Global Convergence of ADMM in Nonconvex Nonsmooth Optimization

Yu Wang · Wotao Yin · Jinshan Zeng†

November 29, 2015

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, $\phi(x_1,\ldots,x_p,y)$, subject to linear equality constraints that couple $x_1,\ldots,x_p,y$, where $p \geq 1$ is an integer. Our ADMM sequentially updates the primal variables in the order $x_1,\ldots,x_p,y$, 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 ($0 < q < 1$) and SCAD, as well as the indicator functions of compact smooth manifolds (e.g., spherical, Stiefel, and Grassman manifolds). 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. Indicated from this example and other analysis in this paper, ADMM might be a better choice than ALM for nonconvex *nonsmooth* problems, because ADMM is not only easier to implement, it is also more likely to converge.

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

The work of W. Yin is supported in part by NSF grants DMS-1317602 and ECCS-1462398 and ONR Grants N000141410683 and N000141210838.

Y. Wang
Department of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
E-mail: shifwang@gmail.com

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: jinshanzeng@jxnu.edu.cn

1 Introduction

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

$$\begin{aligned} \underset{x_1,\ldots,x_p,y}{\text{minimize}} \quad & \phi(x_1,\ldots,x_p,y) \\ \text{subject to} \quad & A_1x_1+\cdots+A_px_p+By=b, \end{aligned} \tag{1.1}$$

where ϕ is a continuous function, $x_i \in \mathbb{R}^{n_i}$ are variables along with their coefficient matrices $A_i \in \mathbb{R}^{m\times n_i}$, $i=1,\ldots,p$, and $y\in\mathbb{R}^q$ is the other variable with its coefficient matrix $B\in\mathbb{R}^{m\times q}$. Clearly, the model is as general if y and By are removed; however, as y and B are treated differently from x_i and A_i in our analysis, keeping them simplifies 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., $b\in \mathrm{Im}(B)$.

In spite of successful applications of ADMM to convex problems, the behavior of ADMM applied to nonconvex problems has been a mystery, especially when there are also nonsmooth terms in the problems. ADMM should generally fail due to nonconvexity but, for so many problems we found it not only worked but also had good performance. Indeed, the nonconvex case of the problem (2.2) has found wide applications in, for example, matrix completion and separation [64,63,68,45,48], asset allocation [56], tensor factorization [28], phase retrieval [57], compressive sensing [7], optimal power flow [65], direction fields correction [25], noisy color image restoration [25], image registration [5], network inference [33], and global conformal mapping [25]. Partially motivated by the recent work [20], we present our Algorithm 1, where L_β denotes the augmented Lagrangian, and show that it converges for a large class of problems. For simplicity, Algorithm 1 only uses the standard ADMM subproblems, which minimize the augmented Lagrangian L_β with all but one variable fixed. It is possible to extend it to inexact and prox-gradient types of subproblems as long as a few key principles (cf. §3.1) are preserved.

In the aforementioned applications, the objective function f 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 compact smooth manifold.

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 [8]. Nonetheless, we still find that it often 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?

In this paper, under some assumptions on the objective and matrices, Algorithm 1 is proved to converge. Even if the objective functions contain sparsity-inducing functions, indicator functions [3] of smooth manifolds or piecewise linear functions, ADMM is still able to converge to a stationary point of the augmented Lagrangian. This paper also extends the theory of coordinate descent methods because Algorithm 1 applies cyclic coordinate descent to the primal variables of our model, yet our model includes the linear constraints that couple all the variables, and such constraints, treated as (nonsmooth) indicator functions, will break the existing analysis of coordinate descent algorithms.

Algorithm 1 Nonconvex ADMM for (1.1)

Initialize $x_2^0, \dots, x_p^0, y^0, w^0$ such that $B^T w^0 = -\nabla h(y^0)$
while stopping criteria not satisfied **do**
  for $i = 1, \dots, p$ **do**
    $x_i^{k+1} \leftarrow \mathrm{argmin}_{x_i} L_\beta(x_{<i}^{k+1}, x_i, x_{>i}^k, y^k, w^k)$;
  end for
  $y^{k+1} \leftarrow \mathrm{argmin}_y L_\beta(\mathbf{x}^{k+1}, y, w^k)$;
  $w^{k+1} \leftarrow w^k + \beta\left(\mathbf{A}\mathbf{x}^{k+1} + By^{k+1}\right)$;
  $k \leftarrow k+1$;
end while
return $x_1^k, \dots, x_p^k$ and y^k.

1.1 Proposed algorithm

Denote the variable $\mathbf{x} := [x_1; \dots; x_p] \in \mathbb{R}^n$ where $n = \sum_{i=1}^p n_i$. Let $\mathbf{A} := [A_1 \ \cdots \ A_p] \in \mathbb{R}^{m \times n}$ and $\mathbf{A}\mathbf{x} := \sum_{i=1}^p A_i x_i \in \mathbb{R}^m$. To present our algorithm, we define the augmented Lagrangian:

$$L_\beta(\mathbf{x}, y, w) := \phi(\mathbf{x}, y) + \langle w, \mathbf{A}\mathbf{x} + By \rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} + By\|^2. \tag{1.2}$$

The proposed Algorithm 1 extends the standard ADMM to have multiple variable blocks. It also extends the *cooridnate descent* algorithms to include linear constraints. We let $x_{<i} := (x_1, \dots, x_{i-1})$ and $x_{>i} := (x_{i+1}, \dots, x_p)$. The convergence of Algorithm 1 will be given in Theorem 2.1.

1.2 Relation to the augmented Lagrangian method (ALM)

ADMM is an approximation to ALM by splitting up the ALM subproblem and sequentially updating each primal variable. Surprisingly, there is a very simple nonconvex problem on which ALM diverges yet ADMM converges.

Proposition 1.1 *For the problem*

$$\begin{aligned} \underset{x,y \in \mathbb{R}}{\text{minimize}} \quad & x^2 - y^2 \\ \text{subject to} \quad & x = y,\ x \in [-1, 1], \end{aligned} \tag{1.3}$$

it holds that

1. *for any fixed $\beta > 0$, ALM generates a divergent sequence;*
2. *for any fixed $\beta > 1$, ADMM generates a sequence that converges to a solution, in finitely many steps.*

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

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

where ι_S denotes the indicator function of set S: $\iota_S(x) = 0$ if $x \in S$; otherwise, equals infinity. It turns out that ADMM solves (1.3) by performing the coordinate descent iteration to $\text{minimize}_{x,y}\ \rho(x, y)$:

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

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

1.3 Related literature

Our work is related to block coordinate-descent (BCD) algorithms for *unconstrained* optimization. BCD can be traced back to [35] for solving linear systems and also to [19,54,2,44], where the objective function f is assumed to be convex (or quasi-convex or hemivariate), differentiable, or has bounded level sets. When f is nonconvex, the original BCD may cycle and stagnate [39]. When f is nonsmooth, the original BCD can get stuck at a non-stationary point [2, Page 94], but not if the nonsmooth part is separable (see [17,31,49, 50,60,41,40,61,62] for results on different forms of f and different variants of BCD). If f is differentiable, multi-convex or has separable non-smooth parts, then every limit point of the sequence generated by the *proximal* BCD is a critical point [16,50,60,61], but again, the nonsmooth terms cannot couple more than one block of variable. When there are linear constraints like in those in (2.2), BCD algorithms may fail [47].

The original ADMM was proposed in [15,13]. For convex problems, its convergence was established first in [14] and its convergence rates given in [18,11,10] in different settings. When the objective function is nonconvex, the recent results [68,63,23,30] directly make assumptions on the iterates $(\mathbf{x}^k, y^k, w^k)$. Hong et al. [20] deals with the nonconvex separable objective functions for some specific A_i, which forms the sharing and consensus problem. Li and Pong [26] 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 and include ADMM as a special case. We review their results and compare to ours in Section 3.4 below.

1.4 Organization

The remainder of this paper is organized as follows. Section 2 presents the main convergence theorem. Section 3 gives the detailed proofs of our theorem and discusses the tightness of our assumptions. Section 4 applies our theorem in some typical applications and obtains novel convergence results. Finally, section 5 concludes this paper.

2 Main results

2.1 Definitions

In these definitions, ∂f denotes the set of general subgradients of f in [42, Definition 8.3]. We call a function *Lipschitz differentiable* if it is differentiable and the gradient is Lipschitz continuous. The functions given in the next two definitions are permitted in our model.

Definition 2.1 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 $a_i \in \mathbb{R}^n$, and points $b_i \in \mathbb{R}$ 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 $x \in U_i$.

Definition 2.2 (Restricted prox-regularity) Let $M \in \mathbb{R}_+$, $f : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, and define the exclusion set

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

We say that f is *restricted prox-regular* if, for a sufficiently large M, S_M is strictly contained in $\mathrm{dom}(\partial f)$, and for any bounded set $T \subset \mathbb{R}^N$, there exists $\gamma > 0$ such that

$$f(y) + \frac{\gamma}{2}\|x - y\|^2 \geq f(x) + \langle d, y - x\rangle, \quad \forall\ x, y \in T \setminus S_M,\ d \in \partial f(x),\ \|d\| \leq M. \tag{2.1}$$

(If $T \setminus S_M$ is empty, (2.1) is satisfied.) □

Throughout the paper, $\|\cdot\|$ represents the Euclidean norm. Definition 2.2 is related to, but different from, the concepts *prox-regularity* [38], *hypomonotonicity* [42, Example 12.28] and *semi-convexity* [32,22,24,34], all of which impose global conditions. Definition 2.2 only requires (2.1) to hold over a subset. Note that many concerned examples in our paper, such as ℓ_q quasi-norms ($0 < q < 1$), Schatten-q quasi-norms ($0 < q < 1$), indicator functions of compact smooth manifolds and etc., are *not* prox-regular, hypomonotone or semiconvex. According to Definition 2.2, all proper closed convex functions satisfy the inequality in (2.1) with $\gamma = 0$. The definition introduces functions that do not satisfy (2.1) globally *only because* they are asymptotically "steep" in the exclusion set S_M. Such functions include the ℓ_q quasi-norm ($0 < q < 1$), 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. The exclusion set can be excluded because the sequence x_i^k of Algorithm 1 never enters the set S_M for each f_i when M and k are large enough; therefore, we only need (2.1) to hold elsewhere.

2.2 Main theorems

We consider two different scenarios

- Theorem 2.1 below considers the scenario where $\mathbf{x}$ and y are decoupled in the objective function;
- Theorem 2.2 below considers the scenarios where $\mathbf{x}$ and y are possibly coupled but their function $\phi(x_1, \ldots, x_p, y)$ is Lipschitz differentiable.

The model in the first scenario is

$$\begin{aligned} \underset{x_1,\ldots,x_p,y}{\text{minimize}} \quad & f(x_1, \ldots, x_p) \quad + h(y) \\ \text{subject to} \quad & A_1x_1 + \cdots + A_px_p + By = b, \end{aligned} \tag{2.2}$$

where the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ ($n = \sum_{i=1}^p n_i$) is continuous, proper, possibly nonsmooth and nonconvex, and the function $h : \mathbb{R}^q \to \mathbb{R}$ is proper, differentiable, and possibly nonconvex.

Theorem 2.1 *Suppose that*

A1 ***(coercivity)*** *The feasible set $\mathcal{F} := \{(\mathbf{x}, y) \in \mathbb{R}^{n+q} : \mathbf{A}\mathbf{x} + By = 0\}$ is nonempty, and the objective function $f(\mathbf{x}) + h(y)$ is coercive over this set, that is, $f(\mathbf{x}) + h(y) \to \infty$ if $(\mathbf{x}, y) \in \mathcal{F}$ and $\|(\mathbf{x}, y)\| \to \infty$; (The objective function does* not *need to be coercive over $\mathbb{R}^{n+q}$.)*

A2 ***(feasibility)*** $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(B)$*, where* $\mathrm{Im}(\cdot)$ *returns the image of a matrix;*
A3 ***(objective-f regularity)*** f *has the form*

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

where
(i) $g(\mathbf{x})$ *is Lipschitz differentiable with constant* L_g*,*
(ii) $f_i(x_i)$ *is either restricted prox-regular (definition 2.2) for* $i = 1, \ldots, p$*, or continuous and piecewise linear (definition 2.1) for* $i = 1, \ldots, p$*;*
A4 ***(objective-h regularity)*** $h(y)$ *is Lipschitz differentiable with constant* L_h*;*
A5 ***(Lipschitz sub-minimization paths)***
(a) There exists a Lipschitz continuous map $H : \mathrm{Im}(B) \to \mathbb{R}^q$ *obeying* $H(u) = \mathrm{argmin}_y\{h(y) : By = u\}$*,*
(b) for $i = 1, \ldots, p$ *and any* $x_{<i}$ *and* $x_{>i}$*, there exists a Lipschitz continuous map* $F_i : \mathrm{Im}(A_i) \to \mathbb{R}^{n_i}$ *obeying* $F_i(u) = \mathrm{argmin}_{x_i}\{f(x_{<i}, x_i, x_{>i}) : A_i x_i = u\}$*,*
and that the above F_i *and* H *have a universal Lipschitz constant* $\bar{M} > 0$*.*

Then, Algorithm 1 converges globally for any sufficiently large β *(the lower bound is given in Lemma 3.9), that is, starting from any* $x_2^0, \ldots, x_p^0, y^0, w^0$*, 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* L_β*, namely,* $0 \in \partial L_\beta(\mathbf{x}^*, y^*, w^*)$*.*

In addition, if L_β *is a Kurdyka-Łojasiewicz (KŁ) function [29, 4, 1], then* $(\mathbf{x}^k, y^k, w^k)$ *converges globally to the unique limit point* $(\mathbf{x}^*, y^*, w^*)$*.*

Assumptions A3 and A4 regulate the objective functions. None of the functions needs to be convex. The non-Lipschitz differentiable parts $f_1, \ldots, f_n$ of f shall satisfy either Definition 2.1 or Definition 2.2. Under assumptions A3 and A4, the augmented Lagrangian function L_β 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 establish before that property is used in our proof. Consequently, g can be functions such as e^x, whose derivative is not globally Lipschitz.

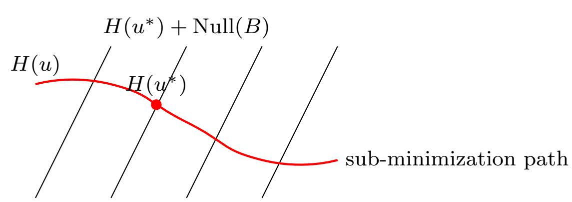

Fig. 2.1 Illustration of assumption A5, which assume that $H(u) = \mathrm{argmin}\{h(y) : By = u\}$ is a Lipschitz manifold [43].

Assumption A5 weakens the full column rank assumption typically assumed for matrices A_i and B. When A_i and B do have full column rank, their null spaces are trivial and, therefore, F_i, H reduce to linear operators and satisfy A5. However, assumption A5 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 A5 for any matrix A.

Functions satisfying the KŁ inequality include real analytic functions, semialgebraic 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 do not need to be decoupled. However, the objective function needs to be smooth.

Theorem 2.2 *Suppose that*

B1 ***(coercivity)*** *ϕ is coercive over the nonempty feasible set $\mathcal{F} := \{(\mathbf{x}, y) \in \mathbb{R}^{n+q} : \mathbf{A}\mathbf{x} + By = 0\}$;*
B2 ***(feasibility)*** $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(B)$*;*
B3 ***(smoothness)*** *ϕ is Lipschitz differentiable with constant L_ϕ;*
B4 ***(Lipschitz sub-minimization paths)***
  (a) For any $\mathbf{x}$, there exists a Lipschitz continuous map $H : \mathrm{Im}(B) \to \mathbb{R}^q$ obeying $H(u) = \mathrm{argmin}_y\{\phi(\mathbf{x}, y) : By = u\}$,
  (b) For $i = 1, \ldots, p$ and any $x_{<i}$, $x_{>i}$ and y, there exists a Lipschitz continuous map $F_i : \mathrm{Im}(A_i) \to \mathbb{R}^{n_i}$ obeying $F_i(u) = \mathrm{argmin}_{x_i}\{\phi(x_{<i}, x_i, x_{>i}, y) : A_i x_i = u\}$,
  and that the above F_i and H have a universal Lipschitz constant $\bar{M} > 0$.

Then, Algorithm 1 converges globally for any sufficiently large β, that is, starting from any $x_2^0, \ldots, x_p^0, y^0, w^0$, 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 L_β, namely, $0 \in \partial L_\beta(\mathbf{x}^*, y^*, w^*)$.

In addition, if L_β is a Kurdyka-Łojasiewicz (KŁ) function, then $(\mathbf{x}^k, y^k, w^k)$ converges globally to the unique limit point $(\mathbf{x}^*, y^*, w^*)$.

Although Theorems 2.1 and 2.2 impose different conditions on the objective functions, their proofs are very similar. Therefore, we will focus on proving Theorem 2.1 in this paper and leave the proof of Theorem 2.2 to the Appendix.

3 Proof and discussion

3.1 Keystones

The following properties hold for Algorithm 1 under our assumptions. Here, we first list them and present Proposition 3.1, 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 $L_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded;
P2 **(sufficient descent)** there is $C_1 > 0$ such that for all sufficiently large k, we have

$$L_\beta(\mathbf{x}^k, y^k, w^k) - L_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \geq C_1\big(\|B(y^{k+1} - y^k)\|^2 + \sum_{i=1}^{p} \|A_i(x_i^k - x_i^{k+1})\|^2\big), \tag{3.1}$$

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

$$\|d\| \leq C_2\big(\|B(y^{k+1} - y^k)\| + \sum_{i=1}^{p} \|A_i(x_i^{k+1} - x_i^k)\|\big). \tag{3.2}$$

Proposition 3.1 *Suppose that when an algorithm is applied to the problem (2.2), 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 solution. That is,* $0 \in \partial L_\beta(\mathbf{x}^*, y^*, w^*)$*, or equivalently,*

$$0 = \mathbf{A}\mathbf{x}^* + By^*, \tag{3.3a}$$

$$0 \in \partial f(\mathbf{x}^*) + \mathbf{A}^T w^*, \tag{3.3b}$$

$$0 \in \partial h(y^*) + B^T w^*. \tag{3.3c}$$

Moreover, if L_β *is a KŁ function, then* $(\mathbf{x}^k, y^k, w^k)$ *converges globally to the unique point* $(\mathbf{x}^*, y^*, w^*)$.

Proof 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, $L_\beta(\mathbf{x}^k, y^k, w^k)$ 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 $d^k \in \partial L_\beta(\mathbf{x}^k, y^k, w^k)$ such that $\|d^k\| \to 0$. In particular, $\|d^{k_s}\| \to 0$ as $s \to \infty$. By definition of general subgradient [42, Definition 8.3], we have $0 \in \partial L_\beta(\mathbf{x}^*, y^*, w^*)$.

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 L_β. □

In P2, the sufficient descent inequality 3.1 does *not* necessarily hold for 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 should consider P1–P3 when generalizing Algorithm 1, for example, by replacing the direct minimization subproblems to prox-gradient subproblems.

3.2 Preliminaries

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

$$(\mathbf{x}^+, y^+, w^+) := (\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}). \tag{3.4}$$

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

Lemma 3.1 *Under assumptions A1–A5, if* $\beta > \bar{M}^2 L_h$*, all the subproblems in Algorithm 1 are well defined.*

Proof First of all, let us prove the y-subproblem is well defined. By A5, we know $h(y) \geq h(H(By))$. Because of A4 and A5, we have

$$h(H(By)) - h(H(0)) - \big\langle \nabla h(H(0)), H(By) - H(0) \big\rangle \geq -\frac{L_h}{2}\|H(By) - H(0)\|^2 \geq -\frac{L_h \bar{M}^2}{2}\|By\|^2$$

and

$$\big\langle \nabla h(H(0)), H(By) - G(0) \big\rangle \geq -\|\nabla h(H(0))\| \cdot \bar{M} \cdot \|By\|.$$

Thus we have

$$\begin{aligned} & f(\mathbf{x}^+) + h(y) + \big\langle w^k, \mathbf{A}\mathbf{x}^+ + By \big\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^+ + By\|^2 \\ \geq & f(\mathbf{x}^+) + h(H(0)) + \big\langle w^k, \mathbf{A}\mathbf{x}^+ \big\rangle - (\|\nabla h(H(0))\|\bar{M} + \|w^k\|) \cdot \|By\| + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^+ + By\|^2 - \frac{L_h \bar{M}^2}{2}\|By\|^2. \end{aligned} \tag{3.5}$$

As $\|By\|$ goes to ∞, this lower bound goes to ∞, so the y-subproblem is well defined.

As for the x_i-subproblem, ignoring some constants yields

$$\begin{aligned}
&\text{argmin } L_\beta(x^+_{<i}, x_i, x^k_{>i}, y^k, w^k) && (3.6)\\
&= \text{argmin } f(x^+_{<i}, x_i, x^k_{>i}) + \frac{\beta}{2}\|\frac{1}{\beta}w^k + A_{<i}x^+_{<i} + A_{>i}x^k_{>i} + A_i x_i + By^k\|^2 && (3.7)\\
&= \text{argmin } f(x^+_{<i}, x_i, x^k_{>i}) + h(u) - h(u) + \frac{\beta}{2}\|Bu - By^k - \frac{1}{\beta}w^k\|^2. && (3.8)
\end{aligned}$$

where $u = H(-A_{<i}x^+_{<i} - A_{>i}x^k_{>i} - A_i x_i)$. The first two terms are lower 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. This shows all subproblems are well defined. □

Lemma 3.2 *It holds that, $\forall k_1, k_2 \in \mathbb{N}$,*

$$\|y^{k_1} - y^{k_2}\| \le \bar{M}\|By^{k_1} - By^{k_2}\|, \tag{3.9}$$

$$\|x_i^{k_1} - x_i^{k_2}\| \le \bar{M}\|A_i x_i^{k_1} - A_i x_i^{k_2}\|, \quad i = 1, \ldots, p, \tag{3.10}$$

where $\bar{M}$ is given in A5.

Proof By the definitions of H in A5(a) and y^k, we have $y^k = H(By^k)$. Therefore, $\|y^{k_1} - y^{k_2}\| = \|H(By^{k_1}) - H(By^{k_2})\| \le \bar{M}\|By^{k_1} - By^{k_2}\|$. Similarly, by the optimality of x_i^k, we have $x_i^k = F_i(A_i x_i^k)$. Therefore, $\|x_i^{k_1} - x_i^{k_2}\| = \|F_i(A_i x_i^{k_1}) - F_i(A_i x_i^{k_2})\| \le \bar{M}\|A_i x_i^{k_1} - A_i x_i^{k_2}\|$. □

Lemma 3.3 *It holds that*

a. $B^T w^k = -\nabla h(y^k)$ for all $k \in \mathbb{N}$.
b. there exists a constant $C > 0$ such that for all $k \in \mathbb{N}$

$$\|w^+ - w^k\| \le C\|By^+ - By^k\|.$$

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

Part b. Let $\lambda_{++}(B^TB)$ denote the smallest strictly-positive eigenvalue of B^TB, $C_1 := \lambda_{++}^{-1/2}(B^TB)$, and $C := L_h\bar{M}C_1$. Since $w^+ - w^k = \beta(A\mathbf{x}^+ + By^+) \in \mathrm{Im}(B)$, we get $\|w^+ - w^k\| \le C_1\|B(w^+ - w^k)\| = C_1\|\nabla h(y^+) - \nabla h(y^k)\| \le C\|By^+ - By^k\|$, where the last inequality follows from Lemma 3.2. □

3.3 Main proof

This subsection proves Theorem 2.1 for Algorithm 1 under Assumptions A1–A5. For all $k \in \mathbb{N}$ and $i = 1, \ldots, p$, the general subgradients of f, f_i, and h exist at $\mathbf{x}_k$, x_i^k, and y_k, respectively, and we define the general subgradients

$$\bar{d}_i^k := -(A_i^T w^+ + \beta\rho_i^k) \in \partial_i f(x^+_{<i}, x^+_i, x^k_{>i}), \tag{3.11}$$

$$d_i^k := -\nabla_i g(x^k_{<i}, x^k_i, x^{k-1}_{>i}) + \bar{d}_i^k \in \partial f_i(x_i^+), \tag{3.12}$$

where

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

The next two lemmas study how much $L_\beta(\mathbf{x}, y, w)$ decreases at each iteration.

Lemma 3.4 (decent of L_β during x_i update) *The iterates in Algorithm 1 satisfy*

1. $L_\beta(x^+_{<i}, \boldsymbol{x}^k_i, x^k_{>i}, y^k, w^k) \geq L_\beta(x^+_{<i}, \boldsymbol{x}^+_i, x^k_{>i}, y^k, w^k)$, $i = 1, \ldots, p$;
2. $L_\beta(\mathbf{x}^k, y^k, w^k) \geq L_\beta(\mathbf{x}^+, y^k, w^k)$;
3. $L_\beta(\mathbf{x}^k, y^k, w^k) - L_\beta(\mathbf{x}^+, y^k, w^k) = \sum_{i=1}^p r_i$, *where*
$$r_i := f(x^+_{<i}, x^k_i, x^k_{>i}) - f(x^+_{<i}, x^+_i, x^k_{>i}) - \langle \bar{d}^k_i, x^k_i - x^+_i\rangle + \frac{\beta}{2}\|A_i x^k_i - A_i x^+_i\|^2, \tag{3.13}$$
where $\bar{d}^k_i$ *is defined in* (3.11).
4. *If*
$$f_i(x^k_i) + \frac{\gamma_i}{2}\|x^k_i - x^+_i\|^2 \geq f_i(x^+_i) + \langle d^k_i, x^k_i - x^+_i\rangle, \tag{3.14}$$
holds with constant $\gamma_i > 0$ *(later, this assumption will be shown to hold), then we have*
$$r_i \geq \frac{\beta - \gamma_i \bar{M}^2 - L_g \bar{M}^2}{2}\|A_i x^k_i - A_i x^+_i\|^2, \tag{3.15}$$
where the constants L_g *and* $\bar{M}$ *are defined in Assumptions A3 and A5, respectively.*

Proof **Part 1** follows directly from the definition of x^+_i. **Part 2** is a result of

$$L_\beta(\mathbf{x}^k, y^k, w^k) - L_\beta(\mathbf{x}^+, y^k, w^k) = \sum_{i=1}^p \left(L(x^+_{<i}, x^k_i, x^k_{>i}, y^k, w^k) - L(x^+_{<i}, x^+_i, x^k_{>i}, y^k, w^k)\right),$$

and part 1. **Part 3**: Each term in the sum equals $f(x^+_{<i}, x^k_i, x^k_{>i}) - f(x^+_{<i}, x^+_i, x^k_{>i})$ *plus*

$$\begin{aligned}
&\langle w^k, A_i x^k_i - A_i x^+_i\rangle + \frac{\beta}{2}\|A_{<i}x^+_{<i} + A_i x^k_i + A_{>i}x^k_{>i} + By^k\|^2 - \frac{\beta}{2}\|A_{<i}x^+_{<i} + A_i x^+_i + A_{>i}x^k_{>i} + By^k\|^2 \\
&= \langle w^k, A_i x^k_i - A_i x^+_i\rangle + \langle \beta\left(A_{<i}x^+_{<i} + A_i x^+_i + A_{>i}x^k_{>i} + By^k\right), A_i x^k_i - A_i x^+_i\rangle + \frac{\beta}{2}\|A_i x^k_i - A_i x^+_i\|^2 \\
&= \langle A_i^T w^+ + \beta\rho^k_i, x^k_i - x^+_i\rangle + \frac{\beta}{2}\|A_i x^k_i - A_i x^+_i\|^2
\end{aligned}$$

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

Part 4. Let d^k_i be defined in (3.12). From the inequalities (3.14) and (3.10), we get

$$f_i(x^k_i) - f_i(x^+_i) - \langle d^k_i, x^k_i - x^+_i\rangle \geq -\frac{\gamma_i}{2}\|x^k_i - x^+_i\|^2 \geq -\frac{\gamma_i \bar{M}^2}{2}\|Ax^k_i - Ax^+_i\|^2. \tag{3.16}$$

By Assumption A3 part (i) and inequality (3.10), we also get

$$g(x^+_{<i}, x^k_i, x^k_i) - g(x^+_{<i}, x^+_i, x^k_{>i}) - \langle \nabla_i g(x^+_{<i}, x^+_i, x^k_{>i}), x^k_i - x^+_i\rangle \geq -\frac{L_g}{2}\|x^k_i - x^+_i\|^2 \geq -\frac{L_g \bar{M}^2}{2}\|Ax^k_i - Ax^+_i\|^2. \tag{3.17}$$

Finally, rewriting the expression of r_i and applying (3.16) and (3.17) we obtain

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

□

The assumption (3.14) in part 4 is the same as (2.1) in Definition 2.2 except the latter holds for more functions by excluding the points in S_M. In order to relax (3.14) to (2.1), we must find M and specify the exclusion set S_M. We will achieve this in Lemma 3.9 after necessary properties have been established.

Lemma 3.5 (decent of L_β during y and w updates) *If $\beta > L_h\bar{M}^2 + 1 + C$, where C is the constant in Lemma 3.3 and L_h is the constant in Assumption A4, then there exists constant $C_1 > 0$ such that for $k \in \mathbb{N}$*

$$L(\mathbf{x}^+, y^k, w^k) - L(\mathbf{x}^+, y^+, w^+) \geq C_1\|By^+ - By^k\|^2. \tag{3.18}$$

Proof From Assumption A4 and Lemma 3.3 part 2, it follows

$$\begin{aligned} &L(\mathbf{x}^+, y^k, w^k) - L(\mathbf{x}^+, y^+, w^+) \\ &= h(y^k) - h(y^+) + \langle w^+, By^k - By^+\rangle + \frac{\beta}{2}\|By^+ - By^k\|^2 - \frac{1}{\beta}\|w^+ - w^k\|^2 \quad (3.19)\\ &\geq -\frac{L_h\bar{M}^2}{2}\|By^+ - By^k\|^2 + \frac{\beta}{2}\|By^+ - By^k\|^2 - \frac{C}{\beta}\|By^+ - By^k\|^2 \quad (3.20)\\ &= C_1\|By^+ - By^k\|^2, \end{aligned}$$

where the lower bound of β is picked to make sure $C_1 > 0$. □

Based on Lemma 3.4 and Lemma 3.5, we now establish the following results:

Lemma 3.6 (Monotone, lower–bounded L_β and bounded sequence P1) *If β is set as in Lemma 3.5, then the sequence $(\mathbf{x}^k, y^k, w^k)$ of Algorithm 1 satisfies*

1. $L_\beta(\mathbf{x}^k, y^k, w^k) \geq L_\beta(\mathbf{x}^+, y^+, w^+)$.
2. $L_\beta(\mathbf{x}^k, y^k, w^k)$ *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 3.4 part 2, and Lemma 3.5.

Part 2. By Assumption A2, there exists y' such that $\mathbf{A}\mathbf{x}^k + By' = 0$ and $y' = H(By')$. By 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

$$\begin{aligned} L_\beta(\mathbf{x}^k, y^k, w^k) &= f(\mathbf{x}^k) + h(y^k) + \langle B^T w^k, y^k - y'\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\ &= f(\mathbf{x}^k) + h(y^k) + \langle \nabla h(y^k), y' - y^k\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\ \text{(Lemma 3.2, } \nabla h \text{ is Lipschitz)} \quad &\geq f(\mathbf{x}^k) + h(y') + \frac{\beta - L_h\bar{M}^2}{2}\|\mathbf{A}\mathbf{x}^k + By^k\|^2 \\ &> -\infty. \end{aligned}$$

Part 3. From parts 1 and 2, $L_\beta(\mathbf{x}^k, y^k, w^k)$ is upper bounded by $L_\beta(\mathbf{x}^0, y^0, w^0)$ and so are $f(\mathbf{x}^k) + h(y')$ and $\|\mathbf{A}\mathbf{x}^k + By^k\|^2$. By Assumption A1, $\{\mathbf{x}^k\}$ is bounded and, therefore, $\{By^k\}$ is also bounded. By Lemma 3.2 and Lemma 3.3, we know that $\{y^k\}$ and $\{w^k\}$ are also bounded. □

It is important to remark that, once β is larger than a threshold, the constants and bounds in Lemmas 3.5 and 3.6 can be chosen *independent of* β, which is essential to the proof of Lemma 3.9 below.

Lemma 3.7 (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 3.4, 3.5, and 3.6 (part 2), and the second from Lemma 3.3 part (a) and that ∇h is Lipschitz. □

Lemma 3.8 (Boundedness for piecewise linear f_i's) *Consider the case that f_i, $i = 1, \ldots, p$, are piecewise linear. For any $\epsilon_0 > 0$, when $\beta > \max\{2(M+1)/\epsilon_0^2, L_h\bar{M}^2 + 1 + C\}$, there exists $k_{\text{pl}} \in \mathbb{N}$ such that the followings hold for all $k > k_{\text{pl}}$:*

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

where $\bar{M}$ and L_g are defined in A5 and A3.

Proof Part 1. Since the number K of the linear pieces of f_i is finite and $\{\mathbf{x}^k, y^k, w^k\}$ is bounded (see Lemma 3.6), $\partial_i f(x_{<i}^+, x_i^+, x_{>i}^k)$ 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 (3.11)), the first three terms of r_i (see (3.13)) are bounded:

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

where M is a universal constant independent of β. Hence, as long as $\beta > 2(M+1)/\epsilon_0^2$,

$$\|A_i x_i^+ - A_i x_i^k\| \geq \epsilon_0 \Rightarrow \; r_i \geq \frac{\beta}{2}\epsilon_0^2 - M > 1 \tag{3.21}$$

$$\Rightarrow \; L_\beta(x_{<i}^+, \boldsymbol{x}_i^k, x_{>i}^k, y^k, w^k) - 1 > L_\beta(x_{<i}^+, \boldsymbol{x}_i^+, x_{>i}^k, y^k, w^k). \tag{3.22}$$

By Lemmas 3.4, 3.5, and 3.6, this means $L_\beta(\mathbf{x}^k, y^k, w^k) - 1 > L_\beta(\mathbf{x}^+, y^+, w^+)$. Since $\{L_\beta(\mathbf{x}^k, y^k, w^k)\}$ is lower bounded, $\|A_i x_i^+ - A_i x_i^k\| \geq \epsilon_0$ can only hold for finitely many k. Then, we get part 1, along with Lemma 3.2. 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 3.2. □

Lemma 3.9 (Sufficient descent property P2) *Suppose*

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

where γ_i $(i = 1, \ldots, p)$ and ϵ_0 are constants only depending on f. Then, Algorithm 1 satisfies the sufficient descent property P2.

Proof We will show the lower bound (3.15) for $i = 1, \ldots, p$, which, along with Lemma 3.4 part 3 and Lemma 3.5, establishes the sufficient descent property P2.

We shall obtain the lower bound (3.15) in the backward order $i = p, (p-1), \ldots, 1$. In light of Lemmas 3.4, 3.5, and 3.6, each lower bound (3.15) for r_i gives us $\|A_i x_i^k - A_i x_i^+\| \to 0$ as $k \to \infty$. We will first show (3.15) 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 (3.15) for the next r_i. We must take this backward order since ρ_i^k (see (3.12)) 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 2.2), we will get (3.15) for r_i by validating the condition (3.14) in Lemma 3.4 part 4 for f_i. In Case 2, f_i's are piecewise linear (cf. Definition 2.1), we will show that (3.14) holds for $\gamma_i = 0$ for $k \geq k_{\mathrm{pl}}$, and following the proof of Lemma 3.4 part 4, we directly get (3.15) with $\gamma_i = 0$.

Base step, take $i = p$.
Case 1) f_p is restricted prox-regular. At $i = p$, the inclusion (3.12) simplifies to

$$d_p^k := -\big(\nabla_p g(\mathbf{x}^+) + A_p^T w^+\big) - \beta A_p^T(By^k - By^+) \in \partial f_p(x_p^+). \tag{3.23}$$

By Lemma 3.6 part 3 and the continuity of ∇g, there exists $M > 0$ (independent of β) such that

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

By Lemma 3.7, 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 (3.23) 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 2.2, which qualifies f_p for (2.1) and thus validates the assumption in Lemma 3.4 part 4, proving the lower bound (3.15) for r_p. As already argued, we get $\lim_k \|A_p x_p^k - A_p x_p^+\| = 0$.

Case 2): f_i's are piecewise linear (cf. Definition 2.1). From $\|By^k - By^+\| \to 0$ and $\|w^k - w^+\| \to 0$ (Lemma 3.7) and $\|\nabla g(\mathbf{x}^k) - \nabla g(\mathbf{x}^+)\| < p\bar{M}L_g\epsilon_0$ (Lemma 3.8). In light of (3.23), $d_p^k \in \partial f_p(x_p^+)$, $d_p^+ \in \partial f_p(x_p^{k+2})$ such that $\|d_p^+ - d_p^k\| < 2p\bar{M}L_g\epsilon_0$ 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) $\|d_p^+ - d_p^k\| < 2p\bar{M}L_g\epsilon_0$ and (ii) $\|x_p^+ - x_p^k\| < \bar{M}\epsilon_0$ (Lemma 3.8 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 combine $\overline{U}_{j_1}$ and $\overline{U}_{j_2}$ together as a new polyhedron. 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 (3.14) for $i = p$ with $\gamma_p = 0$. By following the proof for Lemma 3.4 part 4, we get the lower bound (3.15) for r_p with $\gamma_p = 0$. As already argued, we get $\lim_k \|A_p x_p^k - A_p x_p^+\| = 0$.

Inductive step, let $i \in \{p-1, \ldots, 1\}$ and make the inductive assumption: $\lim_k \|A_j x_j^k - A_j x_j^+\| = 0$, $j = p, \ldots, i+1$, which together with $\lim_k \|By^k - By^+\| = 0$ (Lemma 3.7) gives $\lim_k \rho_i^k = 0$ (defined in (3.12)).

Case 1) f_i is restricted prox-regular. From (3.12), we have

$$d_i^k = -\big(\nabla_i g(x_{<i}^+, x_i^+, x_{>i}^k) + A_p^T w^+\big) - \beta\rho_i^k \in \partial f_i(x_i^+). \tag{3.24}$$

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 2.2 for f_i and thus validates the assumption in Lemma 3.4 part 4 for f_i. Therefore, we get the lower bound (3.15) 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 2.1). The argument is the same as in the base step for case 2, except at its beginning we must use d_i^k in (3.24) instead of d_p^k in (3.23). Therefore, we skip this part.

Finally, by combining $r_i \ge C_1\|A_i x_i^k - A_i x_i^+\|^2$, for $i = 1, \ldots, p$, with Lemmas 3.4 and 3.5, we establish the sufficient descent property P2.

□

Lemma 3.10 (Subgradient bound property P3) *Algorithm 1 satisfies the subgradient bound property P3.*

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

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

In order to prove the lemma, we only need to show that each element can be controlled, in particular,

$$\|\nabla_w L\| \le M\beta\|By^+ - By^k\|, \tag{3.25}$$

$$\|\nabla_y L\| \le M\beta\|By^+ - By^k\|, \tag{3.26}$$

and, for $s = 1, \ldots, p$, there exists $d_s \in \frac{\partial L}{\partial x_s}$ such that

$$\|d_s\| \le M\beta\left(\sum_{i=1}^p \|A_i x_i^+ - A_i x_i^k\| + \|By^+ - By^k\|\right), \tag{3.27}$$

In order to prove (3.25), we have $\nabla_w L = \mathbf{A}\mathbf{x}^+ + By^+ = \frac{1}{\beta}(w^+ - w^k)$. By Lemma 3.3, we have $\|\nabla_w L\| \le M\|By^+ - By^k\|$. In order to prove (3.26), we notice that $\nabla_y L = B^T(w^+ - w^k)$ and apply Lemma 3.3.

In order to prove (3.27), observe that

$$\begin{aligned}\frac{\partial L}{\partial x_s} =& \nabla_s g(x^+) + \partial f_s(x_s^+) + A_s^T w^+ + \beta A_s^T\left(Ax^+ + By^+\right)\\ =& \nabla_s g(x_{\le s}^+, x_{>s}^k) + \partial f_s(x_s^+) + A_s^T w^k + \beta A_s^T\left(A_{\le s}x_{\le s}^+ + A_{>s}x_{>s}^k\right) \qquad (3.28)\\ &+ A_s^T(w^+ - w^k) + \beta A_s^T\left(A_{>s}x_{>s}^+ - A_{>s}x_{>s}^k\right) + \nabla_s g(x^+) - \nabla_s g(x_{\le s}^+, x_{>s}^k), \qquad (3.29)\end{aligned}$$

where (3.29) can be controlled by $M\beta\sum_{i=1}^p \|A_i x_i^+ - A_i x_i^k\| + \|By^+ - By^k\|$ naturally. In (3.28), the first order optimal condition for x_s^+ yields

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

This has completed the whole proof. □

Proof (of Theorem 2.1) .

Lemmas 3.5, 3.9, and 3.10 establish the properties P1–P3. Theorem 2.1 follows from Proposition 3.1. □

3.4 Discussion on the assumptions

In this section, we demonstrate the tightness of the assumptions in Theorem 2.1 and compare them with related recent works. We only focus on results that do *not* make assumptions on the iterates themselves.

Hong et al. [20] uses $\nabla h(y^k)$ to bound w^k. This inspired our analysis. They studied ADMM for nonconvex consensus and sharing problem. Their assumptions for the sharing problem are

(i) $f = \sum_i f_i$, f_i is Lipschitz differentiable or convex. $dom(f)$ is a closed bounded set.
(ii) h is Lipschitz differentiable.
(iii) A_i has full column rank, B is the identity matrix.

The boundedness of dom(f) in part (i) implies our assumption A1, (iii) implies A2 and A5, (i) implies A3, and (ii) implies A4. Our assumptions on f and the matrices A, B are much weaker.

Wang et al. [51] studies the so-called Bregman ADMM and includes the standard ADMM as an special case. 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) $f = \sum_{i=1}^{p} f_i(x_i)$ where f_i, $i = 1, \ldots, p$ is strongly convex.

It is easy to see that (a), (b) and (c) imply our assumptions A1 and A5, (a) implies A2, (c) implies A3 and (b) implies A4. 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 [26] studies ADMM and its proximal version for nonconvex objectives. Their assumptions for ADMM are

(1) $p = 1$ and f is lower semi-continuous.
(2) $h \in C^2$ with bounded Hessian matrix $c_2 I \succeq \nabla^2 h \succeq c_1 I$ where $c_2 > c_1 > 0$.
(3) A is the identity matrix, B has full row rank.
(4) h is coercive and f is lower bounded.

The assumptions (3) and (4) imply our assumption A1, (3) implies A2 and A5, and (2) implies A4. They have a more general assumption on f compared to A3, which is because they only have two blocks, corresponding to set $p = 1$. Our assumption on f can also be weakened if we assume $p = 1$. Nonetheless, 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, A3 and A5 can probably be further weakened, A4 and A2 are essential in the convergence of nonconvex ADMM and cannot be completely dropped in general. In [8], their divergence example is

$$\underset{x_1, x_2, y}{\text{minimize}} \quad 0 \tag{3.30a}$$

$$\text{subject to} \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} x_1 + \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} x_2 + \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} y = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{3.30b}$$

Another related example is shown in [26, Example 7].

$$\underset{x_1,x_2,y}{\text{minimize}} \;\; \iota_{S_1}(x_1)+\iota_{S_2}(x_2) \tag{3.31a}$$

$$\text{subject to } x_1 = y \tag{3.31b}$$

$$x_2 = y, \tag{3.31c}$$

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 A3-A5 but fail to satisfy A2. Without A2, ADMM is generally not capable to find a feasible point at all, let alone a stationary point. Therefore, A2 is indispensable.

To see the necessity of A4 (the smoothness of h), consider another divergence example

$$\underset{x,y}{\text{minimize}} \;\; -|x|+|y| \tag{3.32a}$$

$$\text{subject to } x = y,\; x \in [-1,1]. \tag{3.32b}$$

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 A4. Without the smoothness of h, w^k cannot be controlled by y^k anymore. Therefore, A4 is also indispensable.

4 Applications

In this section, we apply the developed convergence results to several well-known applications.

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

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{p} l_i(A_i x - b_i) + r(x), \tag{4.1}$$

where $A_i \in \mathbb{R}^{m_i \times n}$, $b_i \in \mathbb{R}^{m_i}$ and $x \in \mathbb{R}^n$. The first term controls the fitting error while the second term is the regularizer. Examples of the fitting measure l_i include least squares, logistic functions, and other smooth functions. The regularizers can be some sparse inducing functions [9,3,66,67,64] such as MCP, SCAD, ℓ_q etc. Take LASSO as an example,

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

The first term $\|y - \mathbf{A}x\|^2$, where $\mathbf{A}$ stacks all A_i, measures how the linear model $\mathbf{A}x$ fits into y and the second term $\|x\|_1$ aims to obtain a sparse x.

In order to solve (4.1) efficiently, we reformulate it as

$$\begin{aligned} \underset{x,\{z_i\}}{\text{minimize}} \quad & r(x) + \sum_{i=1}^{p} l_i(A_i z_i - b_i), \\ \text{subject to} \quad & x = z_i, \ \forall i = 1, \ldots, p. \end{aligned} \tag{4.2}$$

Algorithm 2 gives the standard ADMM algorithm for this problem.

Algorithm 2 ADMM for (4.2)

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 2.1, we have the corollary.

Corollary 1 *Let $r(x) = \|x\|_q^q = \sum_i |x_i|^q$, $0 < q \le 1$ 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 $(x^k, \mathbf{z}^k, \mathbf{w}^k)$ in the Algorithm 2 has limit points and all of its limit points are stationary points of the augmented Lagrangian L_β.

Proof Rewrite the optimization to a standard form, we have

$$\underset{x,\{z_i\}}{\text{minimize}} \quad r(x) + \sum_{i=1}^{p} l_i(A_i z_i - b_i), \tag{4.3a}$$

$$\text{subject to} \quad Ex + \mathbf{I}_{np} z = 0. \tag{4.3b}$$

where $E = -(\mathbf{I}_n, \ldots, \mathbf{I}_n)^T$, $\mathbf{I}_{np}$ is the identity matrix, and $z = (z_1, \ldots, z_p)^T$. Fitting (4.3) to the standard form (2.2), there are two blocks (x, 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 whether A1–A5 are satisfied. A1 holds because of the coercivity assumption i). A2 holds because $B = \mathbf{I}_{np}$. A4 holds because of the smoothness ii). A5 holds because E and $\mathbf{I}_{np}$ both have full column ranks. Hence, it remains to verify A3, in particular, showing that $r(x) = \sum_i |x_i|^q$ is restricted prox-regular. We only do this for the nonconvex case $0 < q < 1$. The set of general subgradients of r is

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

For any two positive constants C and M, take $\gamma = \max(\frac{4(pC^q+MC)}{c^2}, q(1-q)c^{q-2})$, where $c \triangleq \left(\frac{M}{q}\right)^{\frac{1}{q-1}}$. The exclusion set S_M contains the set $\{x| \min_{x_i \neq 0} |x_i| \leq c\}$. For any two points $z_1, z_2 \in \mathbb{B}(0, C)/S_M$, if $\|z_1 - z_2\| \leq c$, then $\mathrm{supp}(z_1) = \mathrm{supp}(z_2)$ and $\|z_1\|_0 = \|z_2\|_0$ (where $\mathrm{supp}(z)$ denotes the index set of all non-zero elements of z and $\|z\|_0$ denotes the cardinality of $\mathrm{supp}(z)$). Note that $r(x)$ is twice differentiable along the line segment connecting z_1 and z_2, and the second order derivative is no bigger than $q(1-q)c^{q-2}$, so we have

$$\|z_1\|_q^q - \|z_2\|_q^q - \langle d, z_1 - z_2 \rangle \geq -\frac{q(1-q)}{2} c^{q-2} \|z_1 - z_2\|^2. \tag{4.4}$$

If $\|z_1 - z_2\| > c$, then we have

$$\|z_1\|_q^q - \|z_2\|_q^q - \langle d, z_1 - z_2 \rangle \geq -(2pC^q + 2MC) \geq -\frac{2pC^q + 2MC}{c^2} \|z_1 - z_2\|^2. \tag{4.5}$$

Combining (4.4) and (4.5) yields the result. This verifies A3 and completes the proof. □

B) Minimization on compact smooth manifolds

Compact smooth manifolds such as spherical manifolds S^{n-1}, Stiefel manifolds (the set of p orthonormal vectors $x_1, \ldots, x_p \in \mathbb{R}^n$, $p \leq n$) and Grassman 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 [58,25,**?**]. A simple example is:

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & J(x), \\ \text{subject to} \quad & \|x\|^2 = 1, \end{aligned} \tag{4.6}$$

More generally, let S be a compact smooth manifold. We consider the problem

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & J(x), \\ \text{subject to} \quad & x \in S, \end{aligned} \tag{4.7}$$

which can be rewritten to the following form:

$$\underset{x,y}{\text{minimize}} \quad \iota_S(x) + J(y), \tag{4.8a}$$

$$\text{subject to} \quad x - y = 0, \tag{4.8b}$$

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 2.1, 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) in the Algorithm 3 has at least one limit point, and each limit point is a stationary point of the augmented Lagrangian L_β.*

Algorithm 3 ADMM for minimization on a compact smooth manifold (4.8)

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

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

Assumption A1 holds because the feasible set is a bounded set and J is lower bounded on the feasible set. A2 and A5 hold because both A and B are identity matrices. A4 holds because J is Lipschitz differentiable. To verify A3, 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 $\|y - x\| \le \delta$, 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 $d \in \partial\iota_S(x)$, $\|d\| \le M$ and $\|x - y\| \le \delta$,

$$\begin{aligned}
\iota_S(y) - \iota_S(x) - \langle d, y - x\rangle &= - \langle d, h_\eta(\alpha_y) - h_\eta(\alpha_x)\rangle \\
&= - \langle d, h_\eta(\alpha_y) - h_\eta(\alpha_x) - J_{h_\eta}(\alpha_y - \alpha_x)\rangle \\
&\ge - \|d\| \cdot \gamma\|\alpha_y - \alpha_x\|^2 \\
&\ge - M\gamma C^2\|x - y\|^2, \qquad (4.9)
\end{aligned}$$

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

$$\begin{aligned}
\iota_S(y) - \iota_S(x) - \langle d, y - x\rangle &= - \langle d, y - x\rangle \\
&\ge - \|d\| \cdot \|y - x\| \\
&\ge - \frac{M}{\delta}\|y - x\|^2, \qquad (4.10)
\end{aligned}$$

where M is the maximum of $\|d\|$ over $\partial\iota_S(x)$. Combining (4.9) and (4.10) shows that ι_S is restricted prox-regular.

C) Matrix decomposition

ADMM has also been applied to solve matrix related problems, such as sparse principle component analysis (PCA) [21], matrix decomposition [46, 53], matrix completion [6], matrix recovery [36], non-negative matrix factorization [68, 63, 48] and background/foreground extraction [64].

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) move 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:

$$\underset{X,Y,Z}{\text{minimize}} \quad |||X|||_q + \sum_{i=1}^{m-1} \|Y_i - Y_{i+1}\| + |||Z|||_F^2, \tag{4.11}$$

$$\text{subject to} \quad V = X + Y + Z, \tag{4.12}$$

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 $|||\cdot|||_q$ is the Schatten-q quasi-norm ($0 < q \leq 1$):

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

where σ_i is the ith largest singular value of A.

The corresponding ADMM algorithm is given in Algorithm 4.

Algorithm 4 ADMM for (4.11)

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

Corollary 3 *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 L_β.*

Proof Let us verify Assumptions A1–A5. Assumption A1 holds because of the coercivity of $|||\cdot|||_F$ and $|||\cdot|||_q$. A2 and A5 hold because all the coefficient matrices are identity matrices. A4 holds because $|||\cdot|||_F^2$ is Lipschitz differentiable. It remains to verify A3, which amounts to showing 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.

First, let us characterize S_M. Consider the singular value decomposition (SVD) of A

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

where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, $\Sigma \in \mathbb{R}^{n\times n}$ is diagonal, and $\Sigma_1 \in \mathbb{R}^{K\times K}$ is also diagonal whose diagonal elements are $\sigma_i(A)$, $i = 1, \ldots, K$. Note that U, U_1, U_2 and V, V_1, V_2 might not be unique. Define $D \in \mathbb{R}^{K\times K}$ as a diagonal matrix whose ith diagonal element is $d_i = q\sigma_i(A)^{q-1}$. Then the general subgradient of $|||A|||_q^q$ [37,55] is

$$\partial |||A|||_q^q = U_1 D V_1^T + \text{conv}\{U_2 \Gamma V_2^T \,\big|\, A = U_1\Sigma_1 V_1^T, \Gamma \text{ is an arbitrary diagonal matrix}\}.$$

Take any U_1, V_1 satisfying $A = U_1 \Sigma_1 V_1^T$. For any $X \in \partial |||A|||_q^q$, since

$$\left\langle X, U_1 V_1^T \right\rangle = tr(U_1^T X V_1) = tr(D) > \min_{\sigma_i > 0} q\sigma_i^{q-1},$$

we have $\sqrt{n}|||X|||_F = |||X|||_F \left|\left|\left| U_1 V_1^T \right|\right|\right|_F \geq \left\langle X, U_1 V_1^T \right\rangle \geq \min_{\sigma_i>0} q\sigma_i^{q-1}$. This shows that S_M contains $\{A | \min_{\sigma_i>0} \sigma_i(A) > (\sqrt{n}M/q)^{1/(q-1)}\}$, where the smallest nonzero singular value of A is significantly larger than zero.

For any positive parameter M and $P > 0$. For any B, A such that $|||B|||_F < P, |||A|||_F < P$, $A, B \notin S_M$, $T \in \partial |||A|||_q^q, |||T|||_F \leq M$, we shall show that there exists $\gamma > 0$ such that

$$|||B|||_q^q - |||A|||_q^q - \left\langle T, B - A \right\rangle \geq -\frac{\gamma}{2} |||A - B|||_F^2. \tag{4.13}$$

Based on [12, Theorem 4.1], for any bounded A where $A \notin S_M$, let $F(A) = U_1 f(\Sigma_1) V_1^T$ where $f(\Sigma_1) = \text{diag}(q\sigma_1(A)^{q-1}, \dots, q\sigma_1(A)^{q-1})$ (Define $0^{q-1} = 0$). $F(A) \in \partial |||A|||_q^q$ is Lipschitz continuous, i.e., there exists ϵ_0 and $L > 0$, for any two matrices A and B, $|||A - B|||_F < \epsilon_0$ implies $|||F(A) - F(B)|||_F \leq L|||A - B|||_F$. If $|||B - A||| > \epsilon_0$, by the Weyl Theorem, we have

$$|||B|||_q^q - |||A|||_q^q \geq -n \max_i |\sigma_i(B) - \sigma_i(A)|^q \geq -n\epsilon_0^{q-2} |||B - A|||_F^2.$$

Furthermore, $\langle T, B - A \rangle \geq -M\epsilon_0^{-1} |||B - A|||_F^2$. Combining these two terms, we have

$$|||B|||_q^q - |||A|||_q^q - \left\langle T, B - A \right\rangle \geq -\frac{M\epsilon_0^{-1} + n\epsilon_0^{q-2}}{2} |||A - B|||_F^2. \tag{4.14}$$

For any B, A such that $|||B - A|||_F < \epsilon_0$, suppose $T \in \partial |||A|||_q^q$. Then, $T = F(A) + \sum_i \lambda_i U_{2i} \Gamma_i V_{2i}^T$, where $\sum_i \lambda_i = 1$ and $\lambda_i > 0$, Γ_i are all diagonal matrices whose diagonal elements are smaller than M. Because $F(A) \in \partial |||A|||_q^q$ is Lipschitz continuous, we have

$$|||B|||_q^q - |||A|||_q^q - \left\langle F(A), B - A \right\rangle \geq -\frac{L}{2} |||B - A|||_F^2. \tag{4.15}$$

In addition, because $\left|\left|\left| U_{2i}^T U_B \right|\right|\right|_F < |||B - A|||_F / \epsilon_0$ and $\left|\left|\left| V_{2i}^T V_B \right|\right|\right|_F < |||B - A|||_F / \epsilon_0$ (see [27]),

$$\begin{aligned}
\left\langle T - F(A), B - A \right\rangle &= \sum_i \lambda_i \left\langle U_{2i} \Gamma_i V_{2i}^T, B - A \right\rangle \\
&= \sum_i \lambda_i \left\langle \Gamma_i, U_{2i}^T U_B \Sigma_B V_B^T V_{2i} \right\rangle \\
&= \sum_i \lambda_i \left\langle \Gamma_i, U_{2i}^T U_B \Sigma_B V_B^T V_{2i} \right\rangle \\
&\geq -\frac{M^2}{2\epsilon_0} |||B - A|||_F^2.
\end{aligned} \tag{4.16}$$

Combining (4.14) and (4.16), then we finally get (4.13). □

5 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.

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. The underlying technique identifies an exclusion set where the sequence does not enter after finitely many iterations.

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.

The well known divergence example proposed in [8] satisfies all the assumptions except Assumption A2. The smoothness assumption A4 is also essential for the convergence of ADMM.

6 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 (***Proposition 1.1***) Define the augmented Lagrangian function to be

$$L_\beta(x,y,w) = x^2 - y^2 + w(x-y) + \beta\|x-y\|^2.$$

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

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

On one hand,

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

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

$$\begin{aligned}
\sup_{w\in\mathbb{R}} \inf_{x\in[-1,1],y\in\mathbb{R}} L_\beta(x,y,w) &= \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1],t\in\mathbb{R}} t(2x-t)+wt+\beta t^2 && (6.1)\\
&= \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1],t\in\mathbb{R}} (w+2x)t+(\beta-1)t^2 && (6.2)\\
&= \sup_{w\in\mathbb{R}} \inf_{x\in[-1,1]} -\frac{(w+2x)^2}{4(\beta-1)} && (6.3)\\
&= \sup_{w\in\mathbb{R}} -\frac{\max((w-2)^2,(w+2)^2)}{4(\beta-1)} && (6.4)\\
&= -\frac{1}{\beta-1}. && (6.5)
\end{aligned}$$

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

Then let us show ALM does not converge. ALM consists of two steps

1) $(x^{k+1}, y^{k+1}) = \text{argmin} L_\beta(x,y,w^k)$,
2) $w^{k+1} = w^k + \tau(x^{k+1} - y^{k+1})$.

The only difference between this method and ADMM is that this method update x, y simultaneously. Since $(x^{k+1} - y^{k+1}) \in \partial \inf_{x,y} L_\beta(x,y,w^k)$, and we already know

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

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 $\inf_{x,y} L(x,y,0)$ 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 2.1. Another way is to check the iterates directly. The ADMM iterates are

$$\begin{aligned}
x^{k+1} &= \max(-1, \min(1, \frac{\beta}{\beta+1}(y^k - \frac{w^k}{2\beta}))) && (6.6)\\
y^{k+1} &= \frac{\beta}{\beta-1}\left(x^{k+1} + \frac{w^k}{2\beta}\right) && (6.7)\\
w^{k+1} &= w^k + 2\beta(x^{k+1} - y^{k+1}). && (6.8)
\end{aligned}$$

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

$$\begin{aligned}
x^{k+1} &= \max(-1, \min(1, y^k)) && (6.9)\\
y^{k+1} &= \frac{1}{\beta-1}\left(\beta x^{k+1} - y^k\right). && (6.10)
\end{aligned}$$

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, ADMM converges.

*Proof (**Theorem 2.2**)* Similar to the proof of Theorem 2.1, we only need to verify P1-P3 in Proposition 3.1.
Proof of P2: Similar to Lemma 3.4 and Lemma 3.5, we have

$$\begin{aligned} &L_\beta(\mathbf{x}^k, y^k, w^k) - L_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \\ \geq &-\frac{1}{\beta}\|w^k - w^{k+1}\|^2 + \sum_{i=1}^{p} \frac{\beta - L_\phi \bar{M}}{2}\|A_i x_i^k - A_i x_i^{k+1}\|^2 + \frac{\beta - L_\phi \bar{M}}{2}\|By^k - By^{k+1}\|^2. \end{aligned} \tag{6.11}$$

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

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

where $C_1 = \min_{\lambda_i \neq 0} \lambda_i(B^T B)^{-1/2}$, $\lambda_i(B^T B)$ is ith eigenvalue of $B^T B$, and L_ϕ is the Lipschitz constant for ϕ. Therefore, we have

$$\begin{aligned} &L_\beta(\mathbf{x}^k, y^k, w^k) - L_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1}) \\ \geq &\left(\frac{\beta - L_\phi \bar{M}}{2} - \frac{C L_\phi \bar{M}}{\beta} \right) \sum_{i=1}^{p} \|A_i x_i^k - A_i x_i^{k+1}\|^2 + \left(\frac{\beta - L_\phi \bar{M}}{2} - \frac{C_1 L_\phi \bar{M}}{\beta} \right) \|By^k - By^{k+1}\|^2. \end{aligned} \tag{6.12}$$

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 $L_\beta(\mathbf{x}^k, y^k, w^k) \geq L_\beta(\mathbf{x}^{k+1}, y^{k+1}, w^{k+1})$, which means $L_\beta(\mathbf{x}^k, y^k, w^k)$ decreases monotonically. There exists y' such that $\mathbf{A}\mathbf{x}^k + By' = 0$ and $y' = H(By')$. In order to show $L_\beta(\mathbf{x}^k, y^k, w^k)$ is lower bounded, we apply B1-B4 to get

$$L_\beta(\mathbf{x}^k, y^k, w^k) = \phi(\mathbf{x}^k, y^k) + \left\langle w^k, \sum_{i=1}^{p} A_i x_i^k + By^k \right\rangle + \frac{\beta}{2}\|\sum_{i=1}^{p} A_i x_i^k + By^k\|^2 \tag{6.13}$$

$$= \phi(\mathbf{x}^k, y^k) + \left\langle \partial_y \phi(\mathbf{x}^k, y^k), y' - y^k \right\rangle + \frac{\beta}{2}\|By^k - By'\|^2 \tag{6.14}$$

$$\begin{aligned} &\geq \phi(\mathbf{x}^k, y') + \frac{\beta}{4}\|\sum_{i=1}^{p} A_i x_i^k + By^k\|^2 \\ &> -\infty. \end{aligned}$$

This shows that $L(\mathbf{x}^k, y^k, w^k)$ is lower bounded. If we view (6.13) 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 $L_\beta(\mathbf{x}^0, y^0, w^0)$. Then B1 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, forwardbackward splitting, and regularized gaussseidel methods. Mathematical Programming **137**(1-2), 91–129 (2013)
2. Auslender, A.: Optimisation: méthodes numériques. Masson (1976)
3. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends® in Machine Learning **4**(1), 1–106 (2012)
4. 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)
5. Bouaziz, S., Tagliasacchi, A., Pauly, M.: Sparse iterative closest point. In: Computer graphics forum, vol. 32, pp. 113–123. Wiley Online Library (2013)
6. 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)
7. 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)
8. 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)
9. 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)
10. Davis, D.: Convergence rate analysis of primal-dual splitting schemes. arXiv preprint arXiv:1408.4419 (2014). URL `http://arxiv.org/abs/1408.4419`
11. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. arXiv preprint arXiv:1406.4834 (2014). URL `http://arxiv.org/abs/1406.4834`
12. Ding, C., Sun, D., Sun, J., Toh, K.C.: Spectral operators of matrices. arXiv preprint arXiv:1401.2269 (2014)
13. 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)
14. Glowinski, R.: Numerical methods for nonlinear variational problems. Springer series in computational physics. Springer-Verlag, New York (1984)
15. 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)
16. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. Operations Research Letters **26**(3), 127–136 (2000)
17. Han, S.: A successive projection method. Mathematical Programming **40**(1), 1–14 (1988)
18. 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). DOI 10.1137/110836936. URL `http://epubs.siam.org/doi/abs/10.1137/110836936`
19. Hildreth, C.: A quadratic programming procedure. Naval Research Logistics Quarterly **4**(1), 79–85 (1957)
20. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. ArXiv preprint arXiv:1410.1390 (2014)
21. Hu, Y., Chi, E., Allen, G.I.: Admm algorithmic regularization paths for sparse statistical machine learning. arXiv preprint arXiv:1504.06637 (2015)
22. Ivanov, M., Zlateva, N.: Abstract subdifferential calculus and semi-convex functions. Serdica Mathematical Journal **23**(1), 35p–58p (1997)
23. Jiang, B., Ma, S., Zhang, S.: Alternating direction method of multipliers for real and complex polynomial optimization models. Optimization **63**(6), 883–898 (2014)
24. Kryštof, V., Zajíček, L.: Differences of two semiconvex functions on the real line. preprint (2015)
25. Lai, R., Osher, S.: A splitting method for orthogonality constrained problems. Journal of Scientific Computing **58**(2), 431–449 (2014)
26. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. ArXiv e-prints arXiv:1407.0753, to appear in SIOPT (2014)
27. Li, R.C., Stewart, G.: A new relative perturbation theorem for singular subspaces. Linear Algebra and its Applications **313**(1), 41–51 (2000)
28. Liavas, A.P., Sidiropoulos, N.D.: Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers. arXiv preprint arXiv:1409.2383 (2014)
29. Łojasiewicz, S.: Sur la géométrie semi-et sous-analytique. Ann. Inst. Fourier (Grenoble) **43**(5), 1575–1595 (1993)

30. Magnússon, S., Weeraddana, P.C., Rabbat, M.G., Fischione, C.: On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. arXiv preprint arXiv:1409.8033 (2014)
31. Mangasarian, O., Leone, R.: Parallel successive overrelaxation methods for symmetric linear complementarity problems and linear programs. Journal of Optimization Theory and Applications **54**(3), 437–446 (1987)
32. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM Journal on Control and Optimization **15**(6), 959–972 (1977)
33. 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)
34. 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)
35. Ortega, J., Rheinboldt, W.: Iterative solution of nonlinear equations in several variables. Academic Press (1970)
36. 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)
37. Papadopoulo, T., Lourakis, M.I.: Estimating the jacobian of the singular value decomposition: Theory and applications. In: Computer Vision-ECCV 2000, pp. 554–570. Springer (2000)
38. Poliquin, R., Rockafellar, R.: Prox-regular functions in variational analysis. Transactions of the American Mathematical Society **348**(5), 1805–1838 (1996)
39. Powell, M.: On search directions for minimization algorithms. Mathematical Programming **4**(1), 193–201 (1973)
40. Razaviyayn, M., Hong, M., Luo, Z.Q., Pang, J.S.: Parallel successive convex approximation for nonsmooth nonconvex optimization. In: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 1440–1448. Curran Associates, Inc. (2014)
41. Razaviyayn, M., Sanjabi, M., Luo, Z.Q.: A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. arXiv preprint arXiv:1307.4457 (2013)
42. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis (2009)
43. 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)
44. Sargent, R., Sebastian, D.: On the convergence of sequential minimization algorithms. Journal of Optimization Theory and Applications **12**(6), 567–575 (1973)
45. 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)
46. 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)
47. Solodov, M.V.: On the convergence of constrained parallel variable distribution algorithms. SIAM Journal on Optimization **8**(1), 187–196 (1998)
48. 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)
49. Tseng, P.: Dual coordinate ascent methods for non-strictly convex minimization. Mathematical Programming **59**(1), 231–247 (1993)
50. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications **109**(3), 475–494 (2001)
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. Warga, J.: Minimizing certain convex functions. Journal of the Society for Industrial and Applied Mathematics **11**(3), 588–593 (1963)
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.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. arXiv preprint arXiv:1410.1386 (2014)
62. Xu, Y., Yin, W.: Block stochastic gradient iteration for convex and nonconvex optimization. SIAM Journal on Optimization **25**(3), 1686–1716 (2015). DOI 10.1137/140983938. URL `http://epubs.siam.org/doi/10.1137/140983938`
63. 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)
64. Yang, L., Pong, T.K., Chen, X.: Alternating direction method of multipliers for nonconvex background/foreground extraction. arXiv preprint arXiv:1506.07029 (2015)
65. 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)
66. Zeng, J., Lin, S., Xu, Z.: Sparse regularization: Convergence of iterative jumping thresholding algorithm. arXiv preprint arXiv:1402.5744 (2014)
67. Zeng, J., Peng, Z., Lin, S.: A Gauss-Seidel iterative thresholding algorithm for lq regularized least squares regression. arXiv preprint arXiv:1507.03173 (2015)
68. Zhang, Y.: An alternating direction algorithm for nonnegative matrix factorization. preprint (2010)