SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2018, Vol. 61 122101:1–122101:12
https://doi.org/10.1007/s11432-017-9367-6

Convergence of multi-block Bregman ADMM for nonconvex composite problems

Fenghui WANG[1,2], Wenfei CAO[1,3] & Zongben XU[1*]

[1]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an* 710049*, China;*
[2]*Department of Mathematics, Luoyang Normal University, Luoyang* 471022*, China;*
[3]*School of Mathematics and Information Science, Shaanxi Normal University, Xi'an* 710119*, China*

Received 12 September 2017/Revised 5 December 2017/Accepted 26 February 2018/Published online 21 June 2018

Abstract The alternating direction method with multipliers (ADMM) is one of the most powerful and successful methods for solving various composite problems. The convergence of the conventional ADMM (i.e., 2-block) for convex objective functions has been stated for a long time, and its convergence for nonconvex objective functions has, however, been established very recently. The multi-block ADMM, a natural extension of ADMM, is a widely used scheme and has also been found very useful in solving various nonconvex optimization problems. It is thus expected to establish the convergence of the multi-block ADMM under nonconvex frameworks. In this paper, we first justify the convergence of 3-block Bregman ADMM. We next extend these results to the N-block case ($N \geqslant 3$), which underlines the feasibility of multi-block ADMM applications in nonconvex settings. Finally, we present a simulation study and a real-world application to support the correctness of the obtained theoretical assertions.

Keywords nonconvex regularization, alternating direction method, subanalytic function, K-L inequality, Bregman distance

Citation Wang F H, Cao W F, Xu Z B. Convergence of multi-block Bregman ADMM for nonconvex composite problems. Sci China Inf Sci, 2018, 61(12): 122101, https://doi.org/10.1007/s11432-017-9367-6

1 Introduction

Many problems arising in the fields of signal & image processing and machine learning involve finding a minimizer of the sum of N ($N \geqslant 2$) functions with linear equality constraint [1]. If $N = 2$, the problem then consists of solving

$$\min\ f(x) + g(y) \quad \text{s.t. } Ax + By = 0, \tag{1}$$

where $A \in \mathbb{R}^{m\times n_1}$ and $B \in \mathbb{R}^{m\times n_2}$ are given matrices, $f : \mathbb{R}^{n_1} \to \mathbb{R}$ and $g : \mathbb{R}^{n_2} \to \mathbb{R}$ are proper lower semicontinuous functions. Because of its separable structure, problem (1) can be efficiently solved by ADMM, namely, through the procedure:

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{R}^{n_1}} L_\alpha(x, y^k, p^k), \\ y^{k+1} = \arg\min\limits_{y\in\mathbb{R}^{n_2}} L_\alpha(x^{k+1}, y, p^k), \\ p^{k+1} = p^k + \alpha(Ax^{k+1} + By^{k+1}), \end{cases} \tag{2}$$

[*] Corresponding author (email: zbxu@mail.xjtu.edu.cn)

where α is a penalty parameter and

$$L_\alpha(x,y,p) := f(x) + g(y) + \langle p, Ax+By\rangle + \frac{\alpha}{2}\|Ax+By\|^2$$

is the associated augmented Lagrangian function with multiplier p. So far, various variants of the conventional ADMM have been suggested. Among such varieties, Bregman ADMM (BADMM) is designed to improve the performance of procedure (2) [2]. More specifically, BADMM takes the following iterative form:

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{R}^{n_1}} L_\alpha(x,y^k,p^k) + \triangle_\phi(x,x^k), \\ y^{k+1} = \arg\min\limits_{y\in\mathbb{R}^{n_2}} L_\alpha(x^{k+1},y,p^k) + \triangle_\psi(y,y^k), \\ p^{k+1} = p^k + \alpha(Ax^{k+1}+By^{k+1}), \end{cases} \tag{3}$$

where $\triangle_\phi$ and $\triangle_\psi$ are the Bregman distance with respect to functions ϕ and ψ, respectively. ADMM was introduced in the early 1970s, and its convergence properties for convex objective functions have been extensively studied [3,4]. It has been shown that ADMM can converge at a sublinear rate of $\mathcal{O}(1/k)$ [5], and $\mathcal{O}(1/k^2)$ for the accelerated version [6]. The convergence of BADMM for convex objective functions has been examined in [2].

Recently, there has been an increasing interest in the study of ADMM for nonconvex objective functions. On one hand, the ADMM algorithm is highly successful in solving various nonconvex examples ranging from nonnegative matrix factorization, distributed matrix factorization, distributed clustering, sparse zero variance discriminant analysis, tensor decomposition, to matrix completion (see [7–9]). On the other hand, the convergence analysis of nonconvex ADMM is generally very difficult, due to the failure of the Fejér monotonicity of iterates. Very recently, the convergence of ADMM as well as BADMM for nonconvex objective functions has been established in [10–12].

We now consider the 3-block composite optimization problem:

$$\min\ f(x)+g(y)+h(z) \quad \text{s.t. } Ax+By+Cz=0, \tag{4}$$

where $A\in\mathbb{R}^{m\times n_1}$, $B\in\mathbb{R}^{m\times n_2}$ and $C\in\mathbb{R}^{m\times n_3}$ are given matrices, $f:\mathbb{R}^{n_1}\to\mathbb{R}$, $g:\mathbb{R}^{n_2}\to\mathbb{R}$ are proper lower semicontinuous functions, and $h:\mathbb{R}^{n_3}\to\mathbb{R}$ is a continuously differentiable function. To solve this problem, it is thus natural to extend (2) to the following form:

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathbb{R}^{n_1}} L_\alpha(x,y^k,z^k,p^k), \\ y^{k+1} = \arg\min\limits_{y\in\mathbb{R}^{n_2}} L_\alpha(x^{k+1},y,z^k,p^k), \\ z^{k+1} = \arg\min\limits_{z\in\mathbb{R}^{n_3}} L_\alpha(x^{k+1},y^{k+1},z,p^k), \\ p^{k+1} = p^k + \alpha(Ax^{k+1}+By^{k+1}+Cz^{k+1}), \end{cases} \tag{5}$$

where the augmented Lagrangian function $L_\alpha:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\mathbb{R}^{n_3}\times\mathbb{R}^m\to\mathbb{R}$ is defined by

$$L_\alpha(x,y,z,p) := f(x)+g(y)+h(z)+\langle p, Ax+By+Cz\rangle + \frac{\alpha}{2}\|Ax+By+Cz\|^2. \tag{6}$$

However, as shown in [13], the 3-block ADMM (5) does not necessarily converge in general even under the convex frameworks. To guarantee its global convergence, some restrictive conditions are required; for example, the strong convexity condition of all objective functions [14], or at least one function being strongly convex [15,16].

The purpose of the present study is to examine convergence of ADMM with N blocks for nonconvex objective functions. Following the idea of (3), we first propose 3-block BADMM for solving problem (4), and establish its global convergence for some nonconvex functions. Next, we extend the convergence result to the N-block case ($N \geqslant 3$), which underlines the feasibility of multi-block ADMM applications in nonconvex settings. Finally we present a simulation study and a real-world application to support the correctness of the obtained theoretical assertions.

2 Preliminaries

In what follows, $\mathbb{R}^n$ will stand for the n-dimensional Euclidean space,

$$\langle x, y\rangle = x^{\mathrm{T}}y = \sum_{i=1}^{n} x_i y_i,\ \ \|x\| = \sqrt{\langle x, x\rangle},$$

where $x, y \in \mathbb{R}^n$ and T stands for the transpose operation. For convenience, we fix the following notations:

$$\begin{aligned} &u^k = (x^k, y^k, z^k), \quad w^k = (x^k, y^k, z^k, p^k), \quad \hat{w}^k = (x^k, y^k, z^k, p^k, z^{k-1}),\\ &\|w\| = (\|x\|^2 + \|y\|^2 + \|z\|^2 + \|p\|^2)^{1/2}, \quad \|w\|_1 = \|x\| + \|y\| + \|z\| + \|p\|. \end{aligned}$$

2.1 Subdifferentials

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, we denote by dom$f$ the domain of f, namely, dom$f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$. A function f is said to be proper if dom$f \neq \emptyset$; lower semicontinuous at x_0 if $\liminf_{x\to x_0} f(x) \geqslant f(x_0)$. If f is lower semicontinuous at every point of its domain of definition, then it is simply called a lower semicontinuous function.

Definition 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper lower semi-continuous function.

(i) Given $x \in$ dom$f$, the Fréchet subdifferential of f at x, written by $\widehat{\partial} f(x)$, is the set of all elements $u \in \mathbb{R}^n$ which satisfy

$$\lim_{y\neq x}\inf_{y\to x} \frac{f(y) - f(x) - \langle u, y - x\rangle}{\|x - y\|} \geqslant 0.$$

(ii) The limiting subdifferential, or simply subdifferential, of f at x, written by $\partial f(x)$, is defined as

$$\partial f(x) = \left\{ u \in \mathbb{R}^n : \exists x^k \to x, f(x^k) \to f(x), u^k \in \widehat{\partial} f(x^k) \to u, k \to \infty \right\}.$$

(iii) A stationary point of f is a point x^* in the domain of f satisfying $0 \in \partial f(x^*)$.

(iv) f is said to be L-Lipschitz continuous if $\|f(x) - f(y)\| \leqslant L\|x - y\|$, for any $x, y \in$ dom$f$.

Definition 2. An element $w^* := (x^*, y^*, z^*, p^*)$ is called a stationary point of the Lagrangian function L_α defined as in (6) if it satisfies

$$\begin{cases} A^{\mathrm{T}}p^* \in -\partial f(x^*),\ B^{\mathrm{T}}p^* \in -\partial g(y^*),\\ C^{\mathrm{T}}p^* = -\nabla h(z^*),\ Ax^* + By^* + Cz^* = 0. \end{cases} \tag{7}$$

The existence of proper lower semicontinuous functions and properties of subdifferential can be seen from [17]. We particularly collect some basic properties of the subdifferential.

Proposition 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be proper lower semi-continuous functions. Then the following holds:

(i) $\widehat{\partial} f(x) \subset \partial f(x)$ for each $x \in \mathbb{R}^n$. Moreover, the first set is closed and convex, while the second is closed, and not necessarily convex.

(ii) Let (u^k, x^k) be sequences such that $x^k \to x, u^k \to u, f(x^k) \to f(x)$ and $u^k \in \partial f(x^k)$. Then $u \in \partial f(x)$.

(iii) Fermat's rule: if $x_0 \in \mathbb{R}^n$ is a local minimizer of f, then x_0 is a stationary point of f, that is, $0 \in \partial f(x_0)$.

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

2.2 Kurdyka-Łojasiewicz inequality

The Kurdyka-Łojasiewicz (K-L) inequality was first introduced by Łojasiewicz [18] for real analytic functions, and then was extended by Kurdyka [19] to smooth functions whose graph belongs to an o-minimal structure.

Definition 3 (K-L inequality)**.** A function $f:\mathbb{R}^n\to\mathbb{R}$ is said to have the K-L property at $\tilde{x}$ if there exists $\eta>0,\delta>0,\varphi\in\mathscr{A}_\eta$, such that for all $x\in\mathcal{O}(\tilde{x},\delta)\cap\{x:f(\tilde{x})<f(x)<f(\tilde{x})+\eta\}$,

$$\varphi'(f(x)-f(\tilde{x}))\mathrm{dist}(0,\partial f(x))\geqslant 1,$$

where $\mathrm{dist}(\tilde{x},\partial f(x)):=\inf\{\|\tilde{x}-y\|:y\in\partial f(x)\}$, and $\mathscr{A}_\eta$ stands for the class of functions $\varphi:[0,\eta)\to\mathbb{R}^+$ with the properties: (i) φ is continuous on $[0,\eta)$; (ii) φ is smooth concave on $(0,\eta)$; (iii) $\varphi(0)=0,\varphi'(x)>0,\forall x\in(0,\eta)$.

Let Φ be a proper lower semicontinuous function, and a,b be two fixed positive constants. In the sequel, we consider a sequence $\{x^k\}$ satisfying the following conditions:

(H1) For each $k\in\mathbb{N}$, $\Phi(x^{k+1})\leqslant\Phi(x^k)-a\|x^k-x^{k+1}\|^2$;

(H2) For each $k\in\mathbb{N}$, $\mathrm{dist}(0,\partial\Phi(x^{k+1}))\leqslant b\|x^k-x^{k+1}\|$;

(H3) There exists a subsequence $\{x^{k_j}\}$ converging to $\tilde{x}$ such that $\Phi(x^{k_j})\to\Phi(\tilde{x})$ as $j\to\infty$.

Lemma 1 ([20])**.** Let $\{x^k\}$ be a sequence that satisfies H1–H3. If Φ has the K-L property, then the sequence $\{x^k\}$ converges to $\tilde{x}$, which is a stationary point of Φ. Moreover, the sequence $\{x^k\}$ has a finite length, i.e., $\sum_{k=1}^\infty\|x^{k+1}-x^k\|_1<\infty$.

Typical functions satisfying the K-L inequality include strongly convex functions, real analytic functions, semi-algebraic functions and subanalytic functions.

A differentiable function f is called convex if the following inequality holds for all x,y in its domain:

$$f(y)\geqslant f(x)+\langle\nabla f(x),y-x\rangle;$$

ρ-strongly convex with $\rho>0$ if the following inequality holds for all x,y in its domain:

$$f(y)\geqslant f(x)+\langle\nabla f(x),y-x\rangle+\frac{\rho}{2}\|y-x\|^2. \tag{8}$$

A subset $C\subset\mathbb{R}^n$ is said to be semi-algebraic if it can be written as

$$C=\bigcup_{j=1}^r\bigcap_{i=1}^s\{x\in\mathbb{R}^n:g_{i,j}(x)=0,h_{i,j}(x)<0\},$$

where $g_{i,j},h_{i,j}:\mathbb{R}^n\to\mathbb{R}$ are real polynomial functions. Then a function $f:\mathbb{R}^n\to\mathbb{R}$ is called semi-algebraic if its graph $\mathcal{G}(f):=\{(x,y)\in\mathbb{R}^{n+1}:f(x)=y\}$ is a semi-algebraic subset in $\mathbb{R}^{n+1}$. For example, the L_q norm $\|x\|_q:=(\sum_i|x_i|^q)^{1/q}$ with $0<q\leqslant1$, the sup-norm $\|x\|_\infty:=\max_i|x_i|$, the Euclidean norm $\|x\|$, $\|Ax-b\|_q^q$, $\|Ax-b\|$ and $\|Ax-b\|_\infty$ are all semi-algebraic functions for any matrix A.

A real function on $\mathbb{R}$ is said to be analytic if it possesses derivatives of all orders and agrees with its Taylor series in a neighborhood of every point. For a real function f on $\mathbb{R}^n$, it is said to be analytic if the function of one variable $g(t):=f(x+ty)$ is analytic for any $x,y\in\mathbb{R}^n$. It is readily seen that real polynomial functions such as quadratic functions $\|Ax-b\|^2$ are analytic. Moreover, the ε-smoothed L_q norm $\|x\|_{\varepsilon,q}:=\sum_i(x_i^2+\varepsilon)^{q/2}$ with $0<q\leqslant1$ and the logistic loss function $\log(1+\mathrm{e}^{-t})$ are all examples for real analytic functions. A subset $C\subset\mathbb{R}^n$ is said to be subanalytic if it can be written as

$$C=\bigcup_{j=1}^r\bigcap_{i=1}^s\{x\in\mathbb{R}^n:g_{i,j}(x)=0,h_{i,j}(x)<0\},$$

where $g_{i,j},h_{i,j}:\mathbb{R}^n\to\mathbb{R}$ are real analytic functions. Then a function $f:\mathbb{R}^n\to\mathbb{R}$ is called subanalytic if its graph $\mathcal{G}(f)$ is a subanalytic subset in $\mathbb{R}^{n+1}$. It is clear that both real analytic and semi-algebraic functions are subanalytic. Generally speaking, the sum of two subanalytic functions is not necessarily subanalytic. It is known, however, that for two subanalytic functions, if at least one function maps bounded sets to bounded sets, then their sum is also subanalytic, as shown in [9]. In particular, the sum of a subanalytic function and an analytic function is subanalytic. Typical subanalytic functions include: $\|Ax-b\|^2+\lambda\|y\|_q^q$; $\|Ax-b\|^2+\lambda\sum_i(y_i^2+\varepsilon)^{q/2}$; $\frac{1}{n}\sum_{i=1}^n\log(1+\exp(-c_i(a_i^{\mathrm{T}}x+b))+\lambda\|y\|_q^q$; and $\frac{1}{n}\sum_{i=1}^n\log(1+\exp(-c_i(a_i^{\mathrm{T}}x+b))+\lambda\sum_i(y_i^2+\varepsilon)^{q/2}$.

2.3 Bregman distance

The Bregman distance plays an important role in various iterative algorithms. As a generalization of squared Euclidean distance, the Bregman distance shares many similar nice properties of the Euclidean distance. However, the Bregman distance is not a real metric, since it does not satisfy the triangle inequality nor symmetry. For a convex differential function ϕ, the associated Bregman distance is defined as

$$\triangle_\phi(x,y)=\phi(x)-\phi(y)-\langle\nabla\phi(y),x-y\rangle.$$

In particular, if we let $\phi(x)=\|x\|^2$ in the above, then it is reduced to $\|x-y\|^2$, namely, the classical Euclidean distance. Moreover, if ϕ is ρ-strongly convex, it follows from (8) that

$$\triangle_\phi(x,y)\geqslant\frac{\rho}{2}\|x-y\|^2. \tag{9}$$

For more information on Bregman distance, we refer the reader to [21, 22].

3 Convergence analysis

Motivated by (3), we propose the following algorithm for solving problem (4):

$$\begin{cases} x^{k+1}=\arg\min\limits_{x\in\mathbb{R}^{n_1}} L_\alpha(x,y^k,z^k,p^k)+\Delta_{\phi_1}(x,x^k),\\ y^{k+1}=\arg\min\limits_{y\in\mathbb{R}^{n_2}} L_\alpha(x^{k+1},y,z^k,p^k)+\Delta_{\phi_2}(y,y^k),\\ z^{k+1}=\arg\min\limits_{y\in\mathbb{R}^{n_3}} L_\alpha(x^{k+1},y^{k+1},z,p^k)+\Delta_{\phi_3}(z,z^k),\\ p^{k+1}=p^k+\alpha(Ax^{k+1}+By^{k+1}+Cz^{k+1}), \end{cases} \tag{10}$$

where $\triangle_{\phi_i}$ is an appropriately chosen Bregman distance with respect to function $\phi_i, i=1,2,3$. Compared with the traditional ADMM, our algorithm has advantages both in effectiveness and efficiency. First, the global convergence of our algorithm does not require any strong convexity of the objective function. Second, a proper choice of Bregman distance will simplify the subproblems, which in turn improve the performance of the algorithm. For example, for the y-subproblem, let $g(y)=\|y\|_{1/2}^{1/2}$. In this situation, the traditional ADMM requires to solve the following optimization problem:

$$\min_{y\in\mathbb{R}^{n_2}}\|y\|_{1/2}^{1/2}+\frac{\alpha}{2}\left\|By+Ax^{k+1}+Cz^k+\frac{p^k}{\alpha}\right\|^2.$$

In general finding a solution to the above problem is not a easy task. However, if we set $\phi_2(y)=\frac{\mu}{2}\|y\|^2-\frac{\alpha}{2}\|By+Ax^{k+1}+Cz^k-p^k/\alpha\|^2$ with $\mu>\|B\|^2$ in our algorithm, then by a simple calculation the y-subproblem is transformed into minimizing:

$$\|y\|_{1/2}^{1/2}+\frac{\mu\alpha}{2}\left\|y-\left(y^k-\mu^{-1}B^{\mathrm{T}}\left(By^k+Ax^{k+1}+Cz^k+\frac{p^k}{\alpha}\right)\right)\right\|^2.$$

This problem can be easily solved since its solution has closed form [23].

In what follows, we assume:

(A1) Φ has the K-L property;

(A2) There is $\sigma>0$ such that $\sigma\|x\|^2\leqslant\|C^{\mathrm{T}}x\|^2,\forall x\in\mathbb{R}^m$;

(A3) h is continuously differentiable such that ∇h is L-Lipschitz continuous;

(A4) ϕ_i is ρ_i-strongly convex and $\nabla\phi_i$ is L_i-Lipschitz continuous for $i=1,2,3$;

(A5) The parameters are chosen so that $\alpha\rho\sigma>6(L^2+2L_3^2)$ where $\rho=\min\{\rho_1,\rho_2,\rho_3\}$.

Also, define a function $\Phi:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\mathbb{R}^{n_3}\times\mathbb{R}^m\times\mathbb{R}^{n_3}\to\mathbb{R}$ by

$$\Phi(x,y,z,p,\hat{z})=L_\alpha(x,y,z,p)+\frac{\tau}{2}\|z-\hat{z}\|^2,$$

where $\tau = 6L_3^2(\alpha\sigma)^{-1}$.

We establish a series of lemmas to support the proof of convergence of procedure (10).

Lemma 2. For each $k \in \mathbb{N}$, there exists $a > 0$ such that $\Phi(\hat{w}^{k+1}) \leqslant \Phi(\hat{w}^k) - a\|\hat{w}^{k+1} - \hat{w}^k\|^2$.

Proof. Applying Fermat's rule to the z-subproblem, we get

$$\nabla h(z^{k+1}) + C^{\mathrm{T}}p^{k+1} + \nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k) = 0. \tag{11}$$

It then follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}\|C^{\mathrm{T}}(p^{k+1} - p^k)\|^2 &= \|(\nabla h(z^{k+1}) - \nabla h(z^k)) + (\nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k)) - (\nabla\phi_3(z^k) - \nabla\phi_3(z^{k-1}))\|^2\\ &\leqslant \|\nabla h(z^{k+1}) - \nabla h(z^k)\|^2 + \|(\nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k)) - (\nabla\phi_3(z^k) - \nabla\phi_3(z^{k-1}))\|^2\\ &\quad + 2\|\nabla h(z^{k+1}) - \nabla h(z^k)\|\|(\nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k)) - (\nabla\phi_3(z^k) - \nabla\phi_3(z^{k-1}))\|\\ &\leqslant 3\|\nabla h(z^{k+1}) - \nabla h(z^k)\|^2 + \frac{3}{2}\|(\nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k)) - (\nabla\phi_3(z^k) - \nabla\phi_3(z^{k-1}))\|^2\\ &\leqslant 3L^2\|z^{k+1} - z^k\|^2 + 3(\|\nabla\phi_3(z^{k+1}) - \nabla\phi_3(z^k)\|^2 + \|\nabla\phi_3(z^k) - \nabla\phi_3(z^{k-1})\|^2)\\ &\leqslant 3(L^2 + L_3^2)\|z^{k+1} - z^k\|^2 + 3L_3^2\|z^k - z^{k-1}\|^2.\end{aligned}$$

Thus, in view of condition (A2), we get

$$\|p^{k+1} - p^k\|^2 \leqslant \frac{3(L^2 + L_3^2)}{\sigma}\|z^{k+1} - z^k\|^2 + \frac{3L_3^2}{\sigma}\|z^k - z^{k-1}\|^2. \tag{12}$$

On the other hand, it follows from (10) and (9) that

$$\begin{aligned}&L_\alpha(x^{k+1}, y^k, z^k, p^k) \leqslant L_\alpha(x^k, y^k, z^k, p^k) - \frac{\rho}{2}\|x^{k+1} - x^k\|^2,\\ &L_\alpha(x^{k+1}, y^{k+1}, z^k, p^k) \leqslant L_\alpha(x^{k+1}, y^k, z^k, p^k) - \frac{\rho}{2}\|y^{k+1} - y^k\|^2,\\ &L_\alpha(x^{k+1}, y^{k+1}, z^{k+1}, p^k) \leqslant L_\alpha(x^{k+1}, y^{k+1}, z^k, p^k) - \frac{\rho}{2}\|z^{k+1} - z^k\|^2,\\ &L_\alpha(x^{k+1}, y^{k+1}, z^{k+1}, p^{k+1}) = L_\alpha(x^{k+1}, y^{k+1}, z^{k+1}, p^k) + \frac{1}{\alpha}\|p^{k+1} - p^k\|^2,\end{aligned}$$

from which we have

$$L_\alpha(w^{k+1}) \leqslant L_\alpha(w^k) - \frac{\rho}{2}\|u^{k+1} - u^k\|^2 + \frac{1}{\alpha}\|p^{k+1} - p^k\|^2. \tag{13}$$

Adding up inequalities (12) and (13), we have

$$L_\alpha(w^{k+1}) + \frac{\tau}{2}\|z^{k+1} - z^k\|^2 \leqslant L_\alpha(\hat{w}^k) + \frac{\tau}{2}\|z^{k-1} - z^k\|^2 - a\|\hat{w}^{k+1} - \hat{w}^k\|^2,$$

where $a := (\rho/2) - 3(L^2 + 2L_3^2)(\alpha\sigma)^{-1}$ is clearly a positive real number.

Lemma 3. If $\{u^k\}$ is bounded, then $\sum_{k=1}^\infty \|w^k - w^{k+1}\|^2 < \infty$. In particular, $\{w^k\}$ is asymptotically regular, namely, $\|w^k - w^{k+1}\| \to 0$ as $k \to \infty$. Moreover, any cluster point of $\{w^k\}$ is a stationary point of the augmented Lagrangian function L_α.

Proof. In view of (11), (A2) and (A4), we have

$$\sqrt{\sigma}\|p^k\| \leqslant \|C^{\mathrm{T}}p^k\| \leqslant \|\nabla h(z^k)\| + L_3\|z^k - z^{k-1}\|.$$

Since ∇h is continuous and $\{u^k\}$ is bounded, this implies that $\{p^k\}$ is bounded, and so are $\{w^k\}$ and $\{\hat{w}^k\}$. Thus, there exists a subsequence $\{\hat{w}^{k_j}\}$ convergent to $\hat{w}^*$. By our hypothesis, the function Φ is lower semicontinuous, which leads to $\liminf_{j\to\infty} \Phi(\hat{w}^{k_j}) \geqslant \Phi(\hat{w}^*)$, so that $\Phi(\hat{w}^{k_j})$ is bounded from below. By Lemma 2, $\Phi(\hat{w}^k)$ is nonincreasing, and thus convergent. Furthermore, $\Phi(\hat{w}^k) \geqslant \Phi(\hat{w}^*)$ for each k, which by Lemma 2, yields

$$a\sum_{i=1}^k \|u^{k+1} - u^k\|^2 \leqslant \Phi(\hat{w}^1) - \Phi(\hat{w}^{k+1}) \leqslant \Phi(\hat{w}^1) - \Phi(\hat{w}^*).$$

This together with (12) implies $\sum_{k=1}^{\infty}\|w^k-w^{k+1}\|^2<\infty$; in particular $\|w^k-w^{k+1}\|\to 0$.

Let $w^*=(x^*,y^*,z^*,p^*)$ be any cluster point of $\{w^k\}$ and let $\{w^{k_j}\}$ be a subsequence of $\{w^k\}$ converging to w^*. It then follows from (10) that

$$\begin{aligned}
&p^{k_j+1}=p^{k_j}+\alpha(Ax^{k_j+1}+By^{k_j+1}+Cz^{k_j+1}),\\
&-\partial f(x^{k_j+1})\ni A^{\mathrm{T}}[p^{k_j+1}+\alpha B(y^{k_j}-y^{k_j+1})+\alpha C(z^{k_j}-z^{k_j+1})]+\nabla\phi_1(x^{k_j+1})-\nabla\phi_1(x^{k_j}),\\
&-\partial g(y^{k_j+1})\ni B^{\mathrm{T}}[p^{k_j+1}+\alpha C(z^{k_j}-z^{k_j+1})]+\nabla\phi_2(y^{k_j+1})-\nabla\phi_2(y^{k_j}),\\
&-\nabla h(z^{k_j+1})=C^{\mathrm{T}}p^{k_j+1}+\nabla\phi_3(z^{k_j+1})-\nabla\phi_3(z^{k_j}).
\end{aligned}$$

As $\nabla\phi_i, i=1,2,3$ is continuous and $\|w^k-w^{k+1}\|\to 0$, letting $j\to\infty$ above yields that w^* is a stationary point of the augmented Lagrangian function L_α.

Lemma 4. There exists $b>0$ such that $\mathrm{dist}(0,\partial\Phi(\hat{w}^{k+1}))\leqslant b\|\hat{w}^k-\hat{w}^{k+1}\|$ for each $k\in\mathbb{N}$.

Proof. By a simple calculation, we have

$$\begin{aligned}
&\partial\Phi_x(\hat{w}^{k+1})\ni\alpha A^{\mathrm{T}}B(y^{k+1}-y^k)+\alpha A^{\mathrm{T}}C(z^{k+1}-z^k)+\nabla\phi_1(x^{k+1})-\nabla\phi_1(x^k)+A^{\mathrm{T}}(p^{k+1}-p^k),\\
&\partial\Phi_y(\hat{w}^{k+1})\ni\alpha B^{\mathrm{T}}C(z^{k+1}-z^k)+B^{\mathrm{T}}(p^{k+1}-p^k)+\nabla\phi_2(y^{k+1})-\nabla\phi_2(y^k),\\
&\partial\Phi_z(\hat{w}^{k+1})=\nabla\phi_3(z^{k+1})-\nabla\phi_3(z^k)+C^{\mathrm{T}}(p^{k+1}-p^k)+\tau(z^{k+1}-z^k),\\
&\partial\Phi_p(\hat{w}^{k+1})=\frac{1}{\alpha}(p^{k+1}-p^k),\quad \partial\Phi_{\hat{z}}(\hat{w}^{k+1})=\tau(z^k-z^{k+1}).
\end{aligned}$$

As matrices A,B,C are all bounded, the above together with (12) and (A4) implies that there exists $b>0$ such that the desired inequality follows.

Theorem 1. Each bounded sequence $\{w^k\}$ generated by procedure (10) converges to a stationary point of L_α. Moreover, $\sum_{k=1}^{\infty}\|w^{k+1}-w^k\|_1<\infty$.

Proof. It is easy to see that conditions H1–H2 in Lemma 1 hold. To verify condition H3, we assume that there exists a subsequence $\{\hat{w}^{k_j}\}$ that converges to $\hat{w}^*=(x^*,y^*,z^*,p^*,z^*)$. By the lower semicontinuity of Φ, $\liminf_{j\to\infty}\Phi(\hat{w}^{k_j})\geqslant\Phi(\hat{w}^*)$. On the other hand, we have

$$\begin{aligned}
&f(x^{k_j+1})+\langle p^k,Ax^{k_j+1}\rangle+\frac{\alpha}{2}\|Ax^{k_j+1}+By^{k_j}+Cz^{k_j}\|^2+\Delta_{\phi_1}(x^{k_j+1},x^{k_j})\\
&\leqslant f(x^*)+\langle p^k,Ax^*\rangle+\frac{\alpha}{2}\|Ax^*+By^{k_j}+Cz^{k_j}\|^2+\Delta_{\phi_1}(x^*,x^{k_j}).
\end{aligned}$$

Since $\{x^k\}$ is asymptotically regular, this implies $\limsup_{j\to\infty}f(x^{k_j+1})\leqslant f(x^*)$. In a similar way, we conclude that $\limsup_{j\to\infty}g(y^{k_j+1})\leqslant g(y^*)$. Since

$$\lim_{j\to\infty}h(z^{k_j+1})=h(z^*)\quad\text{and}\quad\lim_{j\to\infty}\|z^{k_j+1}-z^{k_j}\|=0,$$

we have $\limsup_{j\to\infty}\Phi(\hat{w}^k)\leqslant\Phi(\hat{w}^*)$. Altogether, $\lim_{j\to\infty}\Phi(\hat{w}^{k_j})=\Phi(\hat{w}^*)$. Thus, condition H3 holds.

Applying Lemma 1, we conclude that $\{\hat{w}^k\}$ converges to $\hat{w}^*$, which is a stationary point of Φ. In particular, it is easy to see that $\{w^k\}$ converges to w^*. By Lemma 3, w^* is a stationary point of L_α. Moreover, $\{w^k\}$ has a finite length, i.e., $\sum_{k=1}^{\infty}\|w^{k+1}-w^k\|_1<\infty$.

Remark 1. There are various choices of Bregman distance in (10). For instance, if we let

$$\Delta_{\phi_3}(x,y)=\|x-y\|_Q^2=\langle Qx,x\rangle$$

with Q a symmetric positive definite matrix, then our first assumption A1 is satisfied whenever the objective function $f+g+h$ is subanalytic. Indeed, since the function $\|z-\hat{z}\|_Q^2$ is analytic, Φ is also subanalytic as the sum of a subanalytic function and an analytic function, which in turn implies the K-L property. Typical examples of subanalytic functions are exhibited in the previous section.

We now extend the above result to the N-block case. Thus, let us consider the following composite optimization problem:

$$\begin{aligned}
&\min\ f_1(x_1)+f_2(x_2)+\cdots+f_N(x_N)\\
&\text{s.t. } A_1x_1+A_2x_2+\cdots+A_Nx_N=0,
\end{aligned}\tag{14}$$

where $A_i \in \mathbb{R}^{m\times n_i}$, $f_i : \mathbb{R}^{n_i} \to \mathbb{R}, i = 1, 2, \ldots, N-1$ are proper lower semicontinuous functions, and $f_N : \mathbb{R}^{n_N} \to \mathbb{R}$ is a continuously differentiable function. The Lagrangian function $L_\alpha : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^m \to \mathbb{R}$ of problem (14) is defined by

$$L_\alpha(x_1, x_2, \ldots, x_N, p) = \sum_{i=1}^{N} f_i(x_i) + \sum_{i=1}^{N} \langle p, A_i x_i\rangle + \frac{\alpha}{2}\left\|\sum_{i=1}^{N} A_i x_i\right\|^2. \tag{15}$$

Accordingly, the associated algorithm takes the form:

$$\begin{cases} x_1^{k+1} &= \arg\min_{x_1\in\mathbb{R}^{n_1}} L_\alpha(x_1, x_2^k, \ldots, x_N^k, p^k) + \triangle_{\phi_1}(x_1, x_1^k),\\ x_2^{k+1} &= \arg\min_{x_2\in\mathbb{R}^{n_2}} L_\alpha(x_1^{k+1}, x_2, \ldots, x_N^k, p^k) + \triangle_{\phi_2}(x_2, x_2^k),\\ \vdots & \quad\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_N^{k+1} &= \arg\min_{x_N\in\mathbb{R}^{n_N}} L_\alpha(x_1^{k+1}, \ldots, x_{N-1}^{k+1}, x_N, p^k) + \triangle_{\phi_N}(x_N, x_N^k),\\ p^{k+1} &= p^k + \alpha(A_1x_1^{k+1} + A_2x_2^{k+1} + \cdots + A_Nx_N^{k+1}). \end{cases} \tag{16}$$

Following the idea of Theorem 1, it is not hard to extend the results to the case whenever the followings are satisfied:

(B1) Ψ has the K-L property;

(B2) there is $\sigma > 0$ such that $\sigma\|x\|^2 \leqslant \|A_N^{\mathrm{T}}x\|^2, \forall x \in \mathbb{R}^m$;

(B3) f_N is continuously differentiable such that ∇f_N is L-Lipschitz continuous;

(B4) ϕ_i is ρ_i-strongly convex and $\nabla\phi_i$ is L_i-Lipschitz continuous for $i = 1, 2, \ldots, N$;

(B5) the parameters are chosen so that $\alpha\rho\sigma > 6(L^2 + 2L_N^2)$ where $\rho = \min\{\rho_1, \rho_2, \ldots, \rho_N\}$.

Analogously, we define a function $\Psi : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N} \times \mathbb{R}^m \times \mathbb{R}^{n_N} \to \mathbb{R}$ by

$$\Psi(x_1, x_2, \ldots, x_N, p, \hat{x}_N) = L_\alpha(x_1, x_2, \ldots, x_N, p) + \frac{\tau}{2}\|x_N - \hat{x}_N\|^2,$$

where $\tau = 6L_N^2(\alpha\sigma)^{-1}$.

Theorem 2. If conditions B1–B5 are satisfied, then each bounded sequence $\{x_1^k, x_2^k, \ldots, x_N^k, p^k\}$ generated by procedure (16) converges to a stationary point of L_α defined as in (15).

4 Demonstration examples

Consider the non-convex optimization problem with 3-block variables deduced from matrix decomposition applications (see [24, 25]):

$$\min_{L,S,T} \|L\|_{\circledast} + \lambda\|S\|_1 + \frac{\mu}{2}\|T - M\|_F^2 \quad \text{s.t. } T = L + S, \tag{17}$$

where M is an $m \times n$ observation matrix, $\|L\|_{\circledast} := \sum_{i=1}^{\min(m,n)} |\sigma_i(L)|^{1/2}$, $\|S\|_1 := \sum_{i=1}^{m}\sum_{j=1}^{n} |S_{ij}|$, λ is a trade-off parameter between the low-rank term $\|L\|_{\circledast}$ and the sparse term $\|S\|_1$, and μ is a penalty parameter related to the noise level.

The augmented Lagrangian function of problem (17) is given by

$$L_\alpha(L, S, T, \Lambda) = \|L\|_{\circledast} + \lambda\|S\|_1 + \frac{\mu}{2}\|T - M\|_F^2 + \langle\Lambda, T - (L + S)\rangle + \frac{\alpha}{2}\|T - (L + S)\|_F^2, \tag{18}$$

where Λ is the Lagrangian multiplier. According to the 3-block BADMM (10), the optimization problem (17) can be solved by the following procedure:

$$\begin{cases} L^{k+1} = \arg\min_{L} L_\alpha(L, S^k, T^k, \Lambda^k) + \frac{\rho}{2}\|L - L^k\|_F^2,\\ S^{k+1} = \arg\min_{S} L_\alpha(L^{k+1}, S, T^k, \Lambda^k) + \frac{\rho}{2}\|S - S^k\|_F^2,\\ T^{k+1} = \arg\min_{T} L_\alpha(L^{k+1}, S^{k+1}, T, \Lambda^k) + \frac{\rho}{2}\|T - T^k\|_F^2,\\ \Lambda^{k+1} = \Lambda^k + \alpha\big(T^{k+1} - (L^{k+1} + S^{k+1})\big). \end{cases} \tag{19}$$

Simplifying the procedure (19), we then obtain the closed-form iterative formulas:

$$\begin{cases} L^{k+1} = \mathcal{H}\left(\dfrac{\alpha(T^k - S^k + \frac{\Lambda^k}{\alpha}) + \rho L^k}{\alpha+\rho},\ \dfrac{1}{\alpha+\rho}\right), \\ S^{k+1} = \mathcal{S}\left(\dfrac{\alpha(T^k - L^{k+1} + \frac{\Lambda^k}{\alpha}) + \rho S^k}{\alpha+\rho},\ \dfrac{\lambda}{\alpha+\rho}\right), \\ T^{k+1} = \dfrac{\mu M + \alpha(L^{k+1} + S^{k+1} - \frac{\Lambda^k}{\alpha}) + \rho T^k}{\mu+\alpha+\rho}, \\ \Lambda^{k+1} = \Lambda^k + \alpha\big(T^{k+1} - (L^{k+1} + S^{k+1})\big), \end{cases} \tag{20}$$

where $\mathcal{H}(A,\ \cdot)$ indicates the half shrinkage operator [23, 26] imposed on the singular values of A, and $\mathcal{S}(A,\ \cdot)$ indicates the well-known soft shrinkage operator imposed on the entries of A. The procedure (20) is the specification of BADMM (10) for the solution of problem (17) with functions $f(x)$, $g(y)$, $h(z)$ defined by $f(L) = \|L\|_{\circledast}$, $g(S) = \lambda\|S\|_1$, $h(T) = \frac{\mu}{2}\|T - M\|^2$ and matrices A, B, C defined by $A = I$, $B = -I$, $C = -I$ where I is the identity matrix. It is direct to see that all the assumptions of Theorem 1 are satisfied. Consequently, Theorem 1 can be applied to predict convergence of (20) in theory. We conduct a simulation study and an application example below for support of such theoretical assertion.

4.1 Simulation study

Let $M = L^* + S^* + N$ be an observation matrix, where L^* and S^* are, respectively, the original low-rank and sparse matrices that we wish to recover by the problem (17), and N is the Gaussian noise matrix. In the following, `r` and `spr` represent, respectively, matrix rank and sparsity ratio. The MATLAB script for generating matrix M is as follows:

- $L = \mathrm{randn}(m, \mathtt{r}) * \mathrm{randn}(\mathtt{r}, n)$;
- $S = \mathrm{zeros}(m, n)$; $q = \mathrm{randperm}(m * n)$; $K = \mathrm{round}(\mathtt{spr} * m * n)$; $S(q(1:K)) = \mathrm{randn}(K, 1)$;
- $\sigma = 0$; % Noiseless case; $\sigma = 0.01$; % Gaussian noise; $N = \mathrm{randn}(m, n) * \sigma$;
- $T = L + S$; $M = T + N$.

Specifically, we set $m = n = 100$, and tested

$$(\mathtt{r}, \mathtt{spr}) = (1, 0.05), (5, 0.05), (10, 0.05), (20, 0.05), (1, 0.1), (5, 0.1), (10, 0.1),\ \text{and}\ (20, 0.1),$$

for which the decomposition problem roughly changes from easy to hard. Regarding the implementation issues, we empirically set the parameters $\alpha = 0.3$ and $\rho = \alpha$ in (20). The matrices L, S, and T in the procedure (20) are initialized by zero matrix. We terminated the procedure (20) when the relative change falls below 10^{-8}, i.e.,

$$\mathrm{RelChg} := \frac{\|(L^{k+1}, S^{k+1}, T^{k+1}) - (L^k, S^k, T^k)\|_F}{\|(L^k, S^k, T^k)\|_F + 1} \leqslant 10^{-8},$$

where $\|\cdot\|_F$ indicates the Frobenius norm. Let $\hat{L}$, $\hat{S}$, and $\hat{T}$ be a numerical solution of problem (17) obtained by the proposed BADMM. We will measure the quality of recovery by the relative error to (L^*, S^*, T^*), which is defined by

$$\mathrm{RelErr} := \frac{\|(\hat{L}, \hat{S}, \hat{T}) - (L^*, S^*, T^*)\|_F}{\|(L^*, S^*, T^*)\|_F + 1}.$$

In Table 1, we report the recovery results for the noiseless and Gaussian noise cases. From this table, it can be seen that when the true sparsity ratio `spr` of S increase or the noise is introduced, the relative error RelErr will go down, which suggests that the recovery performance will decline when the decomposition problem changes from easy to hard. In addition, for the noiseless case, the proposed BADMM can exactly recover the rank of L and the sparsity number of S. However, for the Gaussian noise case, since the noise imposes an additional impact on the recovery, the sparsity number of S cannot be exactly recovered.

Table 1 The matrix decomposition results on simulated matrices with the size 100 × 100

	(`r, spr`)	RelErr	Rank(L^*)	Rank($\hat{L}$)	$\|S^*\|_0$	$\|\hat{S}\|_0$
Noiseless case ($\sigma =$ 0)	(1, 0.05)	4.8674E−06	1	1	500	500
	(1, 0.1)	5.0446E−06	1	1	1000	1000
	(5, 0.05)	2.2342E−06	5	5	500	500
	(5, 0.1)	2.4366E−06	5	5	1000	1000
	(10, 0.05)	1.5039E−06	10	10	500	500
	(10, 0.1)	1.8572E−06	10	10	1000	1000
	(20, 0.05)	1.2889E−06	20	20	500	500
	(20, 0.1)	1.6974E−06	20	20	1000	1000
Gauss noise ($\sigma =$ 0.01)	(1, 0.05)	0.0049	1	1	500	1723
	(1, 0.1)	0.0060	1	1	1000	3797
	(5, 0.05)	0.0025	5	5	500	1541
	(5, 0.1)	0.0033	5	5	1000	3551
	(10, 0.05)	0.0022	10	10	500	1318
	(10, 0.1)	0.0024	10	10	1000	3183
	(20, 0.05)	0.0020	20	20	500	1110
	(20, 0.1)	0.0024	20	20	1000	3612

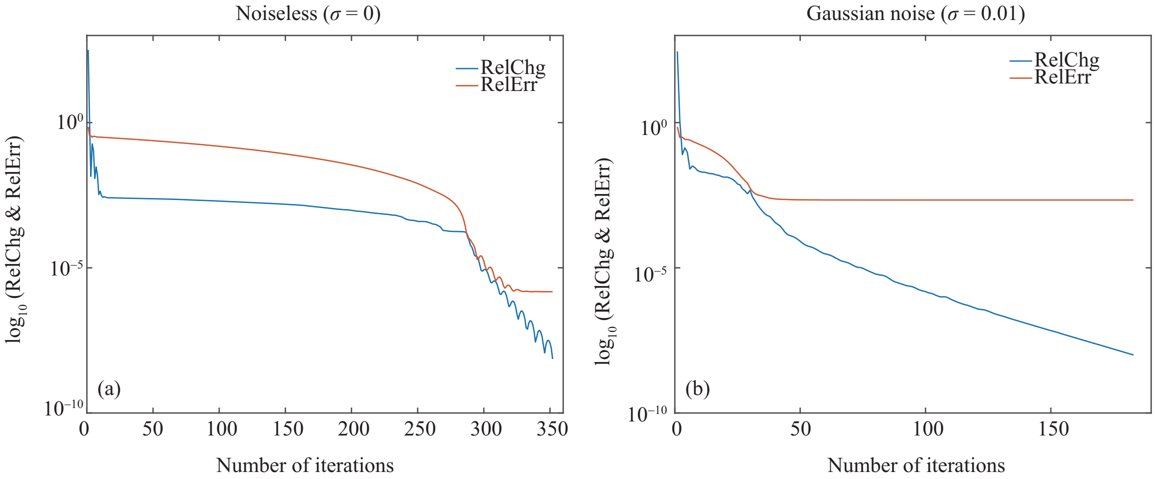


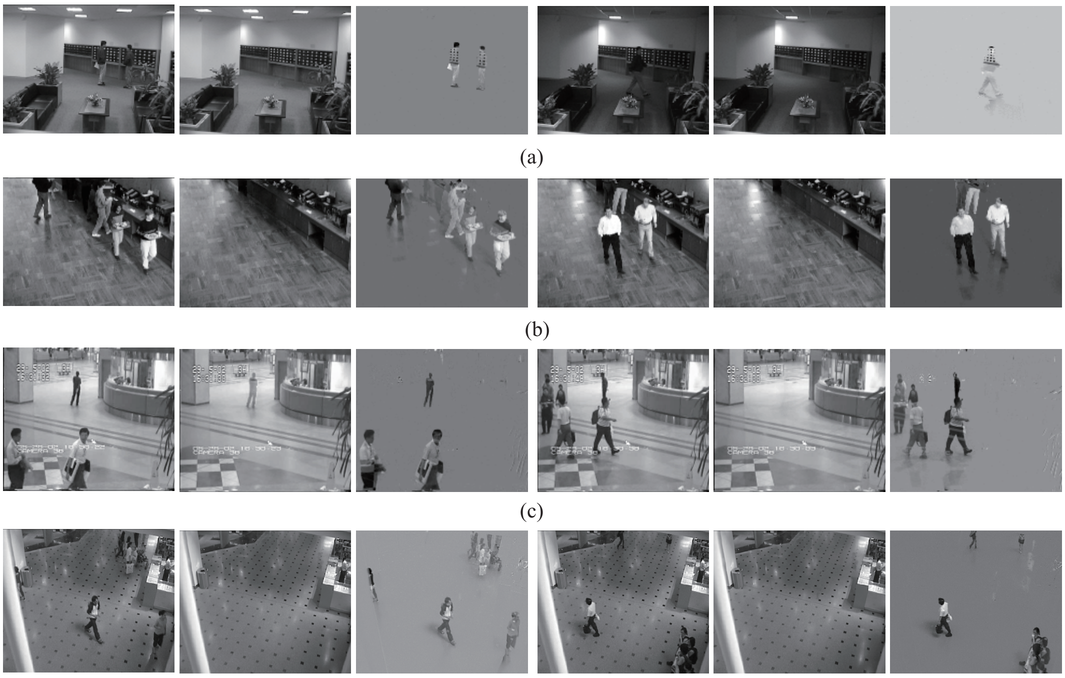
Figure 1 (Color online) Convergence results for (a) the noiseless case and (b) Gaussian noise with the standard deviation $\sigma = 0.01$.

In Figure 1, we further present the convergence results for the (`r`=10, `spr`=0.05) case with no noise and Gaussian noise. From this figure, it can be observed that when the relative change RelChg is less than 10^{-8}, the relative error RelErr will arrive at a stable value, which indicates that the proposed BADMM is convergent.

4.2 An application example

We further applied the model (17) with BADMM (20) to the background subtraction application. Background subtraction is a fundamental task in video surveillance. Its aim is to subtract the background from a video clip and meanwhile detect the anomalies (i.e., moving objects). From the webpage[1)], we download four video clips: Lobby, Bootstrap, Hall, and ShoppingMall. Then we chose 600 frames from each video clip and input these 600 frames into our algorithm. The parameter λ was fixed at the value $\frac{0.1}{\sqrt{\max(m,n)}}$. In Figure 2, we exhibit the separation results of some frames in four video clips. From

[1)] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.



(a)

(b)

(c)

(d)

Figure 2 Background subtraction results in the real-world video clips. (a) Lobby; (b) Bootstrap; (c) Hall; (d) Shopping-Mall.

Figure 2, it can be seen that our algorithm can produce a clean video background and meanwhile detect a satisfactory video foreground, which supports the validity and convergence of the proposed BADMM.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61603235), and Program for Science and Technology Innovation Talents in Universities of Henan Province (Grant No. 15HASTIT013). We thank all anonymous reviewers for their thoughtful and constructive comments that greatly improve the analysis and writing of the manuscript.

References

1 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn, 2011, 3: 1–122

2 Wang H, Banerjee A. Bregman alternating direction method of multipliers. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), Montréal, 2014. 2816–2824

3 Gabay D, Mercier B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput Math Appl, 1976, 2: 17–40

4 Alcouffe A, Enjalbert M, Muratet G. Méthodes de résolution du probléme de transport et de production d'une entreprise á établissements multiples en présence de coûts fixes. RAIRO Recherche opérationnelle, 1975, 9: 41–55

5 He B, Yuan X. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J Numer Anal, 2012, 50: 700–709

6 Goldstein T, O'Donoghue B, Setzer S, et al. Fast alternating direction optimization methods. SIAM J Imag Sci, 2014, 7: 1588–1623

7 Xu Y, Yin W, Wen Z, et al. An alternating direction algorithm for matrix completion with nonnegative factors. Front Math China, 2012, 7: 365–384

8 Bolte J, Sabach S, Teboulle M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math Program, 2014, 146: 459–494

9 Xu Y, Yin W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J Imag Sci, 2013, 6: 1758–1789

10 Hong M, Luo Z Q, Razaviyayn M. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J Optim, 2016, 26: 337–364

11 Li G, Pong T K. Global convergence of splitting methods for nonconvex composite optimization. SIAM J Optim, 2015, 25: 2434–2460

12 Wang F, Xu Z, Xu H. Convergence of bregman alternating direction method with multipliers for nonconvex composite

problems. ArXiv:1410.8625, 2014

13 Chen C, He B, Ye Y, et al. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math Program, 2016, 155: 57–79

14 Han D, Yuan X. A note on the alternating direction method of multipliers. J Optim Theor Appl, 2012, 155: 227–238

15 Cai X, Han D, Yuan X. On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput Optim Appl, 2017, 66: 39–73

16 Li M, Sun D, Toh K C. A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. Asia Pac J Oper Res, 2015, 32: 1550024

17 Mordukhovich B. Variational Analysis And Generalized Differentiation I: Basic Theory. Berlin: Springer, 2006. 30–35

18 Lojasiewicz S. Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles, 1963, 117: 87–89

19 Kurdyka K. On gradients of functions definable in o-minimal structures. Ann de l'institut Fourier, 1998, 48: 769–783

20 Attouch H, Bolte J, Svaiter B F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math Program, 2013, 137: 91–129

21 Si S, Tao D, Geng B. Bregman divergence-based regularization for transfer subspace learning. IEEE Trans Knowl Data Eng, 2010, 22: 929–942

22 Wu L, Hoi S C H, Jin R, et al. Learning Bregman distance functions for semi-supervised clustering. IEEE Trans Knowl Data Eng, 2012, 24: 478–491

23 Xu Z B, Chang X Y, Xu F M, et al. L1/2 regularization: a thresholding representation theory and a fast solver. IEEE Trans Neural Netw Learning Syst, 2012, 23: 1013–1027

24 Behmardi B, Raich R. On provable exact low-rank recovery in topic models. In: Proceedings of IEEE Statistical Signal Processing Workshop (SSP), Nice, 2011. 265–268

25 Xu H, Caramanis C, Mannor S. Outlier-robust PCA: the high-dimensional case. IEEE Trans Inform Theor, 2013, 59: 546–572

26 Zeng J, Xu Z, Zhang B, et al. Accelerated regularization based SAR imaging via BCR and reduced Newton skills. Signal Process, 2013, 93: 1831–1844