, \ldots, N, \end{array} \tag{4.6}$$

where f is differentiable, r_i is nonsmooth, and $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ is a closed convex set for $i = 1, 2, \ldots, N$. Note that f and r_i can be nonconvex functions. Our proximal BCD method for solving (4.6) is described in Algorithm 4.

Algorithm 4 A proximal BCD method for solving (4.6)

Require: Given $\left(x_1^0, x_2^0, \cdots, x_N^0\right) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$

for $k = 0, 1, \ldots$ **do**

Update block x_i in a cyclic order, i.e., for $i = 1, \ldots, N$ (H_i positive definite):

$$x_i^{k+1} := \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}} \ F(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_N^k) + \frac{1}{2}\left\| x_i - x_i^k \right\|_{H_i}^2. \tag{4.7}$$

end for

Similar as settings in Table 1, depending on the properties of r_i and $\mathcal{X}_i$, the ϵ-stationary point of (4.6) can be defined as follows.

Definition 4.2 *$(x_1^*, \ldots, x_N^*, \lambda^*)$ is called an ϵ-stationary point of (4.6), if*

(i) r_i is convex, $\mathcal{X}_i$ is convex and compact, and for any $x_i \in \mathcal{X}_i$, $i = 1, \ldots, N$, it holds that

$$r_i(x_i) - r_i(x_i^*) + (x_i - x_i^*)^\top \left[\nabla_i f(x_1^*, \cdots, x_N^*) - A_i^\top \lambda^*\right] \geq -\epsilon;$$

(ii) or, if r_i is Lipschitz continuous, $\mathcal{X}_i$ is convex and compact, and for any $x_i \in \mathcal{X}_i$, $i = 1, \ldots, N$, it holds that ($g_i = \partial r_i(x_i^*)$ denotes a generalized subgradient of r_i)

$$(x_i - x_i^*)^\top \left[\nabla_i f(x_1^*, \cdots, x_N^*) + g_i - A_i^\top \lambda^*\right] \geq -\epsilon;$$

(iii) or, if r_i is lower semi-continuous, $\mathcal{X}_i = \mathbb{R}^{n_i}$ for $i = 1, \ldots, N$, it holds that

$$\operatorname{dist}\left(-\nabla_i f(x_1^*, \cdots, x_N^*) + A_i^\top \lambda^*, \partial r_i(x_i^*)\right) \leq \epsilon.$$

We now show that the iteration complexity of Algorithm 4 can be obtained from that of proximal ADMM-g. By introducing an auxiliary variable x_{N+1} and an arbitrary vector $b \in \mathbb{R}^m$, problem (4.6) can be equivalently rewritten as

$$\begin{array}{ll} \min & f(x_1, x_2, \cdots, x_N) + \sum_{i=1}^{N} r_i(x_i) \\ \text{s.t.} & x_{N+1} = b, \ x_i \in \mathcal{X}_i, \ i = 1, \ldots, N. \end{array} \tag{4.8}$$

It is easy to see that applying proximal ADMM-g to solve (4.8) (with x_{N+1} being the last block variable) reduces exactly to Algorithm 4. Hence, we have the following iteration complexity result of Algorithm 4 for obtaining an ϵ-stationary point of (4.6).

Theorem 4.3 *Suppose the sequence $\{(x_1^k, \cdots, x_N^k)\}$ is generated by proximal BCD (Algorithm 4). Denote*

$$\kappa_3 := (L + \max_{1 \le i \le N} \|H_i\|_2)^2, \;\; \kappa_4 := \max_{1 \le i \le N} (\mathrm{diam}(\mathcal{X}_i))^2.$$

Letting

$$K := \begin{cases} \left\lceil \frac{\kappa_3 \cdot \kappa_4}{\tau\, \epsilon^2} (\Psi_G(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) - \sum_{i=1}^N r_i^* - f^*) \right\rceil & \text{for } \textbf{Settings 1,2} \\ \left\lceil \frac{\kappa_3}{\tau\, \epsilon^2} (\Psi_G(x_1^1, \cdots, x_N^1, \lambda^1, x_N^0) - \sum_{i=1}^N r_i^* - f^*) \right\rceil & \text{for } \textbf{Setting 3} \end{cases}$$

with τ being defined in (3.20), and $\hat{k} := \min_{1 \le k \le K} \sum_{i=1}^N \left(\|x_i^k - x_i^{k+1}\|^2 \right)$, then $(x_1^{\hat{k}}, \cdots, x_N^{\hat{k}})$ is an ϵ-stationary point of problem (4.6).

Proof. Note that $A_1 = \cdots = A_N = 0$ and $A_{N+1} = I$ in problem (4.8). By applying proximal ADMM-g with $\beta > \max \left\{ 18L, \max_{1 \le i \le N} \left\{ \frac{6L^2}{\sigma_{\min}(H_i)} \right\} \right\}$, Theorem 3.14 holds. In particular, (3.26), (3.27) and (3.28) are valid in different settings with $\beta \sqrt{N} \max_{i+1 \le j \le N+1} [\|A_j\|_2] \, \|A_i\|_2 = 0$ for $i = 1, \ldots, N$, which leads to the choices of κ_3 and κ_4 in the above. Moreover, we do not need to consider the optimality with respect to x_{N+1} and the violation of the affine constraints, thus κ_1 and κ_2 are excluded in the expression of K, and the conclusion follows. □

5 Numerical Results

5.1 Robust Tensor PCA Problem

We consider the following nonconvex and nonsmooth model of robust tensor PCA with ℓ_1 norm regularization for third-order tensor of dimension $I_1 \times I_2 \times I_3$. Given an initial estimate R of the CP-rank, we aim to solve the following problem:

$$\begin{array}{ll} \min_{A,B,C,\mathcal{Z},\mathcal{E},\mathcal{B}} & \|\mathcal{Z} - [\![A, B, C]\!]\|^2 + \alpha \|\mathcal{E}\|_1 + \|\mathcal{B}\|^2 \\ \text{s.t.} & \mathcal{Z} + \mathcal{E} + \mathcal{B} = \mathcal{T}, \end{array} \tag{5.1}$$

where $A \in \mathbb{R}^{I_1 \times R}$, $B \in \mathbb{R}^{I_2 \times R}$, $C \in \mathbb{R}^{I_3 \times R}$, $[\![A, B, C]\!]$ is defined in (1.6). The augmented Lagrangian function of (5.1) is given by

$$\begin{aligned} & \mathcal{L}_\beta(A, B, C, \mathcal{Z}, \mathcal{E}, \mathcal{B}, \Lambda) \\ = \;& \|\mathcal{Z} - [\![A, B, C]\!]\|^2 + \alpha \|\mathcal{E}\|_1 + \|\mathcal{B}\|^2 - \langle \Lambda, \mathcal{Z} + \mathcal{E} + \mathcal{B} - \mathcal{T} \rangle + \frac{\beta}{2} \|\mathcal{Z} + \mathcal{E} + \mathcal{B} - \mathcal{T}\|^2. \end{aligned}$$

The following identities are useful for our presentation later.

$$\|\mathcal{Z} - [\![A, B, C]\!]\|^2 = \|Z_{(1)} - A(C \odot B)^\top\|^2 = \|Z_{(2)} - B(C \odot A)^\top\|^2 = \|Z_{(3)} - C(B \odot A)^\top\|^2,$$

where $Z_{(i)}$ stands for the mode-i unfolding of tensor $\mathcal{Z}$ and $\odot$ stands for the Khatri-Rao product of matrices.

Note that there are six block variables in (5.1), and we choose $\mathcal{B}$ as the last block variable. A typical iteration of proximal ADMM-g for solving (5.1) can be described as follows (we chose

$H_i = \delta_i I$, with $\delta_i > 0, i = 1, \ldots, 5$):

$$\begin{cases} A^{k+1} & = & \left((Z)^k_{(1)}(C^k \odot B^k) + \frac{\delta_1}{2}A^k\right)\left(((C^k)^\top C^k) \circ ((B^k)^\top B^k) + \frac{\delta_1}{2}I_{R\times R}\right)^{-1} \\ B^{k+1} & = & \left((Z)^k_{(2)}(C^k \odot A^{k+1}) + \frac{\delta_2}{2}B^k\right)\left(((C^k)^\top C^k) \circ ((A^{k+1})^\top A^{k+1}) + \frac{\delta_2}{2}I_{R\times R}\right)^{-1} \\ C^{k+1} & = & \left((Z)^k_{(3)}(B^{k+1} \odot A^{k+1}) + \frac{\delta_3}{2}C^k\right)\left(((B^{k+1})^\top B^{k+1}) \circ ((A^{k+1})^\top A^{k+1}) + \frac{\delta_3}{2}I_{R\times R}\right)^{-1} \\ E^{k+1}_{(1)} & = & \mathcal{S}\left(\frac{\beta}{\beta+\delta_4}(T_{(1)} + \frac{1}{\beta}\Lambda^k_{(1)} - B^k_{(1)} - Z^k_{(1)}) + \frac{\delta_4}{\beta+\delta_4}E^k_{(1)}, \frac{\alpha}{\beta+\delta_4}\right) \\ Z^{k+1}_{(1)} & = & \frac{1}{2+2\delta_5+\beta}\left(2A^{k+1}(C^{k+1} \odot B^{k+1})^\top + 2\delta_5\,(Z_{(1)})^k + \Lambda^k_{(1)} - \beta(E^{k+1}_{(1)} + B^k_{(1)} - T_{(1)})\right) \\ B^{k+1}_{(1)} & = & B^k_{(1)} - \gamma\left(2B^k_{(1)} - \Lambda^k_{(1)} + \beta(E^{k+1}_{(1)} + Z^{k+1}_{(1)} + B^k_{(1)} - T_{(1)})\right) \\ \Lambda^{k+1}_{(1)} & = & \Lambda^k_{(1)} - \beta\left(Z^{k+1}_{(1)} + E^{k+1}_{(1)} + B^{k+1}_{(1)} - T_{(1)}\right) \end{cases}$$

where $\circ$ is the matrix Hadamard product and $\mathcal{S}$ stands for the soft shrinkage operator. The updates in proximal ADMM-m are almost the same as proximal ADMM-g except $B_{(1)}$ is updated as

$$B^{k+1}_{(1)} = \frac{1}{L+\beta}\left((L-2)B^k_{(1)} + \Lambda^k_{(1)} - \beta(E^{k+1}_{(1)} + Z^{k+1}_{(1)} - T_{(1)})\right).$$

On the other hand, note that (5.1) can be equivalently written as

$$\min_{A,B,C,\mathcal{Z},\mathcal{E}} \ \|\mathcal{Z} - [\![A,B,C]\!]\|^2 + \alpha\,\|\mathcal{E}\|_1 + \|\mathcal{Z} + \mathcal{E} - \mathcal{T}\|^2, \tag{5.2}$$

which can be solved by the classical BCD method as well as our proximal BCD (Algorithm 4).

In the following we shall compare the numerical performance of BCD, proximal BCD, proximal ADMM-g and proximal ADMM-m for solving (5.1). We let $\alpha = 2/\max\{\sqrt{I_1}, \sqrt{I_2}, \sqrt{I_3}\}$ in model (5.1). We apply proximal ADMM-g and proximal ADMM-m to solve (5.1), and apply BCD and proximal BCD to solve (5.2). In all the four algorithms we set the maximum iteration number to be 2000, and the algorithms are terminated either when the maximum iteration number is reached or when θ_k as defined in (3.22) is less than 10^{-6}. The parameters used in the two ADMM variants are specified in Table 2 and we note that $L = 2$ in (5.1).

	$H_i,\ i = 1,\ldots,5$	β	γ
proximal ADMM-g	$L \cdot I$	$2L$	$\frac{1}{\beta}$
proximal ADMM-m	$L \cdot I$	$2.5L$	-

Table 2: Choices of parameters in the two ADMM variants.

In the experiment, we randomly generate 20 instances for fixed tensor dimension and CP-rank. Suppose the low-rank part $\mathcal{Z}^0$ is of rank R_{CP}. It is generated by

$$\mathcal{Z}^0 = \sum_{r=1}^{R_{CP}} a^{1,r} \otimes a^{2,r} \otimes a^{3,r},$$

where vectors $a^{i,r}$ are generated from standard Gaussian distribution for $i = 1,2,3$, $r = 1,\ldots,R_{CP}$. Moreover, a sparse tensor $\mathcal{E}^0$ is generated with cardinality of $0.001 \cdot I_1 I_2 I_3$ such that each nonzero component follows from standard Gaussian distribution. Finally, we generate noise $\mathcal{B}^0 = 0.001 * \hat{\mathcal{B}}$, where $\hat{\mathcal{B}}$ is a Gaussian tensor. Then we set $\mathcal{T} = \mathcal{Z}^0 + \mathcal{E}^0 + \mathcal{B}^0$ as the observed data in (5.1). We report the average performance of 20 instances of the four algorithms with initial guess $R = R_{CP}$ and $R = R_{CP} + 1$ in Tables 3 and 4, respectively.

In Tables 3 and 4, "Err." denotes the averaged relative error of the low-rank tensor over 20 instances $\frac{\|\mathcal{Z}^* - \mathcal{Z}^0\|_F}{\|\mathcal{Z}^0\|_F}$ where $\mathcal{Z}^*$ is the solution returned by the corresponding algorithm; "Iter."

R_{CP}	proximal ADMM-g			proximal ADMM-m			BCD			proximal BCD		
	Iter.	Err.	Num	Iter.	Err.	Num	Iter.	Err.	Num	Iter.	Err.	Num
Tensor Size $10 \times 20 \times 30$												
3	371.80	0.0362	19	395.25	0.0362	19	678.15	0.7093	1	292.80	0.0362	19
10	632.10	0.0320	17	566.15	0.0320	17	1292.10	0.9133	0	356.00	0.0154	19
15	529.25	0.0165	18	545.05	0.0165	18	1458.65	0.9224	0	753.75	0.0404	15
Tensor Size $15 \times 25 \times 40$												
5	516.30	0.0163	19	636.85	0.0437	17	611.25	0.8597	0	434.25	0.0358	18
10	671.80	0.0345	17	723.20	0.0385	17	1223.60	0.9072	0	592.60	0.0335	17
20	776.70	0.0341	16	922.25	0.0412	15	1716.05	0.9544	0	916.90	0.0416	14
Tensor Size $30 \times 50 \times 70$												
8	909.05	0.1021	13	1004.30	0.1006	13	1094.05	0.9271	0	798.05	0.1059	13
20	1304.65	0.1233	7	1386.75	0.1387	6	1635.80	0.9668	0	1102.85	0.1444	5
40	1261.25	0.0623	10	1387.40	0.0779	7	2000.00	0.9798	0	1096.80	0.0610	9

Table 3: Numerical results for tensor robust PCA with initial guess $R = R_{CP}$

R_{CP}	proximal ADMM-g			proximal ADMM-m			BCD			proximal BCD		
	Iter.	Err.	Num	Iter.	Err.	Num	Iter.	Err.	Num	Iter.	Err.	Num
Tensor Size $10 \times 20 \times 30$												
3	1830.65	0.0032	20	1758.90	0.0032	20	462.90	0.7763	0	1734.85	0.0032	20
10	1493.20	0.0029	20	1586.00	0.0029	20	1277.15	0.9133	0	1137.15	0.0029	20
15	1336.65	0.0078	19	1486.40	0.0031	20	1453.30	0.9224	0	945.05	0.0106	19
Tensor Size $15 \times 25 \times 40$												
5	1267.10	0.0019	20	1291.95	0.0019	20	609.45	0.8597	0	1471.10	0.0019	20
10	1015.25	0.0019	20	1121.00	0.0164	19	1220.50	0.9072	0	1121.40	0.0019	20
20	814.95	0.0019	20	888.40	0.0019	20	1716.30	0.9544	0	736.70	0.0020	20
Tensor Size $30 \times 50 \times 70$												
8	719.45	0.0009	20	608.25	0.0009	20	1094.10	0.9271	0	508.05	0.0327	18
20	726.95	0.0088	19	817.20	0.0220	17	1635.10	0.9668	0	539.25	0.0254	17
40	1063.55	0.0270	16	1122.75	0.0322	15	1998.05	0.9798	0	649.10	0.0246	16

Table 4: Numerical results for tensor robust PCA with initial guess $R = R_{CP} + 1$

denotes the averaged number of iterations over 20 instances; "Num" records the number of solutions (out of 20 instances) that have relative error less than 0.01.

Tables 3 and 4 suggest that BCD mostly converges to a local solution rather than the global optimal solution, while the other three methods are much better in finding the global optimum. It is interesting to note that the results presented in Table 4 are better than that of Table 3 when a slightly larger basis is allowed in tensor factorization. Moreover, in this case, proximal BCD usually consumes less number of iterations than the two ADMM variants.

References

[1] G. Allen. Sparse higher-order principal components analysis. In *The 15th International Conference on Artificial Intelligence and Statistics*, 2012. 3

[2] B. Ames and M. Hong. Alternating direction method of multipliers for sparse zero-variance discriminant analysis and principal component analysis. *Computational Optimization and Applications*, 2016. 3, 19

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010. 3

[4] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):25(1):115–129, 2015, 2015. 5

[5] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for nonstrongly convex functions. *http://arxiv.org/abs/1504.05002*, 2015. 5

[6] W. Bian and X. Chen. Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization. *SIAM Journal on Optimization*, 23:1718–1741, 2013. 3

[7] W. Bian and X. Chen. Feasible smoothing quadratic regularization method for box constrained non-Lipschitz optimization. *Technical Report*, 2014. 3

[8] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Mathematical Programming*, 149:301–327, 2015. 3

[9] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17:1205–1223, 2006. 3

[10] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010. 3

[11] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014. 3

[12] 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*, 3(1):1–122, 2011. 13

[13] K. Bredies. A forward-backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space. *Inverse Problems*, 25(1), 2009. 3

[14] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009. 5

[15] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008. 2

[16] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010. 6

[17] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. part ii: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 130(2):295–319, 2011. 6

[18] C. Cartis, N. I. M. Gould, and Ph. L. Toint. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, 2013. 6, 7

[19] X. Chen, D. Ge, Z. Wang, and Y. Ye. Complexity of unconstrained l_2-l_p minimization. *Mathematical Programming*, 143:371–383, 2014. 3

[20] O. Devolder, G. François, and Yu. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming, Ser. A*, 146:37–75, 2014. 5

[21] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001. 2

[22] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956. 5

[23] R. M. Freund and P. Grigas. New analysis and results for the Frank-Wolfe method. *Mathematical Programming*, 2013. 5

[24] D. Ge, R. He, and S. He. A three criteria algorithm for $l_2 - l_p$ minimization problem with linear constraints. Technical report, Shanghai University of Finance and Economics, 2014. 3, 9

[25] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 2014. 3, 7

[26] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *ICML*, pages 37–45, 2013. 4

[27] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152:75–112, 2015. 5

[28] M. Hong. A distributed, asynchronous and incremental algorithm for nonconvex optimization: An ADMM based approach. *http://arxiv.org/pdf/1412.6058v1.pdf*, 2014. 19

[29] M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *https://arxiv.org/abs/1604.00543*, 2016. 3, 12, 13, 19

[30] M. Hong, Z.-Q. Luo, and M M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 2015. 3, 13, 19

[31] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013. 5

[32] B. Jiang, F. Yang, and S. Zhang. Tensor and its Tucker core: the invariance relationships. *arXiv preprint arXiv:1601.01469*, 2016. 2

[33] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 146:769–783, 1998. 3

[34] G. Lan and Y. Zhou. Conditional gradient sliding for convex optimization. 2014. 5

[35] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. 3, 19

[36] T. Lin, S. Ma, and S. Zhang. Global convergence of unmodified 3-block ADMM for a class of convex minimization problems. *Preprint http://arxiv.org/abs/1505.04252*, 2015. 13

[37] T. Lin, S. Ma, and S. Zhang. Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. *Journal of Scientific Computing*, 2015. 13

[38] Y. Liu, S. Ma, Y. Dai, and S. Zhang. A smoothing SQP framework for a class of composite ℓ_q minimization over polyhedron. *Mathematical Programming Series A*, to appear, 2015. 3

[39] S. Łojasiewicz. *Une propriété topologique des sous-ensembles analytiques réels, Les Équations aux Dérivées Partielles*. Éditions du centre National de la Recherche Scientifique, Paris, 1963. 3

[40] S. Łojasiewicz. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18:556–572, 2007. 3

[41] R. Luss and M. Teboulle. Conditional gradient algorithms for rank one matrix approximations with a sparsity constraint. *SIAM Review*, 55:65–98, 2013. 8

[42] C. Mu, Y. Zhang, J. Wright, and D. Goldfarb. Scalable robust matrix recovery: Frank-Wolfe meets proximal methods. *arXiv preprint arXiv:1403.7588v1*, 2014. 5

[43] Y. E. Nesterov. *Introductory lectures on convex optimization.* Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004. 6

[44] R. T. Rockafellar and R. Wets. *Variational Analysis.* Springer, Volume 317 of Grundlehren der Mathematischen Wissenschafte, 1998. 11

[45] Y. Shen, Z. Wen, and Y. Zhang. Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29(2):239–263, 2014. 2

[46] F. Wang, W. Cao, and Z. Xu. Convergence of multiblock Bregman ADMM for nonconvex composite problems. *http://arxiv.org/abs/1505.03063*, 2015. 3

[47] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *http://arxiv.org/abs/1511.06324*, 2015. 3

[48] Y. Xu. Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation*, 7(1):39–70, 2015. 2

[49] L. Yang, T.K. Pong, and X. Chen. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. *Preprint available at http://arxiv.org/abs/1506.07029*, 2016. 3

[50] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010. 2

[51] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010. 2

[52] T. Zhang. Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293, 2013. 2