Report manuscript No.
(will be inserted by the editor)

Global Convergence of ADMM in Nonconvex Nonsmooth Optimization

Yu Wang · Wotao Yin · Jinshan Zeng†

December 6, 2017

Abstract In this paper, we analyze the convergence of the alternating direction method of multipliers (ADMM) for minimizing a nonconvex and possibly nonsmooth objective function, $\phi(x_0, \ldots, x_p, y)$, subject to coupled linear equality constraints. Our ADMM updates each of the primal variables $x_0, \ldots, x_p, y$, followed by updating the dual variable. We separate the variable y from x_i's as it has a special role in our analysis.

The developed convergence guarantee covers a variety of nonconvex functions such as piecewise linear functions, ℓ_q quasi-norm, Schatten-q quasi-norm $(0 < q < 1)$, minimax concave penalty (MCP), and smoothly clipped absolute deviation (SCAD) penalty. It also allows nonconvex constraints such as compact manifolds (e.g., spherical, Stiefel, and Grassman manifolds) and linear complementarity constraints. Also, the x_0-block can be almost any lower semi-continuous function.

By applying our analysis, we show, for the first time, that several ADMM algorithms applied to solve nonconvex models in statistical learning, optimization on manifold, and matrix decomposition are guaranteed to converge.

Our results provide sufficient conditions for ADMM to converge on (convex or nonconvex) monotropic programs with three or more blocks, as they are special cases of our model.

ADMM has been regarded as a variant to the augmented Lagrangian method (ALM). We present a simple example to illustrate how ADMM converges but ALM diverges with bounded penalty parameter β. Indicated by this example and other analysis in this paper, ADMM might be a better choice than ALM for

The work of W. Yin is supported in part by NSF grants DMS-1317602 and ECCS-1462398. The work of J. Zeng is supported in part by the National Natural Science Foundation of China grants (61603162, 11501440) and the Doctoral start-up foundation of Jiangxi Normal University.

Y. Wang
Department of Statistics, University of California, Berkeley (UCB), Berkeley, CA 94704, USA
E-mail: wang.yu@berkeley.edu

W. Yin
Department of Mathematics, University of California, Los Angeles (UCLA), Los Angeles, CA 90025, USA
E-mail: wotaoyin@ucla.edu

Corresponding author: J. Zeng
College of Computer Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
E-mail: jsh.zeng@gmail.com

some nonconvex *nonsmooth* problems, because ADMM is not only easier to implement, it is also more likely to converge for the concerned scenarios.

Keywords ADMM, nonconvex optimization, augmented Lagrangian method, block coordinate descent, sparse optimization

1 Introduction

In this paper, we consider the (possibly nonconvex and nonsmooth) optimization problem:

$$\begin{aligned} \underset{x_0,x_1,\ldots,x_p,y}{\text{minimize}} \quad & \phi(x_0,x_1,\ldots,x_p,y) \\ \text{subject to} \quad & A_0x_0 + A_1x_1 + \cdots + A_px_p + By = b, \end{aligned} \tag{1}$$

where $\phi : \mathbb{R}^{n_0} \times \cdots \times \mathbb{R}^{n_p} \times \mathbb{R}^q \to \mathbb{R} \cup \{\infty\}$ is a continuous function, $x_i \in \mathbb{R}^{n_i}$ are variables with their coefficient matrices $A_i \in \mathbb{R}^{m \times n_i}$, $i = 0,\ldots,p$, and $y \in \mathbb{R}^q$ is the last variable with its coefficient matrix $B \in \mathbb{R}^{m\times q}$. The model remains as general without y and By; but we keep y and B to simplify the notation.

We set $b = 0$ throughout the paper to simplify our analysis. All of our results still hold if $b \neq 0$ is in the image of the matrix B, i.e., $b \in \mathrm{Im}(B)$.

Besides the linear constraints in (1), any constraint on each variable $x_0, x_1, \ldots, x_p$ and y can be treated as an indicator function and included in the objective function ϕ. Therefore, we do not include constraints like:

$$x_0 \in \mathcal{X}_0, x_1 \in \mathcal{X}_1, \ldots, x_p \in \mathcal{X}_p, y \in \mathcal{Y}.$$

In spite of the success of ADMM on convex problems, the behavior of ADMM on nonconvex problems has been largely a mystery, especially when there are also nonsmooth functions and nonconvex sets in the problems. ADMM generally fails on nonconvexity problems, but it has found to not only work in some applications but often exhibit great performance! Indeed, successful examples include: matrix completion and separation [63,61,47,49], asset allocation [56], tensor factorization [34], phase retrieval [57], compressive sensing [9], optimal power flow [64], direction fields correction [31], noisy color image restoration [31], image registration [6], network inference [39], and global conformal mapping [31]. In these applications, the objective function can be nonconvex, nonsmooth, or both. Examples include the piecewise linear function, the ℓ_q quasi-norm for $q \in (0,1)$, the Schatten-q $(0 < q < 1)$ [59] quasi-norm $f(X) = \sum_i \sigma_i(X)^q$ (where $\sigma_i(X)$ denotes the ith largest singular value of X), and the indicator function $\iota_{\mathcal{B}}$, where $\mathcal{B}$ is a nonconvex set.

The success of these applications can be intriguing, since these applications are far beyond the scope of the theoretical conditions that ADMM is proved to converge. In fact, even the three-block ADMM can diverge on a simple convex problem [10]. Nonetheless, we still find that it works well in practice. This has motivated us to explore in the paper and respond to this question: when will the ADMM type algorithms converge if the objective function includes nonconvex nonsmooth functions?

We present our Algorithm 1, where $\mathcal{L}_\beta$ denotes the augmented Lagrangian (2), and show that it converges for a large class of problems. For simplicity, Algorithm 1 uses the standard ADMM subproblems, which minimize the augmented Lagrangian $\mathcal{L}_\beta$ with all but one variable fixed. It is possible to extend them to inexact, linearized, and/or prox-gradient subproblems as long as a few key principles (cf. §3.1) are preserved.

In this paper, under some assumptions on the objective and matrices, Algorithm 1 is proved to converge. Algorithm 1 is a generalization to the coordinate descent method. By setting $A_0, A_1, \ldots, A_p, B$ to 0, Algorithm 1 reduces to the *cyclic* coordinate descent method.

Algorithm 1 Nonconvex ADMM for (1)

Initialize $x_1^0, \ldots, x_p^0, y^0, w^0$
while stopping criteria not satisfied **do**
  for $i = 0, \ldots, p$ **do**
    $x_i^{k+1} \leftarrow \mathrm{argmin}_{x_i} \mathcal{L}_\beta(x_{<i}^{k+1}, x_i, x_{>i}^k, y^k, w^k)$;
  end for
  $y^{k+1} \leftarrow \mathrm{argmin}_y \mathcal{L}_\beta(\mathbf{x}^{k+1}, y, w^k)$;
  $w^{k+1} \leftarrow w^k + \beta\left(\mathbf{A}\mathbf{x}^{k+1} + By^{k+1}\right)$;
  $k \leftarrow k + 1$;
end while
return $x_1^k, \ldots, x_p^k$ and y^k.

1.1 Proposed algorithm

Our variable is $\mathbf{x} := [x_0; \ldots; x_p] \in \mathbb{R}^n$ where $n = \sum_{i=0}^p n_i$. Let $\mathbf{A} := [A_0 \ \cdots \ A_p] \in \mathbb{R}^{m \times n}$ and $\mathbf{A}\mathbf{x} := \sum_{i=0}^p A_i x_i \in \mathbb{R}^m$. To present our algorithm, we define the augmented Lagrangian:

$$\mathcal{L}_\beta(\mathbf{x}, y, w) := \phi(\mathbf{x}, y) + \langle w, \mathbf{A}\mathbf{x} + By \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} + By\|^2. \tag{2}$$

The proposed Algorithm 1 extends the standard ADMM to multiple variable blocks. It also extends the *coordinate descent* algorithms dealing with linear constraints. We let $x_{<i} := [x_0; \ldots; x_{i-1}] \in \mathbb{R}^{n_0+n_1+\cdots+n_{i-1}}$ and $x_{>i} := [x_{i+1}; \ldots; x_p] \in \mathbb{R}^{n_{i+1}+\cdots+n_p}$ (clearly, $x_{<0}$ and $x_{>p}$ are null variables, which may be used for notational ease). Subvectors $x_{\le i} := [x_{<i}, x_i]$ and $x_{\ge i}$ are defined similarly. The convergence of Algorithm 1 will be given in Theorems 1 and 2.

1.2 Relation to the augmented Lagrangian method (ALM)

ALM is a widely-used method for solving constrained optimization models [22,44]. It applies broadly to nonconvex nonsmooth problems. ADMM is an approximation to ALM by sequentially updating each of the primal variables.

ALM generally uses a sequence of penalty parameters $\{\beta^k\}$, which is nondecreasing and possibly unbounded. When β^k becomes large, the ALM subproblem becomes ill-conditioned. Therefore, using bounded β^k is practically desirable (see [11, Theorem 5.3], [3, Proposition 2.4], or [4, Chapter 7]). For general nonconvex and nonsmooth problems, it is well known that β^k, $k \in \mathbb{N}$ is bounded is not enough for the convergence of ALM [3, Section 2.1]. Proposition 2 below introduces a simple example on which ALM diverges with any bounded β^k. It is surprising, however, that ADMM converges in finite steps for any fixed $\beta > 1$ on this example.

Proposition 1 *Consider the problem*

$$\begin{aligned} \underset{x,y\in\mathbb{R}}{\text{minimize}} \quad & x^2 - y^2 \\ \text{subject to} \quad & x = y, \ x \in [-1, 1]. \end{aligned} \tag{3}$$

It holds that

1. *If $\{\beta^k | k \in \mathbb{N}\}$ is bounded, ALM generates a divergent sequence;*
2. *for any fixed $\beta > 1$, ADMM generates a convergent and finite sequence to a solution.*

The proof is straightforward and included in the Appendix. ALM diverges because $\mathcal{L}_\beta(x, y, w)$ does not have a saddle point, and there is a non-zero duality gap. ADMM, however, is unaffected. As the proof shows, the ADMM sequence satisfies $2y^k = -w^k$, $\forall k$. By substituting $w \equiv -2y$ into $\mathcal{L}_\beta(x, y, w)$, we get a convex function in (x, y)! Indeed,

$$\rho(x,y) := \mathcal{L}_\beta(x,y,w)\big|_{w=-2y} = (x^2 - y^2) + \iota_{[-1,1]}(x) - 2y(x-y) + \frac{\beta}{2}\big|x-y\big|^2 = \frac{\beta+2}{2}|x-y|^2 + \iota_{[-1,1]}(x),$$

where ι_S is the indicator function of set S (that is, $\iota_S(x) = 0$ if $x \in S$; otherwise, equals infinity). It turns out that ADMM solves (3) by performing the following coordinate descent iteration to $\rho(x, y)$:

$$\begin{cases} x^{k+1} = \operatorname{argmin}_x \rho(x, y^k), \\ y^{k+1} = y^k - \frac{\beta}{(\beta+2)^2} \frac{\mathrm{d}}{\mathrm{d}y} \rho(x^{k+1}, y^k). \end{cases}$$

Our analysis for the general case will show that the primal variable y somehow "controls" the dual variable w and reduces ADMM to an iteration that is similar to coordinate descent.

1.3 Related literature

The original ADMM was proposed in [20,18]. For convex problems, its convergence was established firstly in [19] and its convergence rates given in [21,14,15] in different settings. When the objective function is nonconvex, the recent results [61,27,37] directly make assumptions on the iterates $(\mathbf{x}^k, y^k, w^k)$. Hong et al. [23] deals with the nonconvex separable objective functions for some specific A_i, which forms the sharing and consensus problem. Li and Pong [32] studied the convergence of ADMM for some special nonconvex models, where one of the matrices A and B is an identity matrix. Wang et al. [51,52] studied the convergence of the nonconvex Bregman ADMM algorithm, which includes ADMM as a special case. We review their results and compare to ours in §4 below.

1.4 Contribution and novelty

The main contribution of this paper is the establishment of the global convergence of Algorithm 1 under certain assumptions given in Theorems 1 and 2 below. The assumptions apply to largely many nonconvex and nonsmooth objective functions. The developed theoretical results can be extended to the case where subproblems are solved inexactly with summable errors. We also allow the primal block variables $x_1, \ldots, x_p$ to be updated in an arbitrary order as long as x_0 is updated first and y is updated last (just before the w-update). The novelty of this paper can be summarized as follows:

(1) **Weaker assumptions.** Compared to the related works [61,27,37,23,32,51,52], the convergence conditions in this paper are weaker, extending the ADMM theory to significantly more nonconvex functions and nonconvex sets. See Table 1. In addition, we allow the primal variables $x_1, \ldots, x_p$ to be updated in an arbitrary order at each iteration[1], which is new in the ADMM literature. We show that most of our

[1] This is the best that one hope (except for very specific problems) since [62, Section 1] shows a convex 2-block problem, which ADMM fails to converge.

assumptions are necessary by providing counter examples. We also give the first example that causes ADMM to converge but ALM to diverge.

(2) **New examples.** By applying our main theorems, we prove convergence for the nonconvex ADMM applied to the following problems which could not be recovered from previous convergence theory:
   – statistical regression based on nonconvex regularizer such as minimax concave penalty(MCP), smoothly clipped absolute deviation (SCAD), and ℓ_q quasi-norm;
   – minimizing smooth functions subject to norm or Stiefel/Grassmannian manifold constraints;
   – matrix decomposition using nonconvex Schatten-q regularizer;
   – smooth minimization subject to complementarity constraints.

(3) **Novel techniques.** We improve upon the existing analysis techniques and introduce new ones.
   (a) *An induction technique for nonconvex, nonsmooth case.* The analysis uses the augmented Lagrangian as the Lyapunov function: Algorithm 1 produces a sequence of points whose augmented Lagrangian function values are decreasing and lower bounded. This technique appeared first in [23] and also in [32, 51]. However, it has trouble handling nonsmooth functions. An induction technique is introduced to overcome this difficulty and extend the current framework to nonconvex, nonsmooth, multi-block cases. The technique is used in the proof of Lemma 9.
   (b) *Restricted prox-regularity.* Most of the convergence analysis of nonconvex optimization either assumes or proves the sufficient descent and bounded subgradient properties (c.f., [1, 23]). This property is easily obtainable if the objective is smooth. However, some nonconvex and nonsmooth objectives (e.g. nonconvex ℓ_q quasi-norm) violate these properties. We overcome this challenge with the introduced *restricted prox-regularity property* (Definition 2). If the objective satisfies such a property, we prove that the sequence enjoy sufficient descent and bounded subgradients after a finite number of iterations.
   (c) *More general linear mappings.* Most nonconvex ADMM analysis is applied to the primal variables $\mathbf{x}$ and y directly. This requires the matrices $A_0, A_1, \dots, A_p, B$ to either identity or have full column/row rank. In this paper, we introduce techniques to work with possibly rank-deficient $A_0, A_1, \dots, A_p, B$ (see, for example, Lemma 5). This allows us to ensure convergence of ADMM on some important applications in signal processing and statistical learning (see §5), and also gives some positive answers to the open problem presented at the end of [32], i.e., "*it is an interesting problem to study the convergence of ADMM when B is injective*".

   In addition, we use several other techniques that are tailored to relax our convergence assumptions as much as possible.

1.5 Notation and organization

We denote $\mathbb{R}$ as the real number set, $\mathbb{R} \cup \{+\infty\}$ as the extended real number set, $\mathbb{R}_+$ as the positive real number set, and $\mathbb{N}$ as the natural number set. Given a matrix X, $\mathrm{Im}(X)$ denotes its image, $\sigma_i(X)$ denotes its ith largest singular value. $\|\cdot\|$ represents the Euclidean norm for a vector or the Frobenius norm for a matrix. $\mathrm{dom}(f)$ denotes the domain of a function f. For any two square matrices A and B with the same size, $A \succeq B$ means that $A - B$ is positively semi-definite.

The remainder of this paper is organized as follows. Section 2 presents the main convergence analysis. Section 3 gives the detailed proofs. Section 4 discusses the tightness of the assumptions, the primal variable update order, and inexact minimization issues. Section 5 applies the developed theorems in some typical applications and obtains novel convergence results. Finally, Section 6 concludes this paper.

Table 1 Conditions for ADMM convergence (note: $f_0, f_1, \ldots, f_p$ is not required to exist)

<table>
<tr><td></td><td colspan="2">Scenario 1</td><td>Scenario 2</td></tr>
<tr><td>model</td><td colspan="2">$\underset{\mathbf{x}=(x_0,\ldots,x_p),y}{\text{minimize}} \quad \phi(\mathbf{x}, y) := g(\mathbf{x}) + \sum_{i=0}^{p} f_i(x_i) + h(y)$
subject to $\quad \mathbf{A}\mathbf{x} + By = 0$</td><td>$\underset{\mathbf{x},y}{\text{minimize}} \ \phi(\mathbf{x}, y)$
subject to $\mathbf{A}\mathbf{x} + By = 0$</td></tr>
<tr><td>ϕ</td><td colspan="3">coercive over the feasible set $\{(\mathbf{x}, y) : \mathbf{A}\mathbf{x} + By = 0\}$; see assumption A1</td></tr>
<tr><td>g, h</td><td colspan="2">Lipschitz differentiable</td><td rowspan="4">ϕ Lipschitz differentiable</td></tr>
<tr><td></td><td>Scenario 1a</td><td>Scenario 1b</td></tr>
<tr><td>f_0</td><td>lower semi-continuous</td><td>∂f bounded in any bounded set</td></tr>
<tr><td>$f_1, \ldots, f_p$</td><td>restricted prox-regular</td><td>piecewise linear</td></tr>
<tr><td>$\mathbf{A}, B$</td><td colspan="3">$\text{Im}(\mathbf{A}) \subseteq \text{Im}(B)$</td></tr>
<tr><td></td><td colspan="3">solution to each ADMM sub-problem is Lipschitz w.r.t. input (A3)</td></tr>
</table>

2 Main results

2.1 Definitions

In our definitions, ∂f denotes the set of general subgradients of f in [45, Definition 8.3]. We call a function *Lipschitz differentiable* if it is differentiable and its gradient is Lipschitz continuous. The functions given in the next two definitions are permitted in our model.

Definition 1 (Piecewise linear function) A function $f : \mathbb{R}^n \to \mathbb{R}$ is *piecewise linear* if there exist polyhedra $U_1, \ldots, U_K \subset \mathbb{R}^n$, vectors $a_1, \ldots, a_K \in \mathbb{R}^n$, and points $b_1, \ldots, b_K \in \mathbb{R}$ such that $\bigcup_{i=1}^{K} \overline{U_i} = \mathbb{R}^n$, $U_i \bigcap U_j = \emptyset$ $(\forall\, i \neq j)$, and $f(x) = a_i^T x + b_i$ when $x \in U_i$, $i = 1, \ldots, K$.

Definition 2 (Restricted prox-regularity) For a lower semi-continuous function f, let $M \in \mathbb{R}_+$, $f : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, and define the exclusion set

$$S_M := \{x \in \text{dom}(f) : \|d\| > M \text{ for all } d \in \partial f(x)\}.$$

f is called *restricted prox-regular* if, for any $M > 0$ and bounded set $T \subseteq \text{dom} f$, there exists $\gamma > 0$ such that

$$f(y) + \frac{\gamma}{2}\|x - y\|^2 \geq f(x) + \langle d, y - x\rangle, \quad \forall\, x \in T \setminus S_M,\ y \in T,\ d \in \partial f(x),\ \|d\| \leq M. \tag{4}$$

(If $T \setminus S_M$ is empty, (4) is satisfied.)

Definition 2 is related to, but weaker than, the concepts *prox-regularity* [43], *hypomonotonicity* [45, Example 12.28] and *semi-convexity* [38,26,30,40], all of which impose global conditions. Definition 2 only requires (4) to hold over a subset. As shown in Proposition 1, while prox-regular functions include any convex functions and any C^1 functions with Lipschitz continuous gradients, restricted prox-regular functions further include a set of non-smooth non-convex functions such as ℓ_q quasi-norms $(0 < q < 1)$, Schatten-q quasi-norms $(0 < q < 1)$, and indicator functions of compact smooth manifolds.

Proposition 1 ***Examples of (restricted) prox-regular functions*** *The following functions are prox-regular functions, which are thus restricted prox-regular:*

(1) convex functions, including indicator functions of convex sets,
(2) C^1 smooth functions with L-Lipschitz continuous gradient, $L \geq 0$.

The following functions are restricted prox-regular functions, which are not prox-regular:

(a) $\ell_q(x) := \|x\|_q^q$ function for $q \in (0,1)$;
(b) Schatten-q quasi-norm: $\|A\|_q = \sum_{i=1}^n \sigma_i^q$, where $q \in (0,1)$ and σ_i is the ith largest singular value of A;
(c) Indicator functions ι_S of a compact C^2 manifold, such as the unit sphere in a finite Euclidean space.

Definition 2 introduces functions that do not satisfy (4) globally *only because* they are asymptotically "steep" in the exclusion set S_M. Such functions include $|x|^q$ $(0 < q < 1)$, for which S_M has the form $(-\epsilon_M, 0) \cup (0, \epsilon_M)$; the Schatten-$q$ quasi-norm $(0 < q < 1)$, for which $S_M = \{X : \exists i,\ \sigma_i(X) < \epsilon_M\}$ as well as $\log(x)$, for which $S_M = (0, \epsilon_M)$, where ϵ_M is a constant depending on M. We only need (4) because the iterates x_i^k of Algorithm 1, for all large k, never enter the exclusion set S_M.

2.2 Main theorems

To ensure the boundedness of the sequence $(\mathbf{x}^k, y^k, w^k)$, we only need the coercivity of the objective function within the feasible set.

A1 **(coercivity)** Define the feasible set $\mathcal{F} := \{(\mathbf{x}, y) \in \mathbb{R}^{n+q} : \mathbf{A}\mathbf{x} + By = 0\}$. The objective function $\phi(\mathbf{x}, y)$ is coercive over this set, that is, $\phi(\mathbf{x}, y) \to \infty$ if $(\mathbf{x}, y) \in \mathcal{F}$ and $\|(\mathbf{x}, y)\| \to \infty$;

If the feasible set of $(\mathbf{x}, y)$ is bounded, then A1 holds trivially for any continuous objective function. Therefore, A1 is much weaker than assuming that the objective function is coercive over the entire space $\mathbb{R}^{n+q}$. The assumption A1 can be dropped if the boundedness of the sequence can be deducted from other means.

Within the proof, $A_i x_i^k$ and By^k often appear in the first order conditions (e.g. see equations (12), (13)). In order to have a reverse control, i.e., controlling x_i^k, y^k based on $A_i x_i^k, By^k$, we need the following two assumptions on matrices A_i and B.

A2 **(feasibility)** $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(B)$, where $\mathrm{Im}(\cdot)$ returns the image of a matrix;

A3 **(Lipschitz sub-minimization paths)**
(a) For any fixed $\mathbf{x}$, $\mathrm{argmin}_y\{\phi(\mathbf{x}, y) : By = u\}$ has a unique minimizer. $H : \mathrm{Im}(B) \to \mathbb{R}^q$ defined by $H(u) \triangleq \mathrm{argmin}_y\{\phi(\mathbf{x}, y) : By = u\}$ is a Lipschitz continuous map.
(b) For $i = 0, \ldots, p$ and any $x_{<i}$, $x_{>i}$ and y, $\mathrm{argmin}_{x_i}\{\phi(x_{<i}, x_i, x_{>i}, y) : A_i x_i = u\}$ has a unique minimizer and $F_i : \mathrm{Im}(A_i) \to \mathbb{R}^{n_i}$ defined by $F_i(u) \triangleq \mathrm{argmin}_{x_i}\{\phi(x_{<i}, x_i, x_{>i}, y) : A_i x_i = u\}$ is a Lipschitz continuous map.

Moreover, the above F_i and H have a universal Lipschitz constant $\bar{M} > 0$.

These two assumptions allow us to control x_i^k, y^k by $A_i x_i^k, By^k$ as in Lemma 1.

Lemma 1 *It holds that, $\forall k_1, k_2 \in \mathbb{N}$,*

$$\|y^{k_1} - y^{k_2}\| \leq \bar{M}\|By^{k_1} - By^{k_2}\|, \tag{5}$$

$$\|x_i^{k_1} - x_i^{k_2}\| \leq \bar{M}\|A_i x_i^{k_1} - A_i x_i^{k_2}\|, \quad i = 0, 1, \ldots, p, \tag{6}$$

where $\bar{M}$ is given in A3.

They weaken the full column rank assumption typically imposed on matrices A_i and B. When A_i and B have full column rank, their null spaces are trivial and, therefore, F_i, H reduce to linear operators and satisfy A3. However, the assumption A3 allows non-trivial null spaces and holds for more functions. For example, if a function f is a C^2 with its Hessian matrix H bounded everywhere $\sigma_1 I \succeq H \succeq \sigma_2 I$ $(\sigma_1 > \sigma_2 > 0)$, then F satisfies A3 for any matrix A. If the uniqueness fails to hold, i.e., there exists y_1, y_2 such that $By_1 = By_2$ and $\phi(\mathbf{x}, y_1) = \phi(\mathbf{x}, y_2)$, then the augmented Lagrangian cannot distinguish them, causing troubles to the boundedness of the sequence.

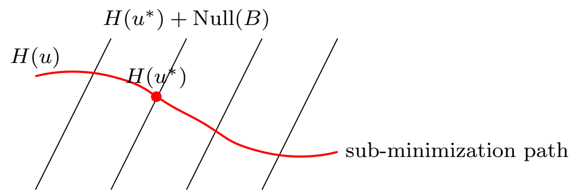

Fig. 1 Illustration of the assumption A3, which assumes that $H(u) = \mathrm{argmin}\{h(y) : By = u\}$ is Lipschitz [46].

As for the objective function, we consider two different scenarios:

- Theorem 1 considers the scenario where $\mathbf{x}$ and y are decoupled in the objective function;
- Theorem 2 considers the scenario where $\mathbf{x}$ and y are possibly coupled but their function $\phi(\mathbf{x}, y)$ is Lipschitz differentiable.

The model in the first scenario is

$$\begin{aligned} \underset{x_0, x_1, \ldots, x_p, y}{\text{minimize}} \quad & f(x_0, x_1, \ldots, x_p) \quad + h(y) \\ \text{subject to} \quad & A_0 x_0 + A_1 x_1 + \cdots + A_p x_p + By = b, \end{aligned} \tag{7}$$

where the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ $(n = \sum_{i=0}^{p} n_i)$ is proper, continuous, and possibly nonsmooth, and the function $h : \mathbb{R}^q \to \mathbb{R}$ is proper and differentiable. Both f and h can be nonconvex.

Theorem 1 *Suppose that A1-A3 and the following assumptions hold.*

A4 ***(objective-f regularity)*** *f has the form*

$$f(\mathbf{x}) := g(\mathbf{x}) + \sum_{i=0}^{p} f_i(x_i)$$

where

(i) $g(\mathbf{x})$ is Lipschitz differentiable with constant L_g,

(ii) Either

a. f_0 is lower semi-continuous, $f_i(x_i)$ is restricted prox-regular (Definition 2) for $i = 1, \ldots, p$; Or,

b. The supremum $\sup\{\|d\| : x_0 \in S, d \in \partial f_0(x_0)\}$ is bounded for any bounded set S, $f_i(x_i)$ is continuous and piecewise linear (Definition 1) for $i = 1, \ldots, p$;

A5 ***(objective-h regularity)*** *$h(y)$ is Lipschitz differentiable with constant L_h;*

Then, Algorithm 1 converges subsequently for any sufficiently large β (the lower bound is given in Lemma 9), that is, starting from any $x_1^0, \dots, x_p^0, y^0, w^0$, it generates a sequence that is bounded, has at least one limit point, and that each limit point $(\mathbf{x}^*, y^*, w^*)$ is a stationary point of $\mathcal{L}_\beta$, namely, $0 \in \partial \mathcal{L}_\beta(\mathbf{x}^*, y^*, w^*)$.

In addition, if $\mathcal{L}_\beta$ is a Kurdyka-Lojasiewicz (KL) function [35, 5, 1], then $(\mathbf{x}^k, y^k, w^k)$ converges globally[2] to the unique limit point $(\mathbf{x}^*, y^*, w^*)$.

Assumptions A4 and A5 regulate the objective functions. None of the functions needs to be convex. f_0 can be any lower semi-continuous function, and the non-Lipschitz differentiable parts $f_1, \dots, f_n$ of f shall satisfy either Definition 1 or Definition 2. Under Assumptions A4 and A5, the augmented Lagrangian function $\mathcal{L}_\beta$ is continuous.

It will be easy to see, from our proof in Section 3.3, that the Lipschitz differentiable assumption on g can be relaxed to hold just in any bounded set, since the boundedness of $\{\mathbf{x}^k\}$ is established before that property is used in our proof. Consequently, g can be functions like e^x, whose derivative is not globally Lipschitz.

Functions satisfying the KŁ inequality include real analytic functions, semi-algebraic functions and locally strongly convex functions (more information can be referred to Sec. 2.2 in [60] and references therein).

In the second scenario, $\mathbf{x}$ and y can be coupled in the objective as shown in (1), but the objective needs to be smooth.

Theorem 2 *Suppose that A1-A3 hold and ϕ in (1) is Lipschitz differentiable with constant L_ϕ. Then, Algorithm 1 has the same subsequential and global convergence results as stated in Theorem 1.*

Although Theorems 1 and 2 impose different conditions on the objective functions, their proofs are similar. Hence, we will focus on proving Theorem 1 first and leave the proof of Theorem 2 to the Appendix.

3 Proof

3.1 Keystones

The following properties hold for Algorithm 1 under our assumptions. Here, we first list them and present Proposition 2, which establishes convergence assuming these properties. Then in the next two subsections, we prove these properties.

P1 **(Boundedness)** $\{\mathbf{x}^k, y^k, w^k\}$ is bounded, and $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded.

P2 **(Sufficient descent)** There is a constant $C_1(\beta) > 0$ such that for all sufficiently large k, we have

$$\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \geq C_1(\beta)\Big(\|B(y^{k+1} - y^k)\|^2 + \sum_{i=1}^{p} \|A_i(x_i^k - x_i^{k+1})\|^2\Big). \tag{8}$$

P3 **(Subgradient bound)** There exists $C_2(\beta) > 0$ and $d^{k+1} \in \partial \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1})$ such that

$$\|d^{k+1}\| \leq C_2(\beta)\Big(\|B(y^{k+1} - y^k)\| + \sum_{i=1}^{p} \|A_i(x_i^{k+1} - x_i^k)\|\Big). \tag{9}$$

It is our intention to start i at 1, thus skipping the x_0-block, in (8) and (9).

[2] *"Globally"* here means regardless of where the initial point is.

The proposition below is standard and not new though it does not appear exactly in the literature.

Proposition 2 *Suppose that when an algorithm is applied to the problem (7), its sequence* $(\mathbf{x}^k, y^k, w^k)$ *satisfies P1–P3. Then, the sequence has at least a limit point* $(\mathbf{x}^*, y^*, w^*)$*, and any limit point* $(\mathbf{x}^*, y^*, w^*)$ *is a stationary point. That is,* $0 \in \partial\mathcal{L}_\beta(\mathbf{x}^*, y^*, w^*)$*, or equivalently,*

$$0 = \mathbf{A}\mathbf{x}^* + By^*, \tag{10a}$$

$$0 \in \partial f(\mathbf{x}^*) + \mathbf{A}^T w^*, \tag{10b}$$

$$0 \in \partial h(y^*) + B^T w^*. \tag{10c}$$

Furthermore, the running best rates[3] *of the sequences* $\{\|B(y^{k+1} - y^k)\|^2 + \sum_{i=1}^p \|A_i(x_i^k - x_i^{k+1})\|^2\}$ *and* $\{\|d^{k+1}\|\}$ *are* $o(\frac{1}{k})$ *and* $o(\frac{1}{\sqrt{k}})$*, respectively. Moreover, if* $\mathcal{L}_\beta$ *is a KŁ function, then* $(\mathbf{x}^k, y^k, w^k)$ *converges globally to the unique point* $(\mathbf{x}^*, y^*, w^*)$.

Proof The proof is standard. Similar steps are found in, for example, [1,60].

By P1, the sequence $(\mathbf{x}^k, y^k, w^k)$ is bounded, so there exist a convergent subsequence and a limit point, denoted by $(\mathbf{x}^{k_s}, y^{k_s}, w^{k_s})_{s\in\mathbb{N}} \to (\mathbf{x}^*, y^*, w^*)$ as $s \to +\infty$. By P1 and P2, $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is monotonically nonincreasing and lower bounded, and therefore $\|A_i x_i^k - A_i x_i^{k+1}\| \to 0$ and $\|By^k - By^{k+1}\| \to 0$ as $k \to \infty$. Based on P3, there exists $d^k \in \partial\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ such that $\|d^k\| \to 0$. In particular, $\|d^{k_s}\| \to 0$ as $s \to \infty$. By definition of general subgradient [45, Definition 8.3], we have $0 \in \partial\mathcal{L}_\beta(\mathbf{x}^*, y^*, w^*)$.

The running best rate of the sequence $\{\|B(y^{k+1} - y^k)\|^2 + \sum_{i=1}^p \|A_i(x_i^k - x_i^{k+1})\|^2\}$ can be easily obtained via taking advantage of [16, Lemma 1.2] or [29, Theorem 3.3.1]. By (9), it is obvious that the running best rate of the sequence $\{\|d^{k+1}\|\}$ is $o(\frac{1}{\sqrt{k}})$.

Similar to the proof of Theorem 2.9 in [1], we can claim the global convergence of the considered sequence $(\mathbf{x}^k, y^k, w^k)_{k\in\mathbb{N}}$ under the KŁ assumption of $\mathcal{L}_\beta$. □

In P2, the sufficient descent inequality (8) is only required for any sufficiently large k, not all k. In our analysis, P1 gives subsequence convergence, P2 measures the augmented Lagrangian descent, and P3 bounds the subgradient by total point changes. The reader may still obtain P1–P3 when generalizing Algorithm 1, for example, by replacing the direct minimization subproblems to prox-gradient or inexact subproblems and by relaxing the ordering in which the primal variables are updated.

3.2 Preliminaries

In this subsection, we give some useful lemmas that will be used in the main proof. To save space, throughout this section we assume assumptions A1–A5 hold, and let

$$(\mathbf{x}^+, y^+, w^+) := (\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}). \tag{11}$$

In addition, we let $A_{<s}x_{<s} := \sum_{i<s} A_i x_i$ and, in a similar fashion, $A_{>s}x_{>s} := \sum_{i>s} A_i x_i$.

Lemma 2 *If* $\beta > \bar{M}^2 L_h$ *(*$\bar{M}$ *is defined in A3), all the subproblems in Algorithm 1 are well defined.*

This lemma is on its own, so we leave its proof to the appendix.

[3] A nonnegative sequence a_k induces its running best sequence $b_k = \min\{a_i : i \le k\}$; therefore, a_k has running best rate of $o(1/k)$ if $b_k = o(1/k)$.

Lemma 3 (bound dual by primal) *Let $\lambda_{++}(B^TB)$ be the smallest strictly-positive eigenvalue of B^TB, $C \triangleq L_h\bar{M}\lambda_{++}^{-1/2}(B^TB)$. For all $k \in \mathbb{N}$, it holds that*

(a) $B^Tw^k = -\nabla h(y^k)$.
(b) $\|w^+ - w^k\| \leq C\|By^+ - By^k\|$.

Proof Part (a) follows directly from the optimality condition of y^k: $0 = \nabla h(y^k) + B^Tw^{k-1} + \beta B^T(A\mathbf{x}^k + By^k)$, and $w^k = w^{k-1} + \beta\left(\mathbf{A}\mathbf{x}^k + By^k\right)$.

Then let us prove Part (b). Since $w^+ - w^k = \beta(A\mathbf{x}^+ + By^+) \in \mathrm{Im}(B)$, we get

$$\|w^+ - w^k\| \leq \lambda_{++}^{-1/2}(B^TB)\|B^T(w^+ - w^k)\| = \lambda_{++}^{-1/2}(B^TB)\|\nabla h(y^+) - \nabla h(y^k)\| \leq C\|By^+ - By^k\|.$$

The last inequality follows from the Lipschitz property of ∇h and Lemma 1. □

3.3 Main proof

This subsection proves Theorem 1 for Algorithm 1 under Assumptions A1–A5. For all $k \in \mathbb{N}$ and $i = 0, \ldots, p$, because of the optimality of x_i^k, we can introduce the following *general subgradients* d_i^k and $\bar{d}_i^k$,

$$\bar{d}_i^k := -(A_i^Tw^+ + \beta\rho_i^k) \in \partial_i f(x_{<i}^+, x_i^+, x_{>i}^k), \tag{12}$$

$$d_i^k := -\nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k) + \bar{d}_i^k \in \partial f_i(x_i^+), \tag{13}$$

where

$$\rho_i^k := A_i^T(A_{>i}x_{>i}^k - A_{>i}x_{>i}^+) + A_i^T(By^k - By^+).$$

The next two lemmas estimate the descent of $\mathcal{L}_\beta(\mathbf{x}, y, w)$ at each iteration.

Lemma 4 (descent of $\mathcal{L}_\beta$ during x_i update) *The iterates in Algorithm 1 satisfy*

1. $\mathcal{L}_\beta(x_{<i}^+, \boldsymbol{x}_i^k, x_{>i}^k, y^k, w^k) \geq \mathcal{L}_\beta(x_{<i}^+, \boldsymbol{x}_i^+, x_{>i}^k, y^k, w^k)$, $i = 0, \ldots, p$;
2. $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) \geq \mathcal{L}_\beta(\mathbf{x}^+, y^k, w^k)$;
3. $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^+, y^k, w^k) = \sum_{i=0}^p r_i$, *where*

$$r_i := f(x_{<i}^+, x_i^k, x_{>i}^k) - f(x_{<i}^+, x_i^+, x_{>i}^k) - \langle \bar{d}_i^k, x_i^k - x_i^+\rangle + \frac{\beta}{2}\|A_ix_i^k - A_ix_i^+\|^2 \geq 0, \tag{14}$$

where $\bar{d}_i^k$ is defined in (12).

4. *For $i = 1, \ldots, p$ (without the block $i = 0$), if*

$$f_i(x_i^k) + \frac{\gamma_i}{2}\|x_i^k - x_i^+\|^2 \geq f_i(x_i^+) + \langle d_i^k, x_i^k - x_i^+\rangle, \tag{15}$$

holds with constant $\gamma_i \geq 0$ (later, this condition will be shown to hold), then we have

$$r_i \geq \frac{\beta - \gamma_i\bar{M}^2 - L_g\bar{M}^2}{2}\|A_ix_i^k - A_ix_i^+\|^2, \tag{16}$$

where the constants L_g and $\bar{M}$ are defined in Assumptions A4 and A3, respectively.

Proof **Part 1** follows directly from the minimization subproblems, which give x_i^+. **Part 2** is a result of

$$\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^+, y^k, w^k) = \sum_{i=0}^{p} \left(\mathcal{L}_\beta(x_{<i}^+, x_i^k, x_{>i}^k, y^k, w^k) - \mathcal{L}_\beta(x_{<i}^+, x_i^+, x_{>i}^k, y^k, w^k)\right),$$

and part 1. **Part 3**: Each term in the sum equals $f(x_{<i}^+, x_i^k, x_{>i}^k) - f(x_{<i}^+, x_i^+, x_{>i}^k)$ *plus*

$$\begin{aligned}
&\langle w^k, A_i x_i^k - A_i x_i^+\rangle + \frac{\beta}{2}\|A_{<i}x_{<i}^+ + A_i x_i^k + A_{>i}x_{>i}^k + By^k\|^2 - \frac{\beta}{2}\|A_{<i}x_{<i}^+ + A_i x_i^+ + A_{>i}x_{>i}^k + By^k\|^2 \\
&= \langle w^k, A_i x_i^k - A_i x_i^+\rangle + \langle \beta\left(A_{<i}x_{<i}^+ + A_i x_i^+ + A_{>i}x_{>i}^k + By^k\right), A_i x_i^k - A_i x_i^+\rangle + \frac{\beta}{2}\|A_i x_i^k - A_i x_i^+\|^2 \\
&= \langle A_i^T w^+ + \beta\rho_i^k, x_i^k - x_i^+\rangle + \frac{\beta}{2}\|A_i x_i^k - A_i x_i^+\|^2
\end{aligned}$$

where the first equality follows from the cosine rule: $\|b+c\|^2 - \|a+c\|^2 = \|b-a\|^2 + 2\langle a+c, b-a\rangle$ with $b = A_i x_i^k$, $a = A_i x_i^+$, and $c = A_{<i}x_{<i}^+ + A_{>i}x_{>i}^k + By^k$.

Part 4. Let d_i^k be defined in (13). From the inequalities (6) and (15), we get

$$f_i(x_i^k) - f_i(x_i^+) - \langle d_i^k, x_i^k - x_i^+\rangle \geq -\frac{\gamma_i}{2}\|x_i^k - x_i^+\|^2 \geq -\frac{\gamma_i \bar{M}^2}{2}\|Ax_i^k - Ax_i^+\|^2. \tag{17}$$

By the assumption A4 part (i) and (6), we also get

$$g(x_{<i}^+, x_i^k, x_{>i}^k) - g(x_{<i}^+, x_i^+, x_{>i}^k) - \langle \nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k), x_i^k - x_i^+\rangle \geq -\frac{L_g}{2}\|x_i^k - x_i^+\|^2 \geq -\frac{L_g \bar{M}^2}{2}\|Ax_i^k - Ax_i^+\|^2. \tag{18}$$

Finally, rewriting the expression of r_i and applying (17) and (18) we obtain

$$\begin{aligned}
r_i &= \left(g(x_{<i}^+, x_i^k, x_{>i}^k) - g(x_{<i}^+, x_i^+, x_{>i}^k) - \langle \nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k), x_i^k - x_i^+\rangle\right) \\
&\quad + \left(f_i(x_i^k) - f_i(x_i^+) - \langle d_i^k, x_i^k - x_i^+\rangle\right) + \frac{\beta}{2}\|Ax_i^k - Ax_i^+\|^2 \\
&\geq \frac{\beta - \gamma_i \bar{M}^2 - L_g \bar{M}^2}{2}\|A_i x_i^k - A_i x_i^+\|^2.
\end{aligned}$$

□

The assumption (15) in the part 4 of Lemma 4 is the same as (4) in Definition 2 except the latter holds for more functions due to the exclusion set S_M. In order to relax (15) to (4), we must find M and specify the exclusion set S_M. (This complicates our analysis but is necessary for many nonconvex functions such as the ℓ_q quasi-norm.) We will finally achieve this relaxation in Lemma 9.

Lemma 5 (descent of $\mathcal{L}_\beta$ due to y and w updates) *If $\beta > 2(L_h\bar{M}^2 + 1 + C)$, where C is the constant specified in Lemma 3 and L_h is the Lipschitz constant in Assumption A5, then for any $k \in \mathbb{N}$*

$$\mathcal{L}_\beta(\mathbf{x}^+, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+) \geq \|By^+ - By^k\|^2. \tag{19}$$

Proof Because $\beta/2 > L_h \bar{M}^2 + 1 + C$ and $\beta^{-1} < 1/C$, we know

$$\frac{\beta}{2} - \frac{C^2}{\beta} - \frac{L_h \bar{M}^2}{2} > L_h \bar{M}^2 + 1 + C - C - \frac{L_h \bar{M}^2}{2} > 1. \tag{20}$$

From the assumption A5 and Lemma 3(b), it follows

$$\begin{aligned}
&\mathcal{L}_\beta(\mathbf{x}^+, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+) \\
&= h(y^k) - h(y^+) + \langle w^+, By^k - By^+ \rangle + \frac{\beta}{2}\|By^+ - By^k\|^2 - \frac{1}{\beta}\|w^+ - w^k\|^2 \qquad (21)\\
&\geq -\frac{L_h \bar{M}^2}{2}\|By^+ - By^k\|^2 + \frac{\beta}{2}\|By^+ - By^k\|^2 - \frac{C^2}{\beta}\|By^+ - By^k\|^2 \qquad (22)\\
&\geq \|By^+ - By^k\|^2,
\end{aligned}$$

The last inequality holds because of (20). □

Based on Lemma 4 and Lemma 5, we now establish the following results:

Lemma 6 (Monotone, lower–bounded $\mathcal{L}_\beta$ and (P1) bounded sequence) *If $\beta > 2(L_h \bar{M}^2 + 1 + C)$ as in Lemma 5, then the sequence $(\mathbf{x}^k, y^k, w^k)$ generated by Algorithm 1 satisfies*

1. $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) \geq \mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+)$.
2. $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ *is lower bounded for all* $k \in \mathbb{N}$ *and converges as* $k \to \infty$.
3. $\{\mathbf{x}^k, y^k, w^k\}$ *is bounded.*

Proof Part 1. It is a direct result of Lemma 4 part 2, and Lemma 5.

Part 2. By the assumption A2, there exists y' such that $\mathbf{A}\mathbf{x}^k + By' = 0$ and $y' = H(By')$. By the assumptions A1–A2, we have

$$f(\mathbf{x}^k) + h(y') \geq \min_{\mathbf{x},y}\{f(\mathbf{x}) + h(y) : \mathbf{A}\mathbf{x} + By = 0\} > -\infty.$$

Then we have

$$\begin{aligned}
\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) &= f(\mathbf{x}^k) + h(y^k) + \langle B^T w^k, y^k - y' \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\
&= f(\mathbf{x}^k) + h(y^k) + \langle \nabla h(y^k), y' - y^k \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\
\text{(Lemma 1, } \nabla h \text{ is Lipschitz)} \quad &\geq f(\mathbf{x}^k) + h(y') + \frac{\beta - L_h \bar{M}^2}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\
&> -\infty.
\end{aligned}$$

Part 3. From parts 1 and 2, $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is upper bounded by $\mathcal{L}_\beta(\mathbf{x}^0, y^0, w^0)$ and so are $f(\mathbf{x}^k) + h(y')$ and $\|\mathbf{A}\mathbf{x}^k + By^k\|^2$. By the assumption A1, $\{\mathbf{x}^k\}$ is bounded and, therefore, $\{By^k\}$ is also bounded. By Lemma 1, we know that $\{y^k\}$ is bounded. By Lemma 3, $\{B^T w^k\}$ is also bounded. Similar to the proof in Lemma 3(b), $w^k - w^0 \in \mathrm{Im}(B)$. Therefore, the boundedness of $B^T w^k$ implies the boundedness of w^k. □

It is important to note that, once β is larger than the threshold, the constants and bounds in Lemmas 5 and 6 only rely on the objective $f(x) + h(y)$, matrices $\mathbf{A}$, B, and the initial point $(\mathbf{x}^0, y^0, w^0)$ but will be *independent of* β, which is essential to the proof of Lemma 9 below.

Lemma 7 (Asymptotic regularity) $\lim_{k\to\infty}\|By^k - By^+\| = 0$ *and* $\lim_{k\to\infty}\|w^k - w^+\| = 0$.

Proof The first result follows directly from Lemmas 4, 5, and 6 (part 2), and the second result from Lemma 3 part (b).

The lemma below corresponds to the assumption A4, part(ii)-b.

Lemma 8 (Boundedness for piecewise linear f_i's) *Consider the case that f_i, $i = 1,\ldots,p$, are piecewise linear. There exist constants $M^* > 0$ (independent of β), $\bar{M}$ and L_g defined in A3 and A4, respectively, for any $\epsilon_0 > 0$, when $\beta > \max\{2(M^*+1)/\epsilon_0^2, L_h\bar{M}^2 + 1 + C\}$, there exists $k_{\rm pl} \in \mathbb{N}$ such that the followings hold for all $k > k_{\rm pl}$:*

1. $\|A_i x_i^+ - A_i x_i^k\| < \epsilon_0$ *and* $\|x_i^+ - x_i^k\| < \bar{M}\epsilon_0$, $i = 0,\ldots,p$;
2. $\|\nabla g(\mathbf{x}^k) - \nabla g(\mathbf{x}^+)\| < (p+1)\bar{M}L_g\epsilon_0$.

Proof Part 1. Since the number K of the linear pieces of f_i is finite for $i = 1,\ldots,p$, ∂f_0 is bounded for x in any bounded set S, and $\{\mathbf{x}^k, y^k, w^k\}$ is bounded (see Lemma 6), $\partial_i f(x_{<i}^+, x_i^+, x_{>i}^k)$ are uniformly bounded for all k and i. Since $\bar{d}_i^k \in \partial_i f(x_{<i}^+, x_i^+, x_{>i}^k)$ (see (12)), the first three terms of r_i (see (14)) are bounded by a universal constant M^* independent of β:

$$f(x_{<i}^+, x_i^k, x_{>i}^k) - f(x_{<i}^+, x_i^+, x_{>i}^k) - \langle \bar{d}_i^k, x_i^k - x_i^+\rangle \in [-M^*, M^*].$$

Hence, as long as $\beta > 2(M^*+1)/\epsilon_0^2$,

$$\|A_i x_i^+ - A_i x_i^k\| \geq \epsilon_0 \Rightarrow \; r_i \geq \frac{\beta}{2}\epsilon_0^2 - M^* > 1 \tag{23}$$

$$\Rightarrow \; \mathcal{L}_\beta(x_{<i}^+, \boldsymbol{x}_i^k, x_{>i}^k, y^k, w^k) - 1 > \mathcal{L}_\beta(x_{<i}^+, \boldsymbol{x}_i^+, x_{>i}^k, y^k, w^k). \tag{24}$$

By Lemmas 4, 5, and 6, this means $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - 1 > \mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+)$. Since $\{\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)\}$ is lower bounded, $\|A_i x_i^+ - A_i x_i^k\| \geq \epsilon_0$ can only hold for finitely many k. Thus for $i = 1,\ldots,p$, we have

$$\|A_i x_i^+ - A_i x_i^k\| < \epsilon_0.$$

As for $i = 0$, because of Lemma 7, we know

$$\limsup_k \|A_0 x_0^+ - A_0 x_0^k\| \leq \limsup_k \|\sum_{i=1}^p (A_i x_i^+ - A_i x_i^k) + By^+ - By^k\| \leq p\epsilon_0.$$

Thus for large $k > k_{pl}$, $\|A_0 x_0^+ - A_0 x_0^k\| \leq (p+1)\epsilon_0$ By Lemma 1, we know Part 1 is correct.
Part 2 follows from $\|\nabla g(\mathbf{x}^k) - \nabla g(\mathbf{x}^+)\| \leq L_g\|\mathbf{x}^k - \mathbf{x}^+\|$, part 1 above, and Lemma 1. □

Lemma 9 (Sufficient descent property P2) *Suppose*

$$\beta > \max\Big\{2(M+1)/\epsilon_0^2, L_h\bar{M}^2 + 1 + C, \sum_{i=1}^p \gamma_i \bar{M}^2 + L_g\bar{M}^2\Big\},$$

where γ_i $(i = 1,\ldots,p)$ and ϵ_0 are constants only depending on f, $M > M^*$ is a constant independent of β. Then, Algorithm 1 satisfies the sufficient descent property P2.

It is worth noting that the proof below will be much simpler if there are only two blocks, instead of $p+2$, or if we assume *prox-regular* functions f_i instead of the less restrictive *restricted prox-regular* functions.

Proof We will show the lower bound (16) for $i = 1, \ldots, p$, which, along with Lemma 4 part 3 and Lemma 5, establishes the sufficient descent property P2.

We shall obtain the lower bound (16) in the backward order $i = p, (p-1), \ldots, 1$. In light of Lemmas 4, 5, and 6, each lower bound (16) for r_i gives us $\|A_i x_i^k - A_i x_i^+\| \to 0$ as $k \to \infty$. We will first show (16) for r_p. Then, after we do the same for $r_{p-1}, \ldots, r_{i+1}$, we will get $\|A_j x_j^k - A_j x_j^+\| \to 0$ for $j = p, p-1, \ldots, i+1$, using which we will get the lower bound (16) for the next r_i. We must take this backward order since ρ_i^k (see (13)) includes the terms $A_j x_j^k - A_j x_j^+$ for $j = p, p-1, \ldots, i+1$.

Our proof for each i is divided into two cases. In Case 1, f_i's are restricted prox-regular (cf. Definition 2), we will get (16) for r_i by validating the condition (15) in Lemma 4 part 4 for f_i. In Case 2, f_i's are piecewise linear (cf. Definition 1), we will show that (15) holds for $\gamma_i = 0$ for $k \geq k_{\text{pl}}$, and following the proof of Lemma 4 part 4, we directly get (16) with $\gamma_i = 0$.

Base step, take $i = p$.
Case 1) f_p is restricted prox-regular. At $i = p$, the inclusion (13) simplifies to

$$d_p^k := -\big(\nabla_p g(\mathbf{x}^+) + A_p^T w^+\big) - \beta A_p^T (By^k - By^+) \in \partial f_p(x_p^+). \tag{25}$$

By Lemma 6 part 3 and the Lipschitz continuity of ∇g, there exists a constant $M > M^*$ (independent of β) such that

$$\|\nabla_p g(\mathbf{x}^+) + A_p^T w^+\| \leq M - 1.$$

By Lemma 7, there exists $k_p \in \mathbb{N}$ such that, for $k > k_p$,

$$\beta \|A_p^T (By^k - By^+)\| \leq 1.$$

Then, we apply the triangle inequality to (25) to obtain

$$\|d_p^k\| \leq \|\nabla_p g(\mathbf{x}^+) + A_p^T w^+\| + \beta \|A_p^T (By^k - By^+)\| \leq M.$$

Use this M to define S_M in Definition 2, which qualifies f_p for (4) and thus validates the assumption in Lemma 4 part 4, proving the lower bound (16) for r_p. As already argued, we get $\lim_{k\to\infty} \|A_p x_p^k - A_p x_p^+\| = 0$.

Case 2): f_i's are piecewise linear (cf. Definition 1). From $\|By^k - By^+\| \to 0$ and $\|w^k - w^+\| \to 0$ (Lemma 7) and $\|\nabla g(\mathbf{x}^k) - \nabla g(\mathbf{x}^+)\| < (p+1)\bar{M} L_g \epsilon_0$ (Lemma 8). In light of (25), $d_p^k \in \partial f_p(x_p^+)$, $d_p^+ \in \partial f_p(x_p^{k+2})$ such that $\|d_p^+ - d_p^k\| < 2(p+1)\bar{M} L_g \epsilon_0$ for all sufficiently large k.

Note that $\epsilon_0 > 0$ can be *arbitrarily* small. Given $d_p^k \in \partial f_p(x_p^+)$ and $d_p^+ \in \partial f_p(x_p^{k+2})$, when the following two properties both hold: (i) $\|d_p^+ - d_p^k\| < 2(p+1)\bar{M} L_g \epsilon_0$ and (ii) $\|x_p^+ - x_p^k\| < \bar{M}\epsilon_0$ (Lemma 8 part 1), we can conclude that x_p^+ and x_p^k belongs to the same $\overline{U}_j$. Suppose $x_p^+ \in \overline{U}_{j_1}$ and $x_p^k \in \overline{U}_{j_2}$. Because of (ii), the polyhedron U_{j_1} is adjacent to the polyhedron U_{j_2} or $j_1 = j_2$. If $\overline{U}_{j_1}$ and $\overline{U}_{j_2}$ are adjacent ($j_1 \neq j_2$) and $a_{j_1} = a_{j_2}$, then we can concatenate $\overline{U}_{j_1}$ and $\overline{U}_{j_2}$ together and all the following analysis carries through. If $\overline{U}_{j_1}$ and $\overline{U}_{j_2}$ are adjacent ($j_1 \neq j_2$) and $a_{j_1} \neq a_{j_2}$, then property (i) is only possible if at least one of x_p^+, x_p^k belongs to their intersection $\overline{U}_{j_1} \cap \overline{U}_{j_2}$ so we can include both points in either $\overline{U}_{j_1}$ or $\overline{U}_{j_2}$, again giving us $j_1 = j_2$. Since $x_p^+, x_p^k \in \overline{U}_{j_1}$ and $d_p^k \in \partial f_p(x_p^+)$, from the convexity of the linear function, we have

$$f_p(x_p^k) - f_p(x_p^+) - \langle d_p^k, x_p^k - x_p^+ \rangle \geq 0,$$

which strengthens the inequality (15) for $i = p$ with $\gamma_p = 0$. By following the proof for Lemma 4 part 4, we get the lower bound (16) for r_p with $\gamma_p = 0$. As already argued, we get $\lim_{k\to\infty} \|A_p x_p^k - A_p x_p^+\| = 0$.

Inductive step, let $i \in \{p-1, \ldots, 1\}$ and make the inductive assumption: $\lim_{k\to\infty} \|A_j x_j^k - A_j x_j^+\| = 0$, $j = p, \ldots, i+1$, which together with $\lim_{k\to\infty} \|By^k - By^+\| = 0$ (Lemma 7) gives $\lim_{k\to\infty} \rho_i^k = 0$ (defined in (13)).

Case 1) f_i is restricted prox-regular. From (13), we have

$$d_i^k = -\big(\nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k) + A_p^T w^+\big) - \beta\rho_i^k \in \partial f_i(x_i^+). \tag{26}$$

Following a similar argument in the case $i = p$ above, there exists $k_i \in \mathbb{N}$ such that, for $k > \max\{k_p, k_{p-1}, \ldots, k_i\}$, we have

$$\|d_i^k\| \le \|\nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k) + A_p^T w^+\| + \beta\|\rho_i^k\| \le M.$$

Use this M to define S_M in Definition 2 for f_i and thus validates the assumption in Lemma 4 part 4 for f_i. Therefore, we get the lower bound (16) for r_i and thus $\lim_k \|A_i x_i^k - A_i x_i^+\| = 0$.

Case 2): f_i's are piecewise linear (cf. Definition 1). The argument is the same as in the base step for case 2, except at its beginning we must use d_i^k in (26) instead of d_p^k in (25). Therefore, we omit this part.

Finally, by combining $r_i \ge \frac{\beta - \gamma_i \bar{M}^2 - L_g \bar{M}^2}{2}\|A_i x_i^k - A_i x_i^+\|^2$, for $i = 1, \ldots, p$, with Lemmas 4 and 5, we establish the sufficient descent property P2.

□

Lemma 10 (Subgradient bound property P3) *Algorithm 1 satisfies Property P3.*

Proof Because $f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^p f_i(x_i)$, we know

$$\partial\mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+) = \left(\left\{\frac{\partial\mathcal{L}_\beta}{\partial x_i}\right\}_{i=1}^p, \nabla_y\mathcal{L}_\beta, \nabla_w\mathcal{L}_\beta\right)(\mathbf{x}^+, y^+, w^+).$$

In order to prove the lemma, we only need to show that each block of $\partial\mathcal{L}_\beta$ can be controlled by some constant depending on β. Therefore, it suffices to prove for $s = 0, \ldots, p$, there exists $d_s \in \frac{\partial\mathcal{L}_\beta}{\partial x_s}(\mathbf{x}^+, y^+, w^+)$ such that

$$\|d_s\| \le (\sigma_{\max}(A_s)\beta + L_h\bar{M} + \sigma_{\max}(A_s)C)\left(\sum_{i=1}^p \|A_i x_i^+ - A_i x_i^k\| + \|By^+ - By^k\|\right), \tag{27}$$

and

$$\|\nabla_w\mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+)\| \le \frac{C}{\beta}\|By^+ - By^k\|, \tag{28}$$

$$\|\nabla_y\mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+)\| \le L_h\bar{M}\|By^+ - By^k\|. \tag{29}$$

In order to prove (28), we have $\nabla_w\mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+) = \mathbf{A}\mathbf{x}^+ + By^+ = \frac{1}{\beta}(w^+ - w^k)$. By Lemma 3,

$$\|\nabla_w\mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+)\| \le \frac{C}{\beta}\|By^+ - By^k\|.$$

In order to prove (29), notice that $\nabla_y \mathcal{L}_\beta(\mathbf{x}^+, y^+, w^+) = B^T(w^+ - w^k)$ and apply Lemma 3. In order to prove (27), observe that for $s = 0, \ldots, p$,

$$\begin{aligned}
&\frac{\partial \mathcal{L}_\beta}{\partial x_s}(\mathbf{x}^+, y^+, w^+) \\
&= \nabla_s g(x^+) + \partial f_s(x_s^+) + A_s^T w^+ + \beta A_s^T \left(Ax^+ + By^+\right) && (30) \\
&= \nabla_s g(x_{\le s}^+, x_{>s}^k) + \partial f_s(x_s^+) + A_s^T w^k + \beta A_s^T \left(A_{\le s} x_{\le s}^+ + A_{>s} x_{>s}^k + By^k\right) && (31) \\
&\quad + A_s^T(w^+ - w^k) + \beta A_s^T \left(A_{>s} x_{>s}^+ - A_{>s} x_{>s}^k + By^+ - By^k\right) + \nabla_s g(x^+) - \nabla_s g(x_{\le s}^+, x_{>s}^k). && (32)
\end{aligned}$$

For the parenthesized term in (31), the first order optimal condition for x_s^+ yields

$$0 \in \nabla_s g(x_{\le s}^+, x_{>s}^k) + \partial f_s(x_s^+) + A_s^T w^k + \beta A_s^T \left(A_{\le s} x_{\le s}^+ + A_{>s} x_{>s}^k + By^k\right).$$

Thus for $s = 0, \ldots, p$, we can have d_s as in (33),

$$\begin{aligned}
d_s &:= \left(A_s^T(w^+ - w^k) + \beta A_s^T \left(A_{>s} x_{>s}^+ - A_{>s} x_{>s}^k + By^+ - By^k\right) + \nabla_s g(x^+) - \nabla_s g(x_{\le s}^+, x_{>s}^k)\right) && (33) \\
&\in \frac{\partial \mathcal{L}_\beta}{\partial x_s}(\mathbf{x}^+, y^+, w^+).
\end{aligned}$$

Note that for any s, x_0^k does not appear in any d_s. Also note that $w^+ - w^k$, $A_{>s} x_{>s}^+ - A_{>s} x_{>s}^k$, $By^+ - By^k$, and $\nabla_s g(x^+) - \nabla_s g(x_{\le s}^+, x_{>s}^k)$ can all be bounded by $\left(\sum_{i=1}^p \|A_i x_i^+ - A_i x_i^k\| + \|By^+ - By^k\|\right)$. Therefore, if we define the largest singular value of A_s to be $\sigma_{\max}(A_s)$, we have the bound for d_s:

$$\|d_s\| \le \left(\sigma_{\max}(A_s)\beta + L_h \bar{M} + \sigma_{\max}(A_s) C\right) \left(\sum_{i=1}^p \|A_i x_i^+ - A_i x_i^k\| + \|By^+ - By^k\|\right).$$

That completes the proof. $\square$

Proof (of Theorem 1) .

Lemmas 5, 9, and 10 establish the properties P1–P3. Theorem 1 follows from Proposition 2. $\square$

4 Discussion

4.1 Tightness of assumptions

In this section, we demonstrate the tightness of the assumptions in Theorem 1 and compare them with related recent works. We only focus on results that do *not* make assumptions on the iterates themselves.

Hong et al. [23] uses $\nabla h(y^k)$ to bound w^k. This inspired our analysis. They studied ADMM for nonconvex consensus and sharing problem. Their formulation is

$$\begin{aligned}
&\underset{x_0, \ldots, x_p, y}{\text{minimize}} \sum_{i=0}^p f_i(x_i) + h(y) \\
&\text{subject to } \sum_{i=0}^p A_i x_i - y = 0.
\end{aligned}$$

where h is Lipschitz differentiable, A_i has full column rank and f_i is Lipschitz differentiable or convex for $i = 0, \ldots, p$. Moreover, $\mathrm{dom}(f_i)$ is required to be a closed bounded set for $i = 0, \ldots, p$.

The boundedness of $\mathrm{dom}(f_i)$ implies the assumption A1. The requirement of A_i for $i = 1, \ldots, p$ and $B = -\mathbf{I}$ implies A2 and A3. Moreover, $f(x_0, \ldots, x_p) = \sum_i f_i(x_i)$, which clearly implies A4. h satisfies A5, too. This shows our theorem could fully cover their case.

Wang et al. [51] studies the so-called Bregman ADMM and includes the standard ADMM as a special case. The following formulation is considered:

$$\begin{aligned} \underset{x_0,\ldots,x_p,y}{\text{minimize}} \sum_{i=0}^{p} f_i(x_i) + h(y) \\ \text{subject to } \sum_{i=0}^{p} A_i x_i + By = 0. \end{aligned}$$

By setting all the auxiliary functions in their algorithm to zero, their assumptions for the standard ADMM reduce to

(a) B is invertible.
(b) h is Lipschitz differentiable and lower bounded. There exists $\beta_0 > 0$ such that $h - \beta_0 \nabla h$ is lower bounded.
(c) $f = \sum_{i=0}^{p} f_i(x_i)$ where f_i, $i = 0, \ldots, p$ is strongly convex.

It is easy to see that (a), (b) and (c) imply assumptions A1 and A3, (a) implies A2, (c) implies A4 and (b) implies A5. Therefore, their assumptions are stronger than ours. We have much more relaxed conditions on f, which can have a coupled Lipschitz differentiable term with separable restricted prox-regular or piecewise linear parts. We also have a simpler assumption on the boundedness without using $h - \nabla h$.

Li and Pong [32] studies ADMM and its proximal version for nonconvex objectives. Their formulation is

$$\begin{aligned} \underset{x_0,y}{\text{minimize}} f_0(x_0) + h(y) \\ \text{subject to } x_0 + By = 0. \end{aligned}$$

Their assumptions for ADMM are

(1) f_0 is lower semi-continuous.
(2) $h \in C^2$ with bounded Hessian matrix $c_2 \mathbf{I} \succeq \nabla^2 h \succeq c_1 \mathbf{I}$ where $c_2 > c_1 > 0$.
(3) B is full row rank.
(4) h is coercive and f_0 is lower bounded.

The assumptions (3) and (4) imply our assumptions A1 and A4, (3) implies A2 and A3, and (2) implies A5. Our assumptions on h and the matrices A, B are more general.

In summary, our convergence conditions for ADMM on nonconvex problems are the most general to the best of our knowledge. It is natural to ask whether our assumptions can be further weakened. We will provide some examples to demonstrate that, while A1, A4 and A3 can probably be further weakened, A5 and A2 are essential in the convergence of nonconvex ADMM and cannot be completely dropped in general. In [10], their divergence example is

$$\underset{x_1,x_2,y}{\text{minimize}} \quad 0 \tag{34a}$$

$$\text{subject to} \begin{pmatrix}1\\1\\1\end{pmatrix} x_1 + \begin{pmatrix}1\\1\\2\end{pmatrix} x_2 + \begin{pmatrix}1\\2\\2\end{pmatrix} y = \begin{pmatrix}0\\0\\0\end{pmatrix}. \tag{34b}$$

Another related example is shown in [32, Example 7].

$$\underset{x_1,x_2,y}{\text{minimize}} \quad \iota_{S_1}(x_1) + \iota_{S_2}(x_2) \tag{35a}$$

$$\text{subject to } x_1 = y \tag{35b}$$

$$x_2 = y, \tag{35c}$$

where $S_1 = \{x = (x_1, x_2) \mid x_2 = 0\}$, $S_2 = \{(0,0), (2,1), (2,-1)\}$. These two examples satisfy A1 and A4-A5 but fail to satisfy A2. Without A2, ADMM is generally incapable to find a feasible point at all, let alone a stationary point. Therefore, A2 is indispensable.

To see the necessity of A5 (the smoothness of h), consider another divergence example

$$\underset{x,y}{\text{minimize}} \quad -|x| + |y| \tag{36a}$$

$$\text{subject to } x = y, \; x \in [-1, 1]. \tag{36b}$$

For any $\beta > 0$, with the initial point $(x^0, y^0, w^0) = (-\frac{2}{\beta}, 0, -1)$, we get the sequence $(x^{2k+1}, y^{2k+1}, w^{2k+1}) = (\frac{2}{\beta}, 0, 1)$ and $(x^{2k}, y^{2k}, w^{2k}) = (-\frac{2}{\beta}, 0, -1)$ for $k \in \mathbb{N}$, which diverges. This problem satisfies all the assumptions except A5, without which w^k cannot be controlled by y^k anymore. Therefore, A5 is also indispensable.

Although the assumptions A2 and A5 seem essential for the convergence of ADMM, other assumptions, especially the assumption A4, might be further relaxed. Moreover, our result requires the y-block to be updated at last right before the update of multiplier. Further studies could be carried out to study the case when a different order is used.

4.2 Primal variables' update order in ADMM

We discuss about the update order of $\{x_i\}_{i=0}^p$ and y in this subsection. Theorem 1 and Theorem 2 apply to the ADMM in which the primal variables $x_0, \ldots, x_p$ are sequentially updated in a fixed order. With minor changes to the proof, both theorems still hold for arbitrary update orders of $x_1, \ldots, x_p$, possibly different between iterations, as long as x_0 is always the first and y is always the last primal variable to update, just before w. In particular, $x_1, \ldots, x_p$ could be updated using random scheme or greedy scheme, which may help avoid low-quality local solutions. Recently, randomized ADMM is considered in papers such as [25, 42, 50]. [50] considered the randomly permuted ADMM for solving linear systems, and proved its convergence in the expectation sense. However, in general, permutation including the last block y could cause ADMM to diverge (A convex example can be found in [62]). Consider

$$\underset{x,y\in\mathbb{R}}{\text{minimize}} \quad x(1+y)$$

$$\text{subject to} \quad x - y = 0.$$

It is easy to check that, if we fix the update order to either x, y, w or y, x, w for all iterations, Algorithm 1 converges. However, if we alternate between the two update orders, we obtain (with $\alpha := 1/\beta$) the diverging sequence $(x^{2k+1}, y^{2k+1}, w^{2k+1}) = (2\alpha(\alpha-1), -\alpha, \alpha-1)$ and $(x^{2k}, y^{2k}, w^{2k}) = (-\alpha, 2\alpha(\alpha-1), -\alpha)$. Another divergent example when primal variables' update order alternates is the following convex and nonsmooth problem:

$$\underset{x,y}{\text{minimize}} \;\; 2|x-1| + |y| \tag{37a}$$

$$\text{subject to } x = y. \tag{37b}$$

4.3 Inexact optimization of subproblems

Note that all subproblems in Algorithm 1 should be solved exactly. This might restrict the wide use of the algorithm in real applications. Thus, the convergence of the inexact version of Algorithm 1 is discussed here. We extend the developed convergence results to the following inexact version of Algorithm 1 under some additional assumptions. More specifically, we assume that the sequence $\{\mathbf{x}^k, y^k, w^k\}$ generated by the inexact version of Algorithm 1 satisfies

P1' **(Boundedness)** $\{\mathbf{x}^k, y^k, w^k\}$ is bounded, and $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded;
P2' **(Sufficient descent)** there is a nonnegative sequence $\{\eta_k\}$ and a constant $C_1 > 0$ such that for all sufficiently large k, we have

$$\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \geq C_1\big(\|B(y^{k+1}-y^k)\|^2 + \sum_{i=1}^{p} \|A_i(x_i^k - x_i^{k+1})\|^2\big) - \eta_k, \tag{38}$$

P3' **(subgradient bound)** and there exists $d^{k+1} \in \partial\mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1})$ such that

$$\|d^{k+1}\| \leq C_2\big(\|B(y^{k+1}-y^k)\| + \sum_{i=1}^{p} \|A_i(x_i^{k+1} - x_i^k)\|\big) + \eta_k. \tag{39}$$

When $\sum_k \eta_k < \infty$, the convergence results in Theorem 1 still hold for this sequence. This is because Proposition 2 still holds when the error is summable. However, when a specific algorithm is applied to solve these subproblems inexactly, it might require some additional conditions, and we leave this in the future work.

5 Applications

In this section, we apply the developed convergence results to several well-known applications. To the best of our knowledge, all the obtained convergence results are novel and cannot be recovered from the previous literature.

A) Statistical learning

Statistical learning models often involve two terms in the objective function. The first term is used to measure the fitting error. The second term is a regularizer to control the model complexity. Generally speaking, it can be written as

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{p} l_i(A_i x - b_i) + r(x), \tag{40}$$

where $A_i \in \mathbb{R}^{m_i\times n}$, $b_i \in \mathbb{R}^{m_i}$ and $x \in \mathbb{R}^n$. Examples of the fitting measure l_i include least squares, logistic functions, and other smooth functions. The regularizers can be some sparsity-inducing functions [2,13,63, 66,67,54,65] such as MCP, SCAD, ℓ_q quasi-norms for $0 < q \le 1$. Take LASSO as an example,

$$\underset{x}{\text{minimize}} \ \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1.$$

The first term $\|y - Ax\|^2$ measures the difference between the linear model Ax and outcome y. The second term $\|x\|_1$ measures the sparsity of x.

In order to solve (40) using ADMM, we reformulate it as

$$\begin{aligned} \underset{x,\{z_i\}_{i=1}^p}{\text{minimize}} \quad & r(x) + \sum_{i=1}^{p} l_i(A_i z_i - b_i), \\ \text{subject to} \quad & x = z_i, \ \forall i = 1, \ldots, p. \end{aligned} \tag{41}$$

Algorithm 2 gives the standard ADMM algorithm for this problem.

Algorithm 2 ADMM for (41)

Denote $\mathbf{z} = [z_1; z_2; \cdots; z_p]$, $\mathbf{w} = [w_1; w_2; \cdots; w_p]$.
  Initialize $x^0, \mathbf{z}^0, \mathbf{w}^0$ arbitrarily;
  while stopping criterion not satisfied **do**
    $x^{k+1} \leftarrow \text{argmin}_x\, r(x) + \frac{\beta}{2}\sum_{i=1}^{p}(z_i^k + \frac{w_i^k}{\beta} - x)^2$;
    for $s = 1, \ldots, p$ **do**
      $z_s^{k+1} \leftarrow \text{argmin}_{z_s}\, l_s(A_s z_s - b_s) + \frac{\beta}{2}(z_s + \frac{w_s^k}{\beta} - x^{k+1})^2$;
      $w_s^{k+1} = w_s^k + \beta(z_s^{k+1} - x^{k+1})$;
    end for
    $k \leftarrow k + 1$;
  end while
  return x^k.

Based on Theorem 1, we have the following corollary.

Corollary 1 *Let $r(x) = \|x\|_q^q = \sum_i |x_i|^q$, $0 < q \le 1$, SCAD, MCP, or any piecewise linear function, if*

i) (Coercivity) $r(x) + \sum_i l_i(A_i x - b_i)$ is coercive;
ii) (Smoothness) For each $i = 1, \ldots, p$, l_i is Lipschitz differentiable.

then for sufficiently large β, the sequence $(x^k, \mathbf{z}^k, \mathbf{w}^k)$generated by Algorithm 2 has limit points and all of its limit points are stationary points of the augmented Lagrangian $\mathcal{L}_\beta$.

Proof Rewrite the optimization to a standard form, we have

$$\underset{x,\{z_i\}_{i=1}^p}{\text{minimize}} \quad r(x) + \sum_{i=1}^{p} l_i(A_i z_i - b_i), \tag{42a}$$

$$\text{subject to} \quad Ex + \mathbf{I}_{np}\mathbf{z} = 0. \tag{42b}$$

where $E = -[\mathbf{I}_n; \ldots; \mathbf{I}_n] \in \mathbb{R}^{np \times n}$, $\mathbf{I}_{np} \in \mathbb{R}^{np \times np}$ is the identity matrix, and $\mathbf{z} = [z_1; \ldots; z_p] \in \mathbb{R}^{np}$. Fitting (42) to the standard form (7), there are two blocks $(x, \mathbf{z})$ and $B = \mathbf{I}_{np}$. $f(x) = r(x)$ and $h(z) = \sum_{i=1}^{p} l_i(A_i z_i - b_i)$.

Now let us check A1–A5. A1 holds because of i). A2 holds because $B = \mathbf{I}_{np}$. A5 holds because of ii). A3 holds because E and $\mathbf{I}_{np}$ both have full column ranks. If $r(x)$ is piecewise linear, then A4 holds naturally. If $r(x)$ is MCP

$$P_{\gamma,\lambda}(x) \triangleq \begin{cases} \lambda|x| - \frac{x^2}{2\lambda}, & \text{if } |x| \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |x| \ge \gamma\lambda \end{cases},$$

or SCAD

$$Q_{\gamma,\lambda}(x) \triangleq \begin{cases} \lambda|x|, & \text{if } |x| \le \lambda \\ \lambda|x| - \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2\gamma - 2}, & \text{if } \lambda < |x| \le \gamma\lambda \\ \frac{1}{2}(\gamma+1)\lambda^2, & \text{if } |x| \ge \gamma\lambda \end{cases}.$$

we can verify that those functions are the maximum of a set of quadratic functions. Then by [43, Example 2.9], we know they are prox-regular. Hence, it remains to verify A4(ii)a that $r(x) = \sum_i |x_i|^q$ is restricted prox-regular. When $q = 1$, this is trivial; when $0 < q < 1$, it has been proved in Proposition 1. This verifies A4 and completes the proof. □

B) Minimization on compact manifolds

Compact manifolds and their projection operators such as spherical manifolds S^{n-1}, Stiefel manifolds (the set of p orthonormal vectors $x_1, \ldots, x_p \in \mathbb{R}^n$, $p \le n$) and Grassmann manifolds (the set of subspaces in $\mathbb{R}^n$ of dimension p) often arise in optimization. Some recent studies and algorithms can be found in [58,31,36]. A simple example is:

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & J(x), \\ \text{subject to} \quad & \|x\|^2 = 1, \end{aligned} \tag{43}$$

More generally, let S be any compact set. We consider the problem

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & J(x), \\ \text{subject to} \quad & x \in S, \end{aligned} \tag{44}$$

which can be rewritten to the following form:

$$\underset{x,y}{\text{minimize}} \quad \iota_S(x) + J(y), \tag{45a}$$

$$\text{subject to} \quad x - y = 0, \tag{45b}$$

Algorithm 3 ADMM for minimization on a compact set (45)

Initialize x^0, y^0, w^0 arbitrarily;
while stopping criterion not satisfied **do**
  $x^{k+1} \leftarrow \text{Proj}_S(y^k - \frac{w^k}{\beta})$;
  $y^{k+1} \leftarrow \text{argmin}_y\, J(y) + \frac{\beta}{2}\|y - \frac{w^k}{\beta} - x^{k+1}\|^2$;
  $w^{k+1} \leftarrow w^k + \beta(y^{k+1} - x^{k+1})$;
  $k \leftarrow k + 1$.
end while
return x^k.

where $\iota_S(\cdot)$ is the indicator function: $\iota_S(x) = 0$ if $x \in S$ or ∞ if $x \notin S$. Applying ADMM to solve this problem, we get Algorithm 3.

Based on Theorem 1, we have the following corollary.

Corollary 2 *If J is Lipschitz differentiable, then for any sufficiently large β, the sequence (x^k, y^k, w^k) generated by Algorithm 3 has at least one limit point, and each limit point is a stationary point of the augmented Lagrangian $\mathcal{L}_\beta$.*

Proof To show this corollary, we shall verify assumptions A1–A5.

The assumption A1 holds because the feasible set is a bounded set and J is lower bounded on the feasible set. A2 and A3 hold because both A and B are identity matrices. A5 holds because J is Lipschitz differentiable. A4 holds because ι_S is lower semi-continuous.

C) Smooth Optimization over complementarity constraints

We consider the following optimization problem over complementarity constraints.

$$\begin{aligned} \underset{\{x,y\}}{\text{minimize}} \quad & h(x,y) \\ \text{subject to} \quad & x^T y = 0, x \geq 0, y \geq 0, \end{aligned} \tag{46}$$

where $h(x,y)$ is a smooth function with Lipschitz differentiable gradient. The considered problem is a special case of the mathematical programming with equilibrium constraints (MPEC) [28], and includes the linear complementarity problem (LCP) [12] as a special case. In order to apply the ADMM algorithm to solve this problem, we introduce two auxiliary variables $x', y' \in \mathbb{R}^n$ and define the complementarity set $S \triangleq \{(x,y) : x^T y = 0,\ x \geq 0,\ y \geq 0\}$. With these notations, problem (46) can be reformulated as follows

$$\begin{aligned} \underset{\{x',y',x,y\}}{\text{minimize}} \quad & \iota_S(x',y') + h(x,y) \\ \text{subject to} \quad & x' - x = 0, \quad y' - y = 0, \end{aligned} \tag{47}$$

where $\iota_S(x', y')$ denotes the indicator function of the set S. Furthermore, let $\mathbf{x}_0 = \begin{pmatrix} x' \\ y' \end{pmatrix}$ and the second block $\mathbf{y} = \begin{pmatrix} x \\ y \end{pmatrix}$. Then (47) becomes

$$\begin{aligned} \underset{\mathbf{x}_0, \mathbf{y}}{\text{minimize}} \quad & \iota_S(\mathbf{x}_0) + h(\mathbf{y}) \\ \text{subject to} \quad & \mathbf{x}_0 - \mathbf{y} = 0. \end{aligned} \tag{48}$$

Corollary 3 *Assume that h is Lipschitz differentiable and coercive over the complementarity set, then for sufficiently large β, the sequence $(\mathbf{x}_0^k, \mathbf{y}^k, w^k)$ generated by Algorithm ADMM applied to* (48) *has limit points and all of its limit points are stationary points of the augmented Lagrangian $\mathcal{L}_\beta$.*

Proof In order to prove this corollary, we only need to verify assumptions **A**1-**A**5. **A**1 holds for the coercivity of h over S and the specific form of ι_S. **A**2 is obvious due to in this case $A = \mathbf{I}$ and $B = -\mathbf{I}$. **A**3 holds for both $\mathbf{I}$ and $-\mathbf{I}$ being full column rank. **A**4 can be satisfied by setting $f_0 = \iota_S$ and $g \equiv h$. **A**5 holds due to the Lipschitz differentiability of h. Thus, according to Theorem 1, we complete the proof. □

D) Matrix decomposition

ADMM has also been applied to solve matrix related problems, such as sparse principle component analysis (PCA) [24], matrix decomposition [48,53], matrix completion [7], matrix recovery [41], non-negative matrix factorization [61,49] and background/foreground extraction [8,63].

In the following, we take the video surveillance image-flow problem as an example. A video can be formulated as a matrix V where each column is a vectorized image of a video frame. It can be generally decomposed into three parts, background, foreground, and noise. The background has low rank since it does not move. The derivative of the foreground is small because foreground (such as human beings, other moving objectives) moves relatively slowly. The noise is generally assumed to be Gaussian and thus can be modeled via Frobenius norm.

More specifically, consider the following matrix decomposition model:

$$\underset{X,Y,Z}{\text{minimize}} \quad p(X) + \sum_{i=1}^{m-1} \|Y_i - Y_{i+1}\| + \|Z\|_F^2, \tag{49}$$

$$\text{subject to} \quad V = X + Y + Z, \tag{50}$$

where $X, Y, Z, V \in \mathbb{R}^{n\times m}$, Y_i is the ith column of Y, $\|\cdot\|_F$ is the Frobenius norm, and $p(X)$ is any lower bounded lower semi-continuous penalty function, for example, the Schatten-q quasi-norm $\|X\|_q$ $(0 < q \le 1)$:

$$\|A\|_q = \sum_{i=1}^{n} \sigma_i^q(A),$$

where $\sigma_i(A)$ is the ith largest singular value of A.

The corresponding ADMM algorithm is given in Algorithm 4.

Corollary 4 *For a sufficiently large β, the sequence (X^k, Y^k, Z^k, W^k) generated by Algorithm 4 has at least one limit point, and each limit point is a stationary point of the augmented Lagrangian function $\mathcal{L}_\beta$.*

Algorithm 4 ADMM for (49)

Initialize Y^0, Z^0, W^0 arbitrarily;
while stopping criteria not satisfied **do**
  $X^{k+1} \leftarrow \mathrm{argmin}_X \, p(X) + \frac{\beta}{2}\|X + Y^k + Z^k - V + W^k/\beta\|_F^2$;
  $Y^{k+1} \leftarrow \mathrm{argmin}_Y \sum_{i=1}^m \|Y_i - Y_j\| + \frac{\beta}{2}\|X^{k+1} + Y + Z^k - V + W^k/\beta\|_F^2$;
  $Z^{k+1} \leftarrow \mathrm{argmin}_Z \, \|Z\|_F^2 + \frac{\beta}{2}\|X^{k+1} + Y^{k+1} + Z - V + W^k/\beta\|_F^2$;
  $W^{k+1} \leftarrow W^k + \beta(X^{k+1} + Y^{k+1} + Z^{k+1} - V)$;
  $k \leftarrow k+1$;
end while
return X^k, Y^k, Z^k.

Proof Let us verify assumptions A1–A5. The assumption A1 holds because of the coercivity of $\|\cdot\|_F$ and $\|\cdot\|_q$. A2 and A3 hold because all the coefficient matrices are identity matrices. A5 holds because $\|\cdot\|_F^2$ is Lipschitz differentiable. A4 holds because p is lower semi-continuous.

6 Conclusion

This paper studied the convergence of ADMM, in its multi-block and original cyclic update form, for nonconvex and nonsmooth optimization. The objective can be certain nonconvex and nonsmooth functions while the constraints are coupled linear equalities. Our results theoretically demonstrate that ADMM, as a variant of ALM, may converge under weaker conditions than ALM. While ALM generally requires the objective function to be smooth, ADMM only requires it to have a smooth part $h(y)$ while the remaining part $f(\mathbf{x})$ can be coupled, nonconvex, and include separable nonsmooth functions and indicator functions of constraint sets.

Our results relax the previous assumptions (e.g., semi-convexity) and allow the nonconvex functions such as ℓ_q quasi-norm ($0 < q < 1$), Schatten-q quasi-norm, SCAD, and others that often appear in sparse optimization. They also allow nonconvex constraint sets such as unit spheres, matrix manifolds, and complementarity constraints.

The underlying proof technique identifies an exclusion set where the sequence does not enter after finitely many iterations. We also manage to have a very general first block x_0. We show that while the middle p blocks $x_1, \ldots, x_p$ can be updated in an arbitrary order for different iterations, the first block x_0 should be updated at first and the last block y at last; otherwise, the concerned iterates may diverge according to the existing example.

Our results can be applied to problems in matrix decomposition, sparse recovery, machine learning, and optimization on compact smooth manifolds and lead to novel convergence guarantees.

Acknowledgements

We would like to thank Drs. Wei Shi, Ting Kei Pong, and Qing Ling for their insightful comments, and Drs. Xin Liu and Yangyang Xu for helpful discussions.

Appendix

Proof **Proof of Proposition 1**

The fact that convex functions and the C^1 regular functions are prox-regular has been proved in the previous literature, for example, see [43]. Here we only prove the second part of the proposition.

(1) For functions $r(x) = \sum_i |x_i|^q$ where $0 < q < 1$, the set of general subgradient of $r(\cdot)$ is

$$\partial r(x) = \left\{ d = [d_1; \ldots; d_n] \,\middle|\, d_i = q \cdot \mathrm{sign}(x_i)|x_i|^{q-1} \text{ if } x_i \neq 0;\ d_i \in \mathbb{R} \text{ if } x_i = 0 \right\}.$$

For any two positive constants $C > 0$ and $M > 1$, take $\gamma = \max\left\{\frac{4(nC^q+MC)}{c^2}, q(1-q)c^{q-2}\right\}$, where $c \triangleq \frac{1}{3}(\frac{q}{M})^{\frac{1}{1-q}}$. The exclusion set S_M contains the set $\{x | \min_{x_i \neq 0} |x_i| \leq 3c\}$. For any point $z \in \mathbb{B}(0, C)/S_M$ and $y \in \mathbb{B}(0, C)$, if $\|z - y\| \leq c$, then $\mathrm{supp}(z) \subset \mathrm{supp}(y)$ and $\|z\|_0 \leq \|y\|_0$, where $\mathbb{B}(0, C) \triangleq \{x | \|x\| < C\}$, $\mathrm{supp}(z)$ denotes the index set of all non-zero elements of z and $\|z\|_0$ denotes the cardinality of $\mathrm{supp}(z)$. Define

$$y'_i = \begin{cases} y_i & i \in \mathrm{supp}(z) \\ 0 & i \notin \mathrm{supp}(z) \end{cases},\ i = 1, \ldots, p.$$

Then for any $d \in \partial r(z)$, the following line of proof holds,

$$\begin{aligned} \|y\|_q^q - \|z\|_q^q - \langle d, y - z \rangle &\overset{(a)}{\geq} \|y'\|_q^q - \|z\|_q^q - \langle d, y' - z \rangle \\ &\overset{(b)}{\geq} -\frac{q(1-q)}{2} c^{q-2} \|z - y'\|^2 \\ &\overset{(c)}{\geq} -\frac{q(1-q)}{2} c^{q-2} \|z - y\|^2, \end{aligned} \tag{51}$$

where (a) holds for $\|y\|_q^q = \|y'\|_q^q + \|y - y'\|_q^q$ by the definition of y', (b) holds because $r(x)$ is twice differentiable along the line segment connecting z and y', and the second order derivative is no bigger than $q(1-q)c^{q-2}$, and (c) holds because $\|z - y\| \geq \|z - y'\|$. While if $\|z - y\| > c$, then for any $d \in \partial r(z)$, we have

$$\|y\|_q^q - \|z\|_q^q - \langle d, y - z \rangle \geq -(2nC^q + 2MC) \geq -\frac{2nC^q + 2MC}{c^2} \|y - z\|^2. \tag{52}$$

Combining (51) and (52) yields the result.

(2) We are going to verify that Schatten-q quasi-norm $\|\cdot\|_q$ is restricted prox-regular. Without loss of generality, suppose $A \in \mathbb{R}^{n \times n}$ is a square matrix.

Suppose the singular value decomposition (SVD) of A is

$$A = U \Sigma V^T = [U_1, U_2] \cdot \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}, \tag{53}$$

where $U, V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma_1 \in \mathbb{R}^{K \times K}$ is diagonal whose diagonal elements are $\sigma_i(A)$, $i = 1, \ldots, K$. Then the general subgradient of $\|A\|_q^q$ [55] is

$$\partial \|A\|_q^q = U_1 D V_1^T + \{U_2 \Gamma V_2^T \,|\, \Gamma \text{ is an arbitrary matrix}\},$$

where $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix whose ith diagonal element is $d_i = q\sigma_i(A)^{q-1}$.

Now we are going to prove $\|\cdot\|_q^q$ is restricted prox-regular, i.e., for any positive parameters $M, P > 0$, there exists $\gamma > 0$ such that for any $\|B\|_F < P$, $\|A\|_F < P$, $A \notin S_M = \{A | \forall X \in \partial\|A\|_q^q,\ \|X\|_F > M\}$, and $T = U_1 D V_1^T + U_2 \Gamma V_2^T \in \partial\|A\|_q^q, \|T\|_F \leq M$, we hope to show

$$\|B\|_q^q - \|A\|_q^q - \langle T, B - A\rangle \geq -\frac{\gamma}{2}\|A - B\|_F^2. \tag{54}$$

Let $\epsilon_0 = \frac{1}{3}(M/q)^{1/(q-1)}$. If $\|B - A\| > \epsilon_0$, we have

$$\|B\|_q^q - \|A\|_q^q - \langle T, B - A\rangle \geq - n^2 P^q - M \cdot \|B - A\|_F \geq -(M\epsilon_0^{-1} + n^2 P^q \epsilon_0^{-2})\|A - B\|_F^2. \tag{55}$$

If $\|B - A\|_F < \epsilon_0$, consider the decomposition of $B = U_B \Sigma^B V_B^T = B_1 + B_2$ where $B_1 = U_B \Sigma_1^B V_B^T$, Σ_1^B is the diagonal matrix preserving elements of Σ^B bigger than $\frac{1}{3}(M/q)^{1/(q-1)}$, and $B_2 = U_B \Sigma_2^B V_B^T$ where $\Sigma_2^B = \Sigma^B - \Sigma_1^B$.

Define a set $S' \triangleq \{T \in \mathbb{R}^{n\times n} | \|T\|_F \leq P,\ \min_{\sigma_i > 0} \sigma_i(T) \geq \epsilon_0\}$. Let's prove $A, B_1 \in S'$. If $\min_{\sigma_i>0} \sigma_i(A) < (M/q)^{1/(q-1)}$, then for any $X \in \partial\|A\|_q^q$, $X = U_1 D V_1^T + U_2 \Gamma V_2^T$ and

$$\|X\|_F \geq \|U_1 D V_1^T\|_F \geq \min_{\sigma_i > 0} q\sigma_i^{q-1} \geq M,$$

which contradicts with the face that $A \notin S_M$. As for B_1, because of $\|A - B\|_F \leq \epsilon_0$ and $\min_{\sigma_i>0} \sigma_i(A) < (M/q)^{1/(q-1)}$, using Weyl inequalities will we get $B_1 \in S'$.

Define the function $F : S' \subset \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, for $A = U_1 \Sigma V_1^T$,

$$F(A) = U_1 D V_1^T,$$

where $D = \mathrm{diag}(q\sigma_1(A)^{q-1}, \ldots, q\sigma_1(A)^{q-1})$, $(0^{q-1} = 0)$. Based on [17, Theorem 4.1] and the compactness of S', $F(A)$ is Lipschitz continuous in S', i.e., there exists $L > 0$, for any two matrices $A, B \in S'$, $\|F(A) - F(B)\|_F \leq L\|A - B\|_F$. This implies

$$\|B_1\|_q^q - \|A\|_q^q - \langle U_1 D V_1^T, B_1 - A\rangle \geq -\frac{L}{2}\|B_1 - A\|_F^2. \tag{56}$$

In addition, because $\|U_2^T U_B\|_F < \|B_1 - A\|_F/\epsilon_0$ and $\|V_2^T V_B\|_F < \|B_1 - A\|_F/\epsilon_0$ (see [33]),

$$\langle U_2 \Gamma V_2^T, B_1 - A\rangle = \langle \Gamma, U_2^T U_B \Sigma_B V_B^T V_2\rangle \geq -\frac{M^2}{\epsilon_0^2}\|B_1 - A\|_F^2. \tag{57}$$

Furthermore, $\|B_2\|_q^q - \langle T, B_2\rangle \geq 0$ and $\|B_1 - A\|_F \leq \|B - A\|_F + \|B - B_1\|_F \leq 2\|B - A\|_F$, together with (56) and (57) we have

$$\begin{aligned}\|B\|_q^q - \|A\|_q^q - \langle T, B - A\rangle &= \|B_1\|_q^q - \|A\|_q^q - \langle T, B_1 - A\rangle + \|B_2\|_q^q - \langle T, B_2\rangle \\ &\geq -\left(\frac{L}{2} + \frac{M^2}{\epsilon_0^2}\right)\|B_1 - A\|_F^2 \geq -\left(2L + \frac{4M^2}{\epsilon_0^2}\right)\|B - A\|_F^2.\end{aligned} \tag{58}$$

Combining (55) and (58), we finally prove (54) with appropriate γ.

(3) We need to show that the indicator function ι_S of a p-dimensional compact C^2 manifold S is restricted prox-regular. First of all, by definition, the exclusion set S_M of ι_S is empty for any $M > 0$. Since S is compact and C^2, there are a series of C^2 homeomorphisms $h_\eta : \mathbb{R}^p \mapsto \mathbb{R}^n$, $\eta \in \{1, \ldots, m\}$ and $\delta > 0$ such that for any

x, there exist an η and an α_x satisfying $x = h_\eta(\alpha_x) \in S$. Furthermore, for any $\|y - x\| \leq \delta$, we can find an α_y satisfying $y = h_\eta(\alpha_y)$.

Note that $\partial \iota_S(x) = \mathrm{Im}(J_{h_\eta}(x))^\perp$, where J_{h_η} is the Jacobian of h_η. For any $d \in \partial\iota_S(x)$, $\|d\| \leq M$ and $\|x - y\| \leq \delta$,

$$\begin{aligned}
\iota_S(y) - \iota_S(x) - \langle d, y - x\rangle &= -\langle d, h_\eta(\alpha_y) - h_\eta(\alpha_x)\rangle \\
&= -\langle d, h_\eta(\alpha_y) - h_\eta(\alpha_x) - J_{h_\eta}(\alpha_y - \alpha_x)\rangle \\
&\geq -\|d\| \cdot \gamma \|\alpha_y - \alpha_x\|^2 \\
&\geq -M\gamma C^2 \|x - y\|^2, \qquad (59)
\end{aligned}$$

where γ and C are the Lipschitz constants of ∇h_η and h_η^{-1}, respectively. For any $\|y - x\| \geq \delta$,

$$\begin{aligned}
\iota_S(y) - \iota_S(x) - \langle d, y - x\rangle &= -\langle d, y - x\rangle \\
&\geq -\|d\| \cdot \|y - x\| \\
&\geq -\frac{M}{\delta}\|y - x\|^2, \qquad (60)
\end{aligned}$$

where M is the maximum of $\|d\|$ over $\partial\iota_S(x)$. Combining (59) and (60) shows that ι_S is restricted prox-regular.

*Proof (**Lemma 1**)* By the definitions of H in A3(a) and y^k, we have $y^k = H(By^k)$. Therefore, $\|y^{k_1} - y^{k_2}\| = \|H(By^{k_1}) - H(By^{k_2})\| \leq \bar{M}\|By^{k_1} - By^{k_2}\|$. Similarly, by the optimality of x_i^k, we have $x_i^k = F_i(A_i x_i^k)$. Therefore, $\|x_i^{k_1} - x_i^{k_2}\| = \|F_i(A_i x_i^{k_1}) - F_i(A_i x_i^{k_2})\| \leq \bar{M}\|A_i x_i^{k_1} - A_i x_i^{k_2}\|$. □

*Proof (**Lemma 2**)* Let us first show that the y-subproblem is well defined. To begin with, we will show that $h(y)$ is lower bounded by a quadratic function of By:

$$h(y) \geq h(H(0)) - \left(\bar{M}\|\nabla h(H(0))\|\right) \cdot \|By\| - \frac{L_h \bar{M}^2}{2}\|By\|^2.$$

By A3, we know $h(y)$ is lower bounded by $h(H(By))$:

$$h(y) \geq h(H(By)).$$

Because of A5 and A3, $h(H(By))$ is lower bounded by a quadratic function of By:

$$h(H(By)) \geq h(H(0)) + \langle \nabla h(H(0)), H(By) - H(0)\rangle - \frac{L_h}{2}\|H(By) - H(0)\|^2 \qquad (61)$$

$$\geq h(H(0)) - \|\nabla h(H(0))\| \cdot \bar{M} \cdot \|By\| - \frac{L_h \bar{M}^2}{2}\|By\|^2 \qquad (62)$$

Therefore $h(y)$ is also bounded by the quadratic function:

$$h(y) \geq h(H(0)) - \|\nabla h(H(0))\| \cdot \bar{M} \cdot \|By\| - \frac{L_h \bar{M}^2}{2}\|By\|^2.$$

Recall that y-subproblem is to minimize the Lagrangian function w.r.t. y, by neglecting other constants, it is equivalent to minimize:

$$\operatorname*{argmin}_y P(y) := h(y) + \langle w^k + \beta \mathbf{A}\mathbf{x}^+, By\rangle + \frac{\beta}{2}\|By\|^2. \qquad (63)$$

Because $h(y)$ is lower bounded by $-\frac{L_h \bar{M}^2}{2}\|By\|^2$, when $\beta > L_h \bar{M}$, $P(y) \to \infty$ as $\|By\| \to \infty$. This shows that y-subproblem is coercive with respect to By. Because $P(y)$ is lower semi-continuous and $\operatorname{argmin}_y h(y)$ s.t. $By = u$ has a unique solution for each u, the minimal point of $P(y)$ must exist and the y-subproblem is well defined.

As for the x_i-subproblem, $i = 0, \ldots, p$, ignoring the constants yields

$$\begin{aligned}
&\underset{x_i}{\operatorname{argmin}}\ \mathcal{L}_\beta(x^+_{<i}, x_i, x^k_{>i}, y^k, w^k) \\
&= \underset{x_i}{\operatorname{argmin}}\ f(x^+_{<i}, x_i, x^k_{>i}) + \frac{\beta}{2}\|\frac{1}{\beta}w^k + A_{<i}x^+_{<i} + A_{>i}x^k_{>i} + A_i x_i + By^k\|^2 \\
&= \underset{x_i}{\operatorname{argmin}}\ f(x^+_{<i}, x_i, x^k_{>i}) + h(u) - h(u) + \frac{\beta}{2}\|Bu - By^k - \frac{1}{\beta}w^k\|^2.
\end{aligned}$$

where $u = H(-A_{<i}x^+_{<i} - A_{>i}x^k_{>i} - A_i x_i)$. The first two terms are coercive bounded because $A_{<i}x^+_{<i} + A_{>i}x^k_{>i} + A_i x_i + Bu = 0$ and A1. The third and fourth terms are lower bounded because h is Lipschitz differentiable. Because the function is lower semi-continuous, all the subproblems are well defined. □

*Proof (**Proposition 1**)* Define the augmented Lagrangian function to be

$$\mathcal{L}_\beta(x, y, w) = x^2 - y^2 + w(x - y) + \beta\|x - y\|^2.$$

It is clear that when $\beta = 0$, $\mathcal{L}_\beta$ is not lower bounded for any w. We are going to show that for any $\beta > 1$, the duality gap is not zero.

$$\inf_{x\in[-1,1], y\in\mathbb{R}} \sup_{w\in\mathbb{R}} \mathcal{L}_\beta(x, y, w) > \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1], y\in\mathbb{R}} \mathcal{L}_\beta(x, y, w).$$

On one hand, because $\sup_{w\in\mathbb{R}} \mathcal{L}_\beta(x, y, w) = +\infty$ when $x \neq y$ and $\sup_{w\in\mathbb{R}} \mathcal{L}_\beta(x, y, w) = 0$ when $x = y$, we have

$$\inf_{x\in[-1,1], y\in\mathbb{R}} \sup_{w\in\mathbb{R}} \mathcal{L}_\beta(x, y, w) = 0.$$

On the other hand, let $t = x - y$,

$$\sup_{w\in\mathbb{R}} \inf_{x\in[-1,1], y\in\mathbb{R}} \mathcal{L}_\beta(x, y, w) = \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1], t\in\mathbb{R}} t(2x - t) + wt + \beta t^2 = \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1], t\in\mathbb{R}} (w + 2x)t + (\beta - 1)t^2 \tag{64}$$

$$= \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1]} -\frac{(w + 2x)^2}{4(\beta - 1)} = \sup_{w\in\mathbb{R}} -\frac{\max\{(w - 2)^2, (w + 2)^2\}}{4(\beta - 1)} = -\frac{1}{\beta - 1}. \tag{65}$$

This shows the duality gap is not zero (but it goes to 0 as β tends to ∞).

Then let us show that ALM does not converge if β^k is bounded, i.e., there exists $\beta > 0$ such that $\beta^k \leq \beta$ for any $k \in \mathbb{N}$. Without loss of generality, we assume that β^k equals to the constant β for all $k \in \mathbb{N}$. This will not affect the proof. ALM consists of two steps

1) $(x^{k+1}, y^{k+1}) = \operatorname{argmin}_{x,y}\mathcal{L}_\beta(x, y, w^k)$,
2) $w^{k+1} = w^k + \tau(x^{k+1} - y^{k+1})$.

Since $(x^{k+1}-y^{k+1}) \in \partial\psi(w^k)$ where $\psi(w) = \inf_{x,y}\mathcal{L}_\beta(x,y,w)$, and we already know

$$\inf_{x,y}\mathcal{L}_\beta(x,y,w) = -\frac{\max((w-2)^2,(w+2)^2)}{4(\beta-1)},$$

we have

$$w^{k+1} = \begin{cases} (1-\frac{\tau}{2(\beta-1)})w^k - \frac{\tau}{\beta-1} & \text{if } w^k \geq 0 \\ (1-\frac{\tau}{2(\beta-1)})w^k + \frac{\tau}{\beta-1} & \text{if } w^k \leq 0 \end{cases}.$$

Note that when $w^k = 0$, the optimization problem $\inf_{x,y}\mathcal{L}_\beta(x,y,0)$ has two distinct minimal points which lead to two different values. This shows no matter how small τ is, w^k will oscillate around 0 and never converge.

However, although the duality gap is not zero, ADMM still converges in this case. There are two ways to prove it. The first way is to check all the conditions in Theorem 1. Another way is to check the iterates directly. The ADMM iterates are

$$x^{k+1} = \max\left(-1, \min(1, \frac{\beta}{\beta+1}(y^k - \frac{w^k}{2\beta}))\right), \quad y^{k+1} = \frac{\beta}{\beta-1}(x^{k+1} + \frac{w^k}{2\beta}), \quad w^{k+1} = w^k + 2\beta(x^{k+1}-y^{k+1}). \tag{66}$$

The second equality shows that $w^k = -2y^k$, substituting it into the first and second equalities, we have

$$x^{k+1} = \max\{-1, \min\{1, y^k\}\}, \quad y^{k+1} = \frac{1}{\beta-1}\left(\beta x^{k+1} - y^k\right). \tag{67}$$

Here $|y^{k+1}| \leq \frac{\beta}{\beta-1} + \frac{1}{\beta-1}|y^k|$. Thus after finite iterations, $|y^k| \leq 2$ (assume $\beta > 2$). If $|y^k| \leq 1$, the ADMM sequence converges obviously. If $|y^k| > 1$, without loss of generality, we could assume $2 > y^k > 1$. Then $x^{k+1} = 1$. It means $0 < y^{k+1} < 1$, so the ADMM sequence converges. Thus, we know for any initial point y^0 and w^0, ADMM converges.

Proof ***(Theorem 2)*** Similar to the proof of Theorem 1, we only need to verify P1-P3 in Proposition 2.

Proof of P2: Similar to Lemma 4 and Lemma 5, we have

$$\begin{aligned} &\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \\ \geq & -\frac{1}{\beta}\|w^k - w^{k+1}\|^2 + \sum_{i=1}^{p}\frac{\beta - L_\phi\bar{M}}{2}\|A_i x_i^k - A_i x_i^{k+1}\|^2 + \frac{\beta - L_\phi\bar{M}}{2}\|By^k - By^{k+1}\|^2. \end{aligned} \tag{68}$$

Since $B^T w^k = -\partial_y\phi(\mathbf{x}^k, y^k)$ for any $k \in \mathbb{N}$, we have

$$\|w^k - w^{k+1}\| \leq C_1 L_\phi\left(\sum_{i=0}^{p}\|x_i^k - x_i^{k+1}\| + \|y^k - y^{k+1}\|\right),$$

where $C_1 = \sigma_{\min}(B)$, $\sigma_{\min}(B)$ is the smallest positive singular value of B, and L_ϕ is the Lipschitz constant for ϕ. Therefore, we have

$$\begin{aligned} &\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) - \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \\ \geq &\left(\frac{\beta - L_\phi\bar{M}}{2} - \frac{CL_\phi\bar{M}}{\beta}\right)\sum_{i=0}^{p}\|A_i x_i^k - A_i x_i^{k+1}\|^2 + \left(\frac{\beta - L_\phi\bar{M}}{2} - \frac{C_1 L_\phi\bar{M}}{\beta}\right)\|By^k - By^{k+1}\|^2. \end{aligned} \tag{69}$$

When $\beta > \max\{1, L_\phi \bar{M} + 2C_1 L_\phi \bar{M}\}$, P2 holds.

Proof of P1: First of all, we have already shown $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) \geq \mathcal{L}_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1})$, which means $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ decreases monotonically. There exists y' such that $\mathbf{A}\mathbf{x}^k + By' = 0$ and $y' = H(By')$. In order to show $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded, we apply A1-A3 to get

$$\begin{aligned}\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k) &= \phi(\mathbf{x}^k, y^k) + \left\langle w^k, \sum_{i=0}^{p} A_i x_i^k + By^k \right\rangle + \frac{\beta}{2}\|\sum_{i=0}^{p} A_i x_i^k + By^k\|^2 \\ &= \phi(\mathbf{x}^k, y^k) + \left\langle d_y^k, y' - y^k \right\rangle + \frac{\beta}{2}\|By^k - By'\|^2 \geq \phi(\mathbf{x}^k, y') + \frac{\beta}{4}\|\sum_{i=0}^{p} A_i x_i^k + By^k\|^2 > -\infty,\end{aligned} \tag{70}$$

for some $d_y^k \in \partial_y \phi(\mathbf{x}^k, y^k)$. This shows that $\mathcal{L}_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded. If we view (70) from the opposite direction, it can be observed that

$$\phi(\mathbf{x}^k, y') + \frac{\beta}{4}\|\sum_{i=1}^{p} A_i x_i^k + By^k\|^2$$

is upper bounded by $\mathcal{L}_\beta(\mathbf{x}^0, y^0, w^0)$. Then A1 ensures that $\{\mathbf{x}^k, y^k\}$ is bounded. Therefore, w^k is bounded too.

Proof of P3: This part is trivial as ϕ is Lipschitz differentiable. Hence we omit it.

References

1. Attouch, H., Bolte, J., Svaiter, B.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming **137**(1-2), 91–129 (2013)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning **4**(1), 1–106 (2012)
3. Bertsekas, D.P.: Constrained optimization and Lagrange multiplier methods. Academic press (2014)
4. Birgin, E.G., Martínez, J.M.: Practical augmented Lagrangian methods for constrained optimization, vol. 10. SIAM (2014)
5. Bolte, J., Daniilidis, A., Lewis, A.: The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization **17**(4), 1205–1223 (2007)
6. Bouaziz, S., Tagliasacchi, A., Pauly, M.: Sparse iterative closest point. In: Computer graphics forum, vol. 32, pp. 113–123. Wiley Online Library (2013)
7. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization **20**(4), 1956–1982 (2010)
8. Chartrand, R.: Nonconvex splitting for regularized low-rank+ sparse decomposition. Signal Processing, IEEE Transactions on **60**(11), 5810–5819 (2012)
9. Chartrand, R., Wohlberg, B.: A nonconvex admm algorithm for group sparsity with sparse groups. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 6009–6013. IEEE (2013)
10. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Mathematical Programming pp. 1–23 (2014)
11. Conn, A.R., Gould, N.I., Toint, P.: A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. SIAM Journal on Numerical Analysis **28**(2), 545–572 (1991)
12. Cottle, R., and Dantzig, G.,: Complementary pivot theory of mathematical programming. Linear Algebra and its Applications **1**, 103–125 (1968)
13. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics **63**(1), 1–38 (2010)
14. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. In (Glowinski, R., Osher, S., Yin, W. ed.) Splitting Methods in Communication, Imaging, Science and Engineering. Springer, New York, (2016)

15. Davis, D., Yin, W.: Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. Mathematics of Operations Research **42**(3), 783 - 805, (2017)
16. Deng, W., Lai, M.J., Peng, Z., Yin, W.: Parallel multi-block ADMM with $o(1/k)$ convergence. Journal of Scientific Computing **71**(2), 712 - 736, (2017)
17. Ding, C., Sun, D., Sun, J., Toh, K.C.: Spectral operators of matrices. Mathematical Programming, Ser. B, 1-23 (2017) DOI 10.1007/s10107-017-1162-3
18. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & Mathematics with Applications **2**(1), 17 – 40 (1976)
19. Glowinski, R.: Numerical methods for nonlinear variational problems. Springer series in computational physics. Springer-Verlag, New York (1984)
20. Glowinski, R., Marroco, A.: On the approximation by finite elements of order one, and resolution, penalisation-duality for a class of nonlinear dirichlet problems. ESAIM: Mathematical Modelling and Numerical Analysis - Mathematical Modelling and Numerical Analysis **9**(R2), 41–76 (1975)
21. He, B., Yuan, X.: On the $o(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM Journal on Numerical Analysis **50**(2), 700–709 (2012)
22. Hestenes, M.R.: Multiplier and gradient methods. Journal of optimization theory and applications **4**(5), 303–320 (1969)
23. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. SIAM Journal on Optimization, 26(1), 337-364, (2016)
24. Hu, Y., Chi, E., Allen, G.I.: ADMM algorithmic regularization paths for sparse statistical machine learning. In (Glowinski, R., Osher, S., Yin, W. ed.) Splitting Methods in Communication, Imaging, Science and Engineering. Springer, New York, (2016)
25. Iutzeler, F., Bianchi, P., Ciblat, P., Hachem, W.:Asynchronous Distributed Optimization using a Randomized Alternating Direction Method of Multipliers. In 52nd IEEE Conference on Decision and Control, 3671-3676 (2013)
26. Ivanov, M., Zlateva, N.: Abstract subdifferential calculus and semi-convex functions. Serdica Mathematical Journal **23**(1), 35p–58p (1997)
27. Jiang, B., Ma, S., Zhang, S.: Alternating direction method of multipliers for real and complex polynomial optimization models. Optimization **63**(6), 883–898 (2014)
28. Jiang S., Zhang J., Chen C., G. Lin: Smoothing partial exact penalty splitting method for mathematical programs with equilibrium constraints. Journal of Global Optimization, (2017) DOI 10.1007/s10898-017-0539-4
29. Knopp, K.: Infinite sequences and series. Courier Corporation (1956)
30. Kryštof, V., Zajíček, L.: Differences of two semiconvex functions on the real line. preprint (2015)
31. Lai, R., Osher, S.: A splitting method for orthogonality constrained problems. Journal of Scientific Computing **58**(2), 431–449 (2014)
32. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM Journal on Optimizaiton, 25(4), 2434-2460, (2015)
33. Li, R.C., Stewart, G.: A new relative perturbation theorem for singular subspaces. Linear Algebra and its Applications **313**(1), 41–51 (2000)
34. Liavas, A.P., Sidiropoulos, N.D.: Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers. IEEE Transactions on Signal Processing, 63(20), 5450 – 5463, (2015)
35. Łojasiewicz, S.: Sur la géométrie semi-et sous-analytique. Ann. Inst. Fourier (Grenoble) **43**(5), 1575–1595 (1993)
36. Lu, Z., Zhang, Y.: An augmented lagrangian approach for sparse principal component analysis. Mathematical Programming **135**(1-2), 149–193 (2012).
37. Magnússon, S., Weeraddana, P.C., Rabbat, M.G., Fischione, C.: On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. IEEE Transactions on Control of Network Systems, 3(3), 296 – 309 (2015)
38. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM Journal on Control and Optimization **15**(6), 959–972 (1977)
39. Miksik, O., Vineet, V., Pérez, P., Torr, P.H., Cesson Sévigné, F.: Distributed non-convex ADMM-inference in large-scale random fields. In: British Machine Vision Conference, BMVC (2014)
40. Mo llenhoff, T., Strekalovskiy, E., Moeller, M., Cremers, D.: The primal-dual hybrid gradient method for semiconvex splittings. SIAM Journal on Imaging Sciences **8**(2), 827–857 (2015)
41. Oymak, S., Mohan, K., Fazel, M., Hassibi, B.: A simplified approach to recovery conditions for low rank matrices. In: Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on, pp. 2318–2322. IEEE (2011)
42. Peng, Z., Xu, Y., Yan, M., Yin, W.:ARock: An Algorithmic Framework for Asynchronous Parallel Coordinate Updates. SIAM Journal on Scientific Computing **38**(5), A2851-A2879 (2016)
43. Poliquin, R., Rockafellar, R.: Prox-regular functions in variational analysis. Transactions of the American Mathematical Society **348**(5), 1805–1838 (1996)

44. Powell, M.J.: A method for non-linear constraints in minimization problems. UKAEA (1967)
45. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis (2009)
46. Rosenberg, J., et al.: Applications of analysis on lipschitz manifolds. Proc. Miniconferences on Harmonic Analysis and Operator Algebras (Canberra, t987), Proc. Centre for Math. Analysis **16**, 269–283 (1988)
47. Shen, Y., Wen, Z., Zhang, Y.: Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. Optimization Methods Software **29**(2), 239–263 (2014)
48. Slavakis, K., Giannakis, G., Mateos, G.: Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge. Signal Processing Magazine, IEEE **31**(5), 18–31 (2014)
49. Sun, D.L., Fevotte, C.: Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 6201–6205. IEEE (2014)
50. Sun, R., Luo, Z., Ye, Y.:On the expected convergence of randomly permuted ADMM. arXiv preprint arXiv:1503.06387 (2015)
51. Wang, F., Cao, W., Xu, Z.: Convergence of multi-block Bregman ADMM for nonconvex composite problems. arXiv preprint arXiv:1505.03063 (2015)
52. Wang, F., Xu, Z., Xu, H.K.: Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems. arXiv preprint arXiv:1410.8625 (2014)
53. Wang, X., Hong, M., Ma, S., Luo, Z.Q.: Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers. arXiv preprint arXiv:1308.5294 (2013)
54. Wang, Y., Zeng, J., Peng, Z., Chang, X., Xu, Z.: Linear convergence of adaptively iterative thresholding algorithm for compressed sensing, IEEE Transactions on Signal Processing, **63**(11): 2957-2971 (2015)
55. Watson, G.A.: Characterization of the subdifferential of some matrix norms. Linear Algebra and its Applications **170**, 33–45 (1992)
56. Wen, Z., Peng, X., Liu, X., Sun, X., Bai, X.: Asset Allocation under the Basel Accord Risk Measures. ArXiv preprint ArXiv:1308.1321 (2013)
57. Wen, Z., Yang, C., Liu, X., Marchesini, S.: Alternating direction methods for classical and ptychographic phase retrieval. Inverse Problems **28**(11), 115,010 (2012)
58. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Mathematical Programming **142**(1-2), 397–434 (2013)
59. Wikipedia: Schatten norm — Wikipedia, the free encyclopedia (2015). [Online; accessed 18-October-2015]
60. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on Imaging Sciences **6**(3), 1758–1789 (2013)
61. Xu, Y., Yin, W., Wen, Z., Zhang, Y.: An alternating direction algorithm for matrix completion with nonnegative factors. Frontiers of Mathematics in China **7**(2), 365–384 (2012)
62. Yan, M. and Yin, W.: Self equivalence of the alternating direction method of multipliers. In (Glowinski, R., Osher, S., Yin, W. ed.) Splitting Methods in Communication, Imaging, Science and Engineering. Springer, New York, (2016)
63. Yang, L., Pong, T.K., Chen, X.: Alternating direction method of multipliers for nonconvex background/foreground extraction. SIAM Journal on Imaging Sciences, **10**(1), 74 - 110, (2017)
64. You, S., Peng, Q.: A non-convex alternating direction method of multipliers heuristic for optimal power flow. In: Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on, pp. 788–793. IEEE (2014)
65. Zeng, J., Lin, S., Wang, Y., Xu, Z.: $L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm. IEEE Transactions on Signal Processing, **62**(9): 2317-2329 (2014)
66. Zeng, J., Lin, S., Xu, Z.: Sparse regularization: Convergence of iterative jumping thresholding algorithm. IEEE Transactions on Signal Processing, **64**(19): 5106-5117 (2016)
67. Zeng, J., Peng, Z., Lin, S.: A Gauss-Seidel iterative thresholding algorithm for ℓ_q regularized least squares regression. Journal of Computational and Applied Mathematics, **319**:220-235 (2017)