Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems

Max L.N. Gonçalves * Jefferson G. Melo * Renato D.C. Monteiro †

February 3, 2017

Abstract

This paper establishes convergence rate bounds for a variant of the proximal alternating direction method of multipliers (ADMM) for solving nonconvex linearly constrained optimization problems. The variant of the proximal ADMM allows the inclusion of an over-relaxation stepsize parameter belonging to the interval $(0, 2)$. To the best of our knowledge, all related papers in the literature only consider the case where the over-relaxation parameter lies in the interval $(0, (1 + \sqrt{5})/2)$.

2000 Mathematics Subject Classification: 47J22, 49M27, 90C25, 90C30, 90C60, 65K10.

Key words: alternating direction method of multipliers (ADMM), nonconvex program, pointwise iteration-complexity, first-order methods.

1 Introduction

We consider the following linearly constrained problem

$$\min\{f(x) + g(y) : Ax + By = b,\ x \in \mathbb{R}^n, y \in \mathbb{R}^p\} \tag{1}$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ and $g : \mathbb{R}^p \to (-\infty, \infty]$ are proper lower semicontinuous functions, $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times p}$ and $b \in \mathbb{R}^m$. Optimization problems such as (1) appear in many important applications such as nonnegative matrix factorization, distributed matrix factorization,

*Instituto de Matemática e Estatística, Universidade Federal de Goiás, Campus II- Caixa Postal 131, CEP 74001-970, Goiânia-GO, Brazil. (E-mails: maxlng@ufg.br and jefferson@ufg.br). The work of these authors was supported in part by CNPq Grants 406250/2013-8, 444134/2014-0 and 309370/2014-0.

†School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: monteiro@isye.gatech.edu). The work of this author was partially supported by NSF Grant CMMI-1300221.

distributed clustering, sparse zero variance discriminant analysis, tensor decomposition, and matrix completion, asset allocation (see, e.g., [1, 7, 22, 32, 33, 36, 38]). Moreover, it has observed that (specific variants of) the alternating direction method of multipliers (ADMM) can tackle many of the instances arising in these settings extremely well despite many of them being nonconvex.

A particular ADMM class for solving (1), namely, the proximal ADMM, recursively computes a sequence $\{(s_k, y_k, x_k)\}$ as

$$\begin{aligned} x_k &= \text{argmin}_x \left\{ \mathcal{L}_\beta(x, y_{k-1}, \lambda_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|_G^2 \right\}, \\ y_k &= \text{argmin}_y \left\{ \mathcal{L}_\beta(x_k, y, \lambda_{k-1}) + \frac{1}{2}\|y - y_{k-1}\|_H^2 \right\}, \\ \lambda_k &= \lambda_{k-1} - \theta\beta \left[Ax_k + By_k - b \right] \end{aligned} \tag{2}$$

where $\beta > 0$ is a penalty parameter, $\theta > 0$ is a stepsize parameter, $G \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{p\times p}$ are symmetric and positive semidefinite matrices, and

$$\mathcal{L}_\beta(x, y, \lambda) := f(x) + g(y) - \langle \lambda, Ax + By - b \rangle + \frac{\beta}{2}\|Ax + By - b\|^2$$

is the augmented Lagrangian function for problem (1). If $(H, G) = (0, 0)$ in the above method, we obtain the standard ADMM. Moreover, the above subproblems with suitable choices of G and H are easy to solve or even have closed-form solutions for many relevant instances of (1) (see [4, 17, 31, 34] for more details).

For the case in which f and g in (1) are both convex (e.g., see [11, 16, 17, 24]), the complexity results for the proximal ADMM (2) can be conveniently stated in terms of the following simple termination criterion associated with the optimality condition for (1), namely: for given $\rho, \varepsilon > 0$, terminate with a quintuple $(x, y, \lambda, r_1, r_2) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^p$ satisfying

$$\max\{\|Ax + By - b\|, \|r_1\|, \|r_2\|\} \le \rho, \quad r_1 \in \partial_\varepsilon f(x) - A^*\lambda, \quad r_2 \in \partial_\varepsilon g(y) - B^*\lambda \tag{3}$$

where ∂_ϵ denotes the classical ϵ-subdifferential of convex functions and the norms in the first inequality can be arbitrarily chosen. In terms of this termination criterion, the best ergodic iteration-complexity bound found in the literature is $\mathcal{O}(\max\{\rho^{-1}, \varepsilon^{-1}\})$ while the best pointwise one is $\mathcal{O}(\rho^{-2})$. (The latter bound is independent of ε since, in the pointwise case, the two inclusions above are shown to hold with $\varepsilon = 0$.)

This paper considers the special case of (1) in which f is as stated immediately following (1) (and hence not necessarily convex) and g is a differentiable function whose gradient is Lipschitz continuous on the whole $\mathbb{R}^p$. By considering an extended notion of subdifferential for the nonconvex function f (see for example [25, 27]), this paper establishes an $\mathcal{O}(\rho^{-2})$-pointwise iteration-complexity bound to obtain a quadruple $(x, y, \lambda, r_1) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^n$ satisfying

$$\max\{\|Ax + By - b\|, \|\nabla g(y) - B^*\lambda\|, \|r_1\|\} \le \rho, \quad r_1 \in \partial f(x) - A^*\lambda.$$

for an important subclass of the proximal ADMM (2). The latter subclass has the following properties: the penalty parameter β is sufficiently large (see (6)), G is an arbitrary positive semidefinite matrix, H is a sufficiently large positive multiple of the identity, and the stepsize θ lies in the interval $(0, 2)$. To the best of our knowledge, no iteration-complexity has been established in the literature for a variant of the ADMM with stepsize $\theta > (\sqrt{5}+1)/2$, even for the case in which (1) is assumed to be a convex problem. It is worth pointing out that [6, 9] show that larger choice of θ usually improves the practical performance of the proximal ADMM.

Previous related works. The ADMM was introduced in [8, 10] and is thoroughly discussed in [2, 9]. Even though convergence of the sequence generated by the ADMM has been established in very early papers about it, only recently has its iteration-complexity been established. To discuss this development in the convex case, we use the terminology weak pointwise or strong pointwise bounds to refer to complexity bounds relative to the best of the first k iterates or the last iterate, respectively, to satisfy the termination criterion (3). The first iteration-complexity bound for the ADMM was established in [24] under the assumptions that C is injective. More specifically, the ergodic iteration-complexity for the standard ADMM is derived in [24] for any $\theta \in (0, 1]$ while a weak pointwise iteration-complexity easily follows from the approach in [24] for any $\theta \in (0, 1)$. Subsequently, without assuming that C is injective, [17] established the ergodic iteration-complexity of the proximal ADMM (2) with $G = 0$ and $\theta = 1$ and, as a consequence, of the split inexact Uzawa method [37]. Paper [16] establishes the weak pointwise and ergodic iteration-complexity of another collection of ADMM instances which includes the standard ADMM for any $\theta \in (0, (1+\sqrt{5})/2)$. It should be noted however that [16, 17] do not provide any details on how to obtain an easily verifiable ergodic termination criterion with a well-established iteration-complexity bound. A strong pointwise iteration-complexity bound for the proximal ADMM (2) with $G = 0$ and $\theta = 1$ is derived in [18]. Pointwise and ergodic iteration-complexity results for the whole proximal ADMM (2) and for any $\theta \in (0, (1+\sqrt{5})/2)$ are given in [3, 13]. In addition to providing alternative proofs for these latter results, paper [11] obtains an ergodic iteration complexity bound for the proximal ADMM with $\theta = (1+\sqrt{5})/2$. Finally, a number of papers (see for example [4, 5, 12, 15, 23, 26] and references therein) have obtained similar complexity results in the context of other ADMM classes.

Iteration-complexity analysis of the ADMM has also been established for possibly nonconvex instances of (1) satisfying the same assumptions made on this paper, i.e., f is a proper lower semi-continuous function and g is a continuously differentiable function whose gradient is Lipschitz continuous on the whole $\mathbb{R}^p$. Recently, there have been a lot of interest on the study of ADMM variants for nonconvex problems (see, e.g., [14, 19, 20, 21, 28, 29, 30, 35]). The results developed in [14, 21, 28, 29, 30, 35] establish convergence of the generated sequence to a stationary point of (1) under the assumption that the objective function of (1) satisfies the so-called Kurdyka-Lojasiewicz (K-L) property. However, none of these papers considers the issue of iteration complexity for ADMM although their theoretical analysis are generally half-way or close to accomplishing such goal. Paper [19] analyzes the convergence of ADMM for solving nonconvex consensus and sharing problems and establishes the iteration complexity of ADMM for the consensus problem. Paper [20] studies the iteration-complexity of a multi-block type ADMM method whose two-block special case is a modification of the proximal ADMM in which the function g of the second subproblem in (2) is replaced by its linear approximation, G is positive definite and H is chosen as LI where L is the Lipschitz constant of $\nabla g(\cdot)$.

Organization of the paper. Subsection 1.1 presents some notation and basic results. Section 2 describes the proximal ADMM and presents corresponding convergence rate bounds whose proofs are given in Subsection 3.

1.1 Notation and basic results

This subsection presents some definitions, notation and basic results used in this paper.

Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space with inner product and associated norm denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. We use $\mathbb{R}^{m\times n}$ to denote the set of all $m \times n$ matrices. The image space of a matrix $Q \in \mathbb{R}^{m\times n}$ is defined as $\mathrm{Im}(Q) := \{Qx : x \in \mathbb{R}^n\}$ and $\mathcal{P}_Q$ denotes the Euclidean projection onto $\mathrm{Im}\,(Q)$. The notation $Q \succ 0$ means that Q is a definite positive matrix. The symbol $\lambda_{\min}(Q)$ denotes the minimum eigenvalue of a symmetric matrix Q. If Q is a symmetric and positive semidefinite matrix, the seminorm induced by Q on $\mathbb{R}^n$, denoted by $\|\cdot\|_Q$, is defined as $\|\cdot\|_Q = \langle Q(\cdot), \cdot\rangle^{1/2}$. For a given sequence $\{z_k : k \geq 0\}$, we denote by $\{\Delta z_k\}$ be sequence defined by

$$\Delta z_k := z_k - z_{k-1}, \quad k \geq 1.$$

The domain of a function $h : \mathbb{R}^n \to (-\infty, \infty]$ is the set $\mathrm{dom}\, h := \{x \in \mathbb{R}^n : h(x) < +\infty\}$. Moreover, h is said to be proper if $f(x) < \infty$ for some $x \in \mathbb{R}^n$.

We next recall some definitions and results of subdifferential calculus [25, 27].

Definition 1.1. *Let $h : \mathbb{R}^n \to (-\infty, \infty]$ be a proper lower semi-continuous function.*

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

$$\liminf_{y \neq x\; y \to x} \frac{h(y) - h(x) - \langle u, y - x\rangle}{\|y - x\|} \geq 0.$$

When $x \notin \mathrm{dom}\, h$, we set $\hat{\partial} h(x) = \emptyset$.

(ii) The limiting subdifferential, or simply subdifferential, of h at $x \in \mathrm{dom}\, h$, written by $\partial h(x)$, is defined as

$$\partial h(x) = \{u \in \mathbb{R}^n : \exists\, x_n \to x, h(x_n) \to h(x), u_k \in \hat{\partial} h(x_n),\ \textit{with}\ u_k \to u\}.$$

(iii) A critical (or stationary) point of h is a point x in the domain of h satisfying $0 \in \partial h(x)$.

The following result gives some properties of the subdifferential.

Proposition 1.2. *Let $h : \mathbb{R}^n \to (-\infty, \infty]$ be a proper lower semi-continuous function.*

(a) if $\{(u_k, x_k)\}$ is a sequence such that $x_k \to x$, $u_k \to u$, $h(x_k) \to h(x)$ and $u_k \in \partial h(x_k)$, then $u \in \partial h(x)$;

(b) if $x \in R^n$ is a local minimizer of h, then $0 \in \partial h(x)$;

(c) if $p : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function, then $\partial(h+p)(x) = \partial h(x) + \nabla p(x)$.

We end this section by recalling the definition of critical points of (1).

Definition 1.3. *A triple $(x^*, y^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ is a critical point of problem* (1) *if*

$$0 \in \partial f(x^*) - A^*\lambda^*, \quad 0 = \nabla g(y^*) - B^*\lambda^*, \quad 0 = Ax^* + By^* - b.$$

It is well-known that if (x^*, y^*) is a global minimum of (1), then there exists λ^* such that (x^*, y^*, λ^*) is a critical point of (1).

2 Proximal ADMM and its convergence rate

This section describes the assumptions made on problem (1) and states the variant of the proximal ADMM considered in this paper. It also states the main result of this paper (Theorem 2.1), and a special case of it (Corollary 2.2), both of them describing convergence rate bounds for the aforementioned proximal ADMM variant. The proof of Theorem 2.1 is however postponed to Section 3.

The augmented Lagrangian associated with problem (1) is defined as

$$\mathcal{L}_\beta(x, y, \lambda) := f(x) + g(y) - \langle \lambda, Ax + By - b \rangle + \frac{\beta}{2}\|Ax + By - b\|^2. \tag{4}$$

This paper considers problem (1) under the following set of assumptions:

(A0) $f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper lower semi-continuous function;

(A1) $B \neq 0$ and $\mathrm{Im}(B) \supset \{b\} \cup \mathrm{Im}(A)$;

(A2) $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable everywhere on $\mathbb{R}^p$ and there exists $L > 0$ such that

$$\|\mathcal{P}_{B^*}(\nabla g(y')) - \mathcal{P}_{B^*}(\nabla g(y))\| \leq L\|y' - y\| \quad \forall y, y' \in \mathbb{R}^p;$$

(A3) there exists $\tilde{L} \geq 0$ such that the function $g(\cdot) + \tilde{L}\|\cdot\|^2/2$ is convex, or equivalently,

$$g(y') - g(y) - \langle \nabla g(y), y' - y \rangle \geq -\frac{\tilde{L}}{2}\|y' - y\|^2 \quad \forall y, y' \in \mathbb{R}^p;$$

(A4) there exists $\bar{\beta} \geq 0$ such that

$$\bar{\mathcal{L}} := \inf_{(x,y)} \left\{ f(x) + g(y) + \frac{\bar{\beta}}{2}\|Ax + By - b\|^2 \right\} > -\infty.$$

Some comments are in order. First, due to the generality of (**A0**), problem (1) may include an extra constraint of the form $x \in X$ where X is a closed set since this constraint can be incorporated into f by adding to it the indicator function of X. Second, (**A1**) implies that for every $x \in \mathbb{R}^n$, there exists $y \in \mathbb{R}^p$ such that (x, y) satisfies the (linear) constraint of (1). The extra condition that $B \neq 0$ is very mild since otherwise (1) would be much simpler to solve. Third, if $\nabla g(\cdot)$ is L-Lipschitz continuous, then (**A2**) and (**A3**) with $\tilde{L} = L$ obviously hold. However, conditions (**A2**) and (**A3**) combined are generally weaker than the condition that $\nabla g(\cdot)$ be L-Lipschitz continuous.

Next we state the proximal ADMM for solving problem (1).

Proximal ADMM

---

(0) Let an initial point $(x_0, y_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ and a symmetric positive semi-definite matrix $G \in \mathbb{R}^{n \times n}$ be given. Let a stepsize parameter $\theta \in (0, 2)$ be given and define

$$\gamma := \frac{\theta}{(1 - |\theta - 1|)^2}. \tag{5}$$

Choose scalars $\beta \geq \bar{\beta}$ (see **(A4)**) and $\tau > 0$ such that

$$\delta_1 := \left(\frac{\beta}{4} - \frac{4\gamma\tau^2}{\beta\sigma_B^2}\right) > 0, \qquad \delta_2 := \lambda_{\min}\left[\left(\tau - \frac{2L^2\gamma}{\beta\sigma_B} - \frac{\tilde{L}}{2}\right) I + \delta_1 B^* B\right] > 0, \tag{6}$$

$$\beta B^* B + (\tau - \tilde{L}) I \succ 0 \tag{7}$$

where σ_B denotes the smallest positive eigenvalue of $B^* B$, and set $k = 1$;

(1) compute an optimal solution $x_k \in \mathbb{R}^n$ of the subproblem

$$\min_{x \in \mathbb{R}^n} \left\{ \mathcal{L}_\beta(x, y_{k-1}, \lambda_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|_G^2 \right\} \tag{8}$$

and then compute an optimal solution $y_k \in \mathbb{R}^p$ of the subproblem

$$\min_{y \in \mathbb{R}^p} \left\{ \mathcal{L}_\beta(x_k, y, \lambda_{k-1}) + \frac{\tau}{2}\|y - y_{k-1}\|^2 \right\}; \tag{9}$$

(2) set

$$\lambda_k = \lambda_{k-1} - \theta\beta \left[A x_k + B y_k - b \right] \tag{10}$$

and $k \leftarrow k + 1$, and go to step (1).

end

We now make a few remarks about the proximal ADMM. First, the assumption that $\theta \in (0, 2)$ guarantees that γ in (5) is well-defined and positive. Second, the special case of the proximal ADMM in which $G = 0$ requires only an initial pair (y_0, λ_0) since any of its iteration is independent of x_{k-1}. Third, (7) implies that the objective function of subproblem (9) is strongly convex and hence that y_k is uniquely determined. Fourth, the subproblems (8) and (9) are of the form

$$\min_{x\in\mathbb{R}^n}\left\{f(x)+\langle c,x\rangle+\frac{1}{2}\|x\|^2_{G+\rho A^*A}\right\},\quad \min_{y\in\mathbb{R}^p}\left\{g(y)+\langle d,y\rangle+\frac{1}{2}\|y\|^2_{\tau I+\rho B^*B}\right\}$$

for some $c \in \mathbb{R}^n$ and $d \in \mathbb{R}^p$. For the purpose of this paper, we assume they are easy to solve exactly, possibly by choosing $\tau > 0$, $\beta > 0$ and G appropriately. Fifth, any scalars $\beta \geq \bar{\beta}$ and $\tau > 0$ such that

$$\frac{2L^2\gamma}{\beta\sigma_B}+\tilde{L}\leq\tau<\frac{\beta\sigma_B}{4\sqrt{\gamma}} \tag{11}$$

satisfy conditions (6) and (7), and hence they are compactible conditions. Sixth, even though (11) does not allow $\tau = 0$, conditions (6) and (7) still allow the choice $\tau = 0$ if B^*B is assumed to be invertible (e.g., $B = I$).

Next we define a parameter required in order to present our convergence rate bounds. Define

$$\begin{aligned}\eta_0(y_0,\lambda_0) := \min_{(\Delta y_0,\Delta\lambda_0)} &\ \frac{c_1}{2}\|B^*\Delta\lambda_0\|^2+\frac{\beta}{4}\|B\Delta y_0\|^2\\ \text{s.t.} &\ \ \tau\Delta y_0+(1-1/\theta)B^*\Delta\lambda_0 = B^*\lambda_0-\nabla g(y_0)\end{aligned} \tag{12}$$

where

$$c_1 := \frac{2|\theta-1|}{\beta\theta(1-|\theta-1|)\sigma_B} > 0. \tag{13}$$

We make two remarks about the definition of $\eta_0(y_0, \lambda_0)$. First, if $\tau > 0$, then (12) always has a feasible solution. Second, if $\tau = 0$ and B^*B is invertible, then there are two cases to consider: (i) if $\theta \neq 1$, then (12) always has a feasible solution; and, (ii) if $\theta = 1$, then choosing λ_0 so that $B^*\lambda_0 = \nabla g(y_0)$ makes any pair $(\Delta y_0, \Delta\lambda_0)$ feasible to (12), and hence $\eta_0(y_0, \eta_0) = 0$.

We now state the main convergence rate result for the proximal ADMM under assumptions **(A0)-(A4)**.

Theorem 2.1. *Let $(x_0, y_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ be given and define*

$$\tilde{\mathcal{L}}_0 := \mathcal{L}_\beta(x_0, y_0, \lambda_0) - \bar{\mathcal{L}} + \eta_0 \tag{14}$$

where $\eta_0 := \eta_0(y_0, \lambda_0)$ and $\bar{\mathcal{L}}$ is as in **(A4)** *. If, for every $k \geq 1$, we define*

$$R_k^y := \nabla g(y_k) - B^*\hat{\lambda}_k, \quad R_k^\lambda := Ax_k + By_k - b \tag{15}$$

where

$$\hat{\lambda}_k := \lambda_{k-1} - \beta\left(Ax_k + By_{k-1} - b\right), \tag{16}$$

then we have

$$-G\Delta x_k \in \partial f(x_k) - A^*\hat{\lambda}_k, \tag{17}$$

and there exists $j \le k$ such that

$$\|\Delta x_j\|_G \le 2\sqrt{\frac{\tilde{\mathcal{L}}_0}{k}}, \quad \|R_j^y\| \le (\beta\|B^*B\| + \tau)\sqrt{\frac{2\tilde{\mathcal{L}}_0}{\delta_2 k}}, \quad \|R_j^\lambda\| \le \frac{1}{\beta\theta}\sqrt{\frac{2\tilde{\mathcal{L}}_0}{\delta_3 k}}$$

where δ_1, δ_2 are as in (6)*, and δ_3 is defined as*

$$\delta_3 = \left(\frac{\beta\theta\eta_0}{\tilde{\mathcal{L}}_0} + \frac{2\theta\gamma L^2}{\delta_2\sigma_B} + \frac{8\theta\gamma\tau^2}{\delta_1\sigma_B^2}\right)^{-1}. \tag{18}$$

By relating τ and β in a proper manner, the next result gives a more explicit specialization of the bounds obtained in Theorem 2.1 only in terms of β, L, $\|B^*B\|$ and σ_β. For the sake of simplicity, we view quantities depending only on the initial iterate as being $\mathcal{O}(1)$ in order to obtain the bounds below.

Corollary 2.2. *Choose $\beta \ge \bar{\beta}$ and $\tau > 0$ such that*

$$\frac{4L^2\gamma}{\beta\sigma_B} + \tilde{L} \le \frac{\beta\sigma_B}{4\sqrt{2\gamma}} = \tau$$

and assume that (x_0, y_0) is a feasible point for (1)*. Then, for every $k \ge 1$, inclusion* (17) *holds and there exists $j \le k$ such that*

$$\|\Delta x_j\|_G \le \mathcal{O}\left(\frac{1}{\sqrt{k}}\sqrt{\frac{1}{\beta\sigma_B^2} + 1}\right), \quad \|R_j^y\| \le \mathcal{O}\left[\frac{1}{\sqrt{k}}(\beta\|B^*B\| + \beta\sigma_B)\sqrt{\frac{1}{\beta^2\sigma_B^3} + \frac{1}{\beta\sigma_B}}\right],$$

$$\|R_j^\lambda\| \le \mathcal{O}\left[\frac{1}{\beta\sqrt{k}}\sqrt{\frac{1}{\sigma_B^2} + \beta + \frac{L^2}{\beta^2\sigma_B^4} + \frac{L^2}{\beta\sigma_B^2}}\right].$$

Proof. First note that the assumption on the parameters β and τ together with (6), yield $\tau = \mathcal{O}(\beta\sigma_B)$, $\delta_1 = \mathcal{O}(\beta)$ and $\delta_2 \ge \mathcal{O}(\beta\sigma_B)$. Hence, using (14) and the fact that (x_0, y_0) is feasible to problem (1), we obtain

$$\tilde{\mathcal{L}}_0 = \mathcal{O}(\eta_0 + 1) \le \mathcal{O}\left(\frac{1}{\beta\sigma_B^2} + 1\right), \quad 1/\delta_3 \le \mathcal{O}\left(\beta + \frac{L^2}{\beta\sigma_B^2}\right).$$

Then,

$$\frac{\tilde{\mathcal{L}}_0}{\delta_2} \le \mathcal{O}\left(\frac{1}{\beta^2\sigma_B^3} + \frac{1}{\beta\sigma_B}\right), \quad \frac{\tilde{\mathcal{L}}_0}{\delta_3} \le \mathcal{O}\left(\frac{1}{\sigma_B^2} + \beta + \frac{L^2}{\beta^2\sigma_B^4} + \frac{L^2}{\beta\sigma_B^2}\right).$$

Therefore, the conclusion of the corollary follows by combining the above bounds with Theorem 2.1. □

We now make a few remarks about Corollary 2.2. First, the larger β is, the smaller the bounds on $\|R_j^\lambda\|$ and $\|\Delta x_j\|_G$, and the larger the one on $\|R_j^y\|$ become. Second, the assumption that (x_0, y_0) is a feasible point for (1) in the above corolllary is not crucial. Without this assumption, the dependency of the above bounds on β would change since now $\tilde{\mathcal{L}}_0$ would be $\mathcal{O}(\eta_0 + \beta)$ instead of $\mathcal{O}(\eta_0 + 1)$.

3 Proof of Theorem 2.1

This section gives the proof of Theorem 2.1.

We first establish a few technical lemmas. The first one describes a set of inclusions/equations satisfied by the sequence $\{(x_k, y_k, \lambda_k)\}$ generated by the proximal ADMM.

Lemma 3.1. *Consider the sequence $\{(x_k, y_k, \lambda_k)\}$ generated by the proximal ADMM and let $\{\hat{\lambda}_k\}$ as defined in* (16). *Then, for every $k \geq 1$, the following inclusions hold:*

$$0 \in \left[\partial f(x_k) - A^*\hat{\lambda}_k\right] + G(x_k - x_{k-1}), \tag{19}$$

$$0 = \left[\nabla g(y_k) - B^*\hat{\lambda}_k\right] + \beta B^*B(y_k - y_{k-1}) + \tau(y_k - y_{k-1}), \tag{20}$$

$$0 = [Ax_k + By_k - b] + \frac{1}{\theta\beta}(\lambda_k - \lambda_{k-1}). \tag{21}$$

Proof. The optimality conditions for (8) and (9) imply that

$$\begin{aligned} 0 &\in \partial f(x_k) - A^*(\lambda_{k-1} - \beta(Ax_k + By_{k-1} - b)) + G(x_k - x_{k-1}), \\ 0 &= \nabla g(y_k) - B^*(\lambda_{k-1} - \beta(Ax_k + By_k - b)) + \tau(y_k - y_{k-1}), \end{aligned}$$

respectively. These relations combined with (16) immediately yield (19) and (20). Relation (21) follows immediately from (10). □

The following lemma provides a recursive relation for the sequence $\{\Delta\lambda_k\}$.

Lemma 3.2. *Let $\Delta y_0 \in \mathbb{R}^p$ and $\Delta\lambda_0 \in \mathbb{R}^m$ be such that*

$$\tau\Delta y_0 + (1 - 1/\theta)B^*\Delta\lambda_0 = B^*\lambda_0 - \nabla g(y_0). \tag{22}$$

Then, for every $k \geq 1$, we have

$$B^*\Delta\lambda_k = (1-\theta)B^*\Delta\lambda_{k-1} + \theta u_k \tag{23}$$

where

$$u_k = \nabla g(y_k) - \nabla g(y_{k-1}) + \tau(\Delta y_k - \Delta y_{k-1}). \tag{24}$$

Proof. Using (16) and (21) we easily see that

$$\theta\hat{\lambda}_k := \lambda_k + (\theta - 1)\lambda_{k-1} + \beta\theta B(y_k - y_{k-1}), \quad \forall k \geq 1.$$

This expression together with (20) then imply that

$$B^*\lambda_k = (1-\theta)B^*\lambda_{k-1} + \theta[\nabla g(y_k) + \tau\Delta y_k] \quad \forall k \geq 1. \tag{25}$$

Hence, in view of (24), relation (23) holds for every $k \geq 2$. Also, (24) and (25) both with $k = 1$ imply that

$$B^*\Delta\lambda_1 = B^*(\lambda_1 - \lambda_0) = -\theta B^*\lambda_0 + \theta\left[\nabla g(y_1) + \tau\Delta y_1\right] = -\theta B^*\lambda_0 + \theta\left[u_1 + \nabla g(y_0) + \tau\Delta y_0\right]$$

which, together with the definition of Δy_0 in (22), shows that (23) also holds for $k = 1$. □

The next lemma describes how the sequence $\{(x_k, y_k, \lambda_k)\}$ affects the value of the augmented Lagrangian function defined in (4).

Lemma 3.3. *For every $k \geq 1$, we have*

(a) $\mathcal{L}_\beta(x_k, y_{k-1}, \lambda_{k-1}) - \mathcal{L}_\beta(x_{k-1}, y_{k-1}, \lambda_{k-1}) \leq -\|\Delta x_k\|_G^2/2$;

(b) $\mathcal{L}_\beta(x_k, y_k, \lambda_{k-1}) - \mathcal{L}_\beta(x_k, y_{k-1}, \lambda_{k-1}) \leq (\tilde{L}/2)\|\Delta y_k\|^2 - (\beta/2)\|B\Delta y_k\|^2 - \tau\|\Delta y_k\|^2$;

(c) $\mathcal{L}_\beta(x_k, y_k, \lambda_k) - \mathcal{L}_\beta(x_k, y_k, \lambda_{k-1}) = [1/(\theta\beta)]\|\Delta\lambda_k\|^2$.

Proof. (a) In view of (8), we have $\mathcal{L}_\beta(x_k, y_{k-1}, \lambda_{k-1}) + \|x_k - x_{k-1}\|_G^2/2 \leq \mathcal{L}_\beta(x_{k-1}, y_{k-1}, \lambda_{k-1})$, which, combined with the identity $\Delta x_k = x_k - x_{k-1}$, proves (a).

(b) Observe that the objective function of (9) has the form

$$\mathcal{L}_\beta(x_k, \cdot, \lambda_{k-1}) + \frac{\tau}{2}\|\cdot - y_{k-1}\|^2 = (g + q)(\cdot) \tag{26}$$

where q is a quadratic function whose Hessian is $Q = \beta B^*B + \tau I$. Since $Q - \tilde{L}I \succ 0$ in view of (7), and condition **(A3)** implies that g is a proper lower semi-continuous such that $g + \tilde{L}\|\cdot\|^2$ is convex, it follows from inequality (37) of Lemma A.1 with $y = y_{k-1}$ and $\bar{y} = y_k$ that

$$(g + q)(y_{k-1}) \geq (g + q)(y_k) + \frac{\beta}{2}\|B(y_{k-1} - y_k)\|^2 + \frac{\tau}{2}\|y_{k-1} - y_k\|^2 - \frac{\tilde{L}}{2}\|y_{k-1} - y_k\|^2,$$

which together with (26) immediately implies (b).

(c) This statement follows from (10), the identity $\Delta\lambda_k = \lambda_k - \lambda_{k-1}$ and the fact that (4) implies that

$$\mathcal{L}_\beta(x_k, y_k, \lambda_k) = \mathcal{L}_\beta(x_k, y_k, \lambda_{k-1}) - \langle \lambda_k - \lambda_{k-1}, Ax_k + By_k - b\rangle. \qquad \square$$

Our goal now is to show that a certain sequence associated with $\{\mathcal{L}_\beta(x_k, y_k, \lambda_k)\}$ is monotonically decreasing, namely, the sequence $\{\hat{\mathcal{L}}_k\}$ defined as

$$\hat{\mathcal{L}}_k := \mathcal{L}_\beta(x_k, y_k, \lambda_k) + \eta_k \qquad \forall k \geq 0 \tag{27}$$

where

$$\eta_k := \frac{c_1}{2}\|B^*\Delta\lambda_k\|^2 + \frac{\beta}{4}\|B\Delta y_k\|^2 \qquad \forall k \geq 1, \tag{28}$$

and $\eta_0 = \eta_0(y_0, \lambda_0)$ and c_1 are as defined in (12) and (13), respectively.

Before establishing the monotonicity property of the above sequence, we state three technical results. The first one describes an upper bound on $\hat{\mathcal{L}}_k - \hat{\mathcal{L}}_{k-1}$ in terms of three quantities related to $\{\Delta x_k\}$, $\{\Delta\lambda_k\}$ and $\{\Delta y_k\}$, respectively.

Lemma 3.4. *For every $k \geq 1$,*

$$\hat{\mathcal{L}}_k - \hat{\mathcal{L}}_{k-1} \leq -\frac{1}{2}\|\Delta x_k\|_G^2 + \Theta_k^1 + \Theta_k^2 \tag{29}$$

where

$$\Theta_k^1 := \frac{1}{\beta\theta}\|\Delta\lambda_k\|^2 + \frac{c_1}{2}\left(\|B^*\Delta\lambda_k\|^2 - \|B^*\Delta\lambda_{k-1}\|^2\right) \tag{30}$$

and

$$\Theta_k^2 := \left(\frac{\tilde{L}}{2} - \tau\right)\|\Delta y_k\|^2 - \frac{\beta}{4}\left(\|B\Delta y_{k-1}\|^2 + \|B\Delta y_k\|^2\right) \tag{31}$$

where c_1 is defined in (13).

Proof. The lemma follows by adding the three inequalities given in statements (a), (b) and (c) of the previous lemma and using the definition of $\hat{\mathcal{L}}_k$ in (27). □

The next two results combined provide an upper bound for Θ_k^1 in terms of the sequence $\{\Delta y_k\}$.

Lemma 3.5. *Let u_k and Θ_k^1 be as in* (24) *and* (30)*, respectively. Then,*

$$\Theta_k^1 \le \frac{\gamma}{\beta\sigma_B}\|u_k\|^2$$

where γ is defined in (5).

Proof. Assumption **(A1)** clearly implies that $\Delta\lambda_k = -\beta\theta(Ax_k + By_k - b) \in \mathrm{Im}(B)$. Hence, it follows from Lemma A.2 that

$$\|\Delta\lambda_k\| = \|\mathcal{P}_B(\Delta\lambda_k)\| \le \frac{1}{\sqrt{\sigma_B}}\|B^*\Delta\lambda_k\|$$

where $\mathcal{P}_B(\cdot)$ is defined in Subsection 1.1. Hence, in view of (23) and (30), we have

$$\begin{aligned}\Theta_k^1 &\le \frac{1}{\beta\theta\sigma_B}\|B^*\Delta\lambda_k\|^2 + \frac{c_1}{2}(\|B^*\Delta\lambda_k\|^2 - \|B^*\Delta\lambda_{k-1}\|^2)\\ &= \left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)\|(1-\theta)B^*\Delta\lambda_{k-1} + \theta u_k\|^2 - \frac{c_1}{2}\|B^*\Delta\lambda_{k-1}\|^2.\end{aligned}$$

Note that if $\theta = 1$, then (13) implies that $c_1 = 0$ and the above inequality implies the conclusion of the lemma. We will now establish the conclusion of the lemma for the case in which $\theta \ne 1$. The previous inequality together with the relation $\|s_1 + s_2\|^2 \le (1+t)\|s_1\|^2 + (1+1/t)\|s_2\|^2$ which holds for every $s_1, s_2 \in \mathbb{R}^m$ and $t > 0$ yield

$$\begin{aligned}\Theta_k^1 &\le \left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)\left[(1+t)(\theta-1)^2\|B^*\Delta\lambda_{k-1}\|^2 + \left(1+\frac{1}{t}\right)\theta^2\|u_k\|^2\right] - \frac{c_1}{2}\|B^*\Delta\lambda_{k-1}\|^2\\ &= \left[\left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)(1+t)(\theta-1)^2 - \frac{c_1}{2}\right]\|B^*\Delta\lambda_{k-1}\|^2 + \left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)\left(1+\frac{1}{t}\right)\theta^2\|u_k\|^2\\ &= \left\{\frac{(1+t)(\theta-1)^2}{\beta\theta\sigma_B} - \left[1-(1+t)(\theta-1)^2\right]\frac{c_1}{2}\right\}\|B^*\Delta\lambda_{k-1}\|^2 + \left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)\left(1+\frac{1}{t}\right)\theta^2\|u_k\|^2.\end{aligned}$$

Using the above expression with $t = -1 + 1/|\theta - 1|$ and noting that $t > 0$ in view of the assumption that $\theta \in (0, 2)$, we conclude that

$$\begin{aligned}\Theta_k^1 &\le \left[\frac{1}{\beta\theta\sigma_B}|\theta - 1| - (1 - |\theta - 1|)\,\frac{c_1}{2}\right]\|B^*\Delta\lambda_{k-1}\|^2 + \left(\frac{1}{\beta\theta\sigma_B} + \frac{c_1}{2}\right)\frac{\theta^2}{1 - |\theta - 1|}\|u_k\|^2 \\ &= \frac{1}{\beta\theta\sigma_B}\left(1 + \frac{|\theta - 1|}{1 - |\theta - 1|}\right)\frac{\theta^2}{1 - |\theta - 1|}\|u_k\|^2\end{aligned}$$

where the last equality is due to (13). Hence, in view of (5), the conclusion of the lemma follows. □

Lemma 3.6. *The vector u_k defined in* (24) *satisfies*

$$\|u_k\|^2 \le 2L^2\|\Delta y_k\|^2 + \frac{4\tau^2}{\sigma_B}(\|B\Delta y_k\|^2 + \|B\Delta y_{k-1}\|^2).$$

Proof. Noting that (23) implies that $u_k \in \operatorname{Im} B^*$ and using assumption **(A2)** and Lemma A.2, we obtain

$$\begin{aligned}\|u_k\|^2 &= \|\mathcal{P}_{B^*}(u_k)\|^2 = \|\mathcal{P}_{B^*}\left(\nabla g(y_k) - \nabla g(y_{k-1})\right) + \tau\mathcal{P}_{B^*}\left(\Delta y_k - \Delta y_{k-1}\right)\|^2 \\ &\le \left[L\|\Delta y_k\| + \frac{\tau}{\sqrt{\sigma_B}}\|B\Delta y_k - B\Delta y_{k-1}\|\right]^2 \le 2L^2\|\Delta y_k\|^2 + \frac{2\tau^2}{\sigma_B}(\|B\Delta y_k\| + \|B\Delta y_{k-1}\|)^2 \\ &\le 2L^2\|\Delta y_k\|^2 + \frac{4\tau^2}{\sigma_B}(\|B\Delta y_k\|^2 + \|B\Delta y_{k-1}\|^2)\end{aligned}$$

where the last two inequalities follow from the Cauchy-Schwarz inequality and the relation $(s_1 + s_2)^2 \le 2s_1^2 + 2s_2^2$ for $s_1, s_2 \in \mathbb{R}$. □

Finally, the next proposition shows that the sequence $\{\hat{\mathcal{L}}_k\}$ decreases.

Proposition 3.7. *The sequence $\{(x_k, y_k, \lambda_k)\}$ generated by the proximal ADMM satisfies*

$$\hat{\mathcal{L}}_k - \hat{\mathcal{L}}_{k-1} \le -\frac{1}{2}\|\Delta x_k\|_G^2 - \delta_1\|B\Delta y_{k-1}\|^2 - \delta_2\|\Delta y_k\|^2 \quad \forall k \ge 1$$

where $\hat{\mathcal{L}}_k$ is defined in (27)*, and δ_1 and δ_2 are defined in* (6)*.*

Proof. It follows from Lemmas 3.5 and 3.6 that

$$\Theta_k^1 \le \frac{2\gamma L^2}{\beta\sigma_B}\|\Delta y_k\|^2 + \frac{4\gamma\tau^2}{\beta\sigma_B^2}\left(\|B\Delta y_{k-1}\|^2 + \|B\Delta y_k\|^2\right)$$

and hence, in view of (31), we have

$$\begin{aligned}\Theta_k^1 + \Theta_k^2 &\le \left(\frac{4\gamma\tau^2}{\beta\sigma_B^2} - \frac{\beta}{4}\right)\|B\Delta y_{k-1}\|^2 + \left(\frac{2\gamma L^2}{\beta\sigma_B} + \frac{\tilde{L}}{2} - \tau\right)\|\Delta y_k\| + \left(\frac{4\gamma\tau^2}{\beta\sigma_B^2} - \frac{\beta}{4}\right)\|B\Delta y_k\|^2 \\ &= -\delta_1\|B\Delta y_{k-1}\|^2 + \left\langle\left[\left(\frac{2\gamma L^2}{\beta\sigma_B} + \frac{\tilde{L}}{2} - \tau\right)I - \delta_1 B^*B\right]\Delta y_k, \Delta y_k\right\rangle \\ &\le -\delta_1\|B\Delta y_{k-1}\|^2 - \delta_2\|\Delta y_k\|^2\end{aligned}$$

where the last inequality is due to the definition of δ_2 in (18). Hence, the result follows due to (29). □

The next three lemmas show how to obtain convergence rate bounds for the quantities $\|\Delta x_j\|_G$, $\|\Delta y_j\|$ and $\|\Delta\lambda_j\|$ with the aid of Proposition 3.7. The first one shows that $\{\hat{\mathcal{L}}_k\}$ is bounded below.

Lemma 3.8. *For every $k \geq 1$, we have $\hat{\mathcal{L}}_k \geq \bar{\mathcal{L}}$ where $\hat{\mathcal{L}}_k$ and $\bar{\mathcal{L}}$ are as in* (27) *and* **(A4)**, *respectively.*

Proof. Assume for contradiction that there exists an index $k_0 \geq 0$ such that $\hat{\mathcal{L}}_{k_0+1} < \bar{\mathcal{L}}$. Since $\{\hat{\mathcal{L}}_k\}$ is decreasing (see Proposition 3.7), we obtain

$$\sum_{k=1}^{j}(\hat{\mathcal{L}}_k - \bar{\mathcal{L}}) \leq \sum_{k=1}^{k_0}(\hat{\mathcal{L}}_k - \bar{\mathcal{L}}) + (j-k_0)(\hat{\mathcal{L}}_{k_0+1} - \bar{\mathcal{L}}) \quad \forall j > k_0$$

and hence

$$\lim_{j\to\infty}\sum_{k=1}^{j}(\hat{\mathcal{L}}_k - \bar{\mathcal{L}}) = -\infty.$$

On the other hand, (4), (10), (27) and assumption **(A4)** imply that

$$\begin{aligned}
\hat{\mathcal{L}}_k &= \mathcal{L}_\beta(x_k, y_k, \lambda_k) + \eta_k \geq \mathcal{L}_\beta(x_k, y_k, \lambda_k) \geq \mathcal{L}_{\bar{\beta}}(x_k, y_k, \lambda_k) \\
&= f(x_k) + g(y_k) + \frac{\bar{\beta}}{2}\|Ax_k + By_k - b\|^2 + \frac{1}{\beta\theta}\langle \lambda_k, \lambda_k - \lambda_{k-1}\rangle \\
&\geq \bar{\mathcal{L}} + \frac{1}{2\beta\theta}\left(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2 + \|\lambda_k - \lambda_{k-1}\|^2\right) \geq \bar{\mathcal{L}} + \frac{1}{2\beta\theta}\left(\|\lambda_k\|^2 - \|\lambda_{k-1}\|^2\right)
\end{aligned}$$

and hence that

$$\sum_{k=1}^{j}(\hat{\mathcal{L}}_k - \bar{\mathcal{L}}) \geq \frac{1}{2\beta\theta}\left(\|\lambda_j\|^2 - \|\lambda_0\|^2\right) \geq -\frac{1}{2\beta\theta}\|\lambda_0\|^2 \quad \forall j \geq 1,$$

which yields the desired contradiction. □

Lemma 3.9. *For every $k \geq 1$, we have*

$$\sum_{j=1}^{k}\left(\frac{1}{2}\|\Delta x_j\|_G^2 + \delta_1\|B\Delta y_{j-1}\|^2 + \delta_2\|\Delta y_j\|^2 + \delta_3\|\Delta\lambda_j\|^2\right) \leq 2\tilde{\mathcal{L}}_0 \tag{32}$$

where δ_1, δ_2 and δ_3 are as defined in (6) *and* (18), *and $\tilde{\mathcal{L}}_0$ is as in* (14).

Proof. First note that Proposition 3.7 together with Lemma 3.8 yield, for every $k \geq 1$,

$$\sum_{j=1}^{k}\left(\frac{1}{2}\|\Delta x_j\|_G^2 + \delta_1\|B\Delta y_{j-1}\|^2 + \delta_2\|\Delta y_j\|^2\right) \leq \tilde{\mathcal{L}}_0 \tag{33}$$

which, in particular, implies that

$$\sum_{j=0}^{k} \|B\Delta y_j\|^2 \leq \frac{\tilde{\mathcal{L}}_0}{\delta_1}, \quad \sum_{j=1}^{k} \|\Delta y_j\|^2 \leq \frac{\tilde{\mathcal{L}}_0}{\delta_2}. \tag{34}$$

Due to (33), in order to prove (32), it suffices to show that

$$\sum_{j=1}^{k} \|\Delta \lambda_j\|^2 \leq \frac{1}{\delta_3} \tilde{\mathcal{L}}_0. \tag{35}$$

Then, in the remaining part of the proof we will show that (35) holds. By rewriting (30), we have

$$\|\Delta \lambda_k\|^2 = \beta\theta \left[\frac{c_1}{2} \left(\|B^* \Delta \lambda_{k-1}\|^2 - \|B^* \Delta \lambda_k\|^2 \right) + \Theta_k^1 \right] \qquad \forall k \geq 1,$$

where $\Delta\lambda_0$ is an arbitrary vector satisfying (23). Hence, using (12), we obtain

$$\begin{aligned}
\sum_{j=1}^{k} \|\Delta \lambda_j\|^2 &\leq \beta\theta \left[\frac{c_1}{2} \|B^* \Delta \lambda_0\|^2 + \sum_{j=1}^{k} \Theta_j^1 \right] \leq \beta\theta\eta_0 + \frac{\theta\gamma}{\sigma_B} \sum_{j=1}^{k} \|u_j\|^2 \\
&\leq \beta\theta\eta_0 + \sum_{j=1}^{k} \frac{2\gamma\theta L^2}{\sigma_B} \|\Delta y_j\|^2 + \frac{4\theta\gamma\tau^2}{\sigma_B^2} \sum_{j=1}^{k} \left(\|B\Delta y_{j-1}\|^2 + \|B\Delta y_j\|^2 \right) \\
&\leq \beta\theta\eta_0 + \sum_{j=1}^{k} \frac{2\gamma\theta L^2}{\sigma_B} \|\Delta y_j\|^2 + \sum_{j=0}^{k} \frac{8\theta\gamma\tau^2}{\sigma_B^2} \|B\Delta y_j\|^2 \\
&\leq \frac{\beta\theta\eta_0 \tilde{\mathcal{L}}_0}{\tilde{\mathcal{L}}_0} + \frac{2\theta\gamma L^2}{\sigma_B} \frac{\tilde{\mathcal{L}}_0}{\delta_2} + \frac{8\theta\gamma\tau^2}{\sigma_B^2} \frac{\tilde{\mathcal{L}}_0}{\delta_1} \\
&\leq \left(\frac{\beta\theta\eta_0}{\tilde{\mathcal{L}}_0} + \frac{2\theta\gamma L^2}{\delta_2 \sigma_B} + \frac{8\theta\gamma\tau^2}{\delta_1 \sigma_B^2} \right) \tilde{\mathcal{L}}_0
\end{aligned}$$

where the fifth inequality is due to (34). Hence, (35) follows from the last inequality combined with the definition of δ_3 in (18), and then the proof is concluded. □

Lemma 3.10. *For every $k \geq 1$, there exists $j \leq k$ such that*

$$\|\Delta x_j\|_G \leq 2\sqrt{\frac{\tilde{\mathcal{L}}_0}{k}}, \quad \|\Delta y_j\| \leq \sqrt{\frac{2\tilde{\mathcal{L}}_0}{\delta_2 k}}, \quad \|\Delta \lambda_j\| \leq \sqrt{\frac{2\tilde{\mathcal{L}}_0}{\delta_3 k}}$$

where $\tilde{\mathcal{L}}_0$, δ_2 and δ_3 are as defined in (14), (6) *and* (18), *respectively.*

Proof. The proof of this result follows directly from Lemma 3.9. □

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1: First note that the inclusion (17) follows immediately from (19). Also, taking into account (15), we obtain from (20) and (21) that

$$R_k^y = -(\beta B^*B + \tau)\Delta y_k, \quad R_k^\lambda = -\frac{1}{\beta\theta}\Delta\lambda_k.$$

Hence, to end the proof, just combine the above identities with Lemma 3.10. □

A Auxiliary Results

This section presents some auxiliary results which are used in our presentation.

Lemma A.1. *Assume that, for some $\tilde{L} \geq 0$, $g : \mathbb{R}^p \to [-\infty, \infty]$ is a proper lower semi-continuous function such that $g(\cdot) + \tilde{L}\|\cdot\|^2/2$ is convex and that $q(\cdot)$ is a quadratic function whose Hessian $Q \in \mathbb{R}^{p\times p}$ satisfies $Q - \tilde{L}I \succ 0$. Then, the problem*

$$\min\{(g+q)(y) : y \in \mathbb{R}^p\} \tag{36}$$

has a unique optimal solution $\bar{y}$ and

$$(g+q)(y) \geq (g+q)(\bar{y}) + \frac{1}{2}\|y-\bar{y}\|_Q^2 - \frac{\tilde{L}}{2}\|y-\bar{y}\|^2 \quad \forall y \in \mathbb{R}^p. \tag{37}$$

Proof. Define $\tilde{g} := g + \tilde{L}\|\cdot\|^2/2$, $\tilde{q} = q - \tilde{L}\|\cdot\|^2/2$ and $\tilde{Q} = Q - \tilde{L}I$. Clearly, $\tilde{g}$ is a proper lower semi-continuous convex function and $\tilde{q}$ is a strongly convex quadratic function whose Hessian is $\tilde{Q} \succ 0$. Since $g + q = \tilde{g} + \tilde{q}$, we conclude that the objective function of (36) is strongly convex, and hence that the first statement of the lemma follows. Moreover, we have

$$0 \in \partial(g+q)(\bar{y}) = \partial(\tilde{g}+\tilde{q})(\bar{y}) = \partial\tilde{g}(\bar{y}) + \nabla\tilde{q}(\bar{y})$$

and hence

$$\tilde{g}(y) \geq \tilde{g}(\bar{y}) - \langle\nabla\tilde{q}(\bar{y}), y-\bar{y}\rangle \quad \forall y \in \mathbb{R}^p.$$

On the other hand, the fact that $\tilde{q}$ is a quadratic function implies that

$$\tilde{q}(y) = \tilde{q}(\bar{y}) + \langle\nabla\tilde{q}(\bar{y}), y-\bar{y}\rangle + \frac{1}{2}\|y-\bar{y}\|_{\tilde{Q}}^2 \quad \forall y \in \mathbb{R}^p.$$

Adding the above two relations, and using the fact that $g + q = \tilde{g} + \tilde{q}$ and the definition of $\tilde{Q}$, we conclude that (37) holds. □

Lemma A.2. *Let $S \in \mathbb{R}^{n\times p}$ be a non-zero matrix and let σ_S denote the smallest positive eigenvalue of SS^*. Then, for every $u \in \mathbb{R}^p$, there holds*

$$\|\mathcal{P}_{S^*}(u)\| \leq \frac{1}{\sqrt{\sigma_S}}\|Su\|.$$

Proof. Let r denote the rank of S and let $S = R\Lambda Q^*$ be a partial singular-value decomposition of S where $R \in \mathbb{R}^{n\times r}$ is such that $R^*R = I$, $Q \in \mathbb{R}^{p\times r}$ is such that $Q^*Q = I$ and $\Lambda \in \mathbb{R}^{r\times r}$ is a positive diagonal matrix. It is easy to see that

$$\|\mathcal{P}_{S^*}(u)\| = \|\mathcal{P}_Q(u)\| = \|Q(Q^*Q)^{-1}Q^*u\| = \|Q^*u\| \quad \forall u \in \mathbb{R}^p. \tag{38}$$

Moreover, we have

$$\|Q^*u\| = \|\Lambda^{-1}\Lambda Qu\| \le \|\Lambda^{-1}\|\|\Lambda Q^*u\| = \|\Lambda^{-1}\|\|R\Lambda Q^*u\| = \|\Lambda^{-1}\|\|Su\| \quad \forall u \in \mathbb{R}^p.$$

The result now follows from the above two relations and the fact that $\|\Lambda^{-1}\| = 1/\sqrt{\sigma_S}$. □

References

[1] B. P. W. Ames and M. Hong. Alternating direction method of multipliers for penalized zero-variance discriminant analysis. *Comput. Optim. Appl.*, 64(3):725–754, 2016.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[3] Y. Cui, X. Li, D. Sun, and K. C. Toh. On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. *J. Optim. Theory Appl.*, 169(3):1013–1041, 2016.

[4] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *J. Sci. Comput.*, pages 1–28, 2015.

[5] E. X. Fang, B. He, H. Liu, and X. Yuan. Generalized alternating direction method of multipliers: new theoretical insights and applications. *Math. Prog. Comp.*, 7(2):149–187, 2015.

[6] M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM J. Matrix Anal. Appl.*, 34(3):946–977, 2013.

[7] P. A. Forero, A. Cano, and G. B. Giannakis. Distributed clustering using wireless sensor networks. *IEEE J. Selected Topics Signal Process.*, 5(4):707–724, 2011.

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

[9] R. Glowinski. *Numerical Methods for Nonlinear Variational Problems.* Springer Series in Computational Physics. Springer-Verlag, 1984.

[10] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par penalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. 1975.

[11] M. L. N. Gonçalves, J. G. Melo, and R. D. C. Monteiro. Extending the ergodic convergence rate of the proximal ADMM. *Avaliable on http://www.arxiv.org.*

[12] M. L. N. Gonçalves, J. G. Melo, and R. D. C. Monteiro. Improved pointwise iteration-complexity of a regularized ADMM and of a regularized non-euclidean HPE framework. *To appear in SIAM J. Optim.*

[13] Y. Gu, B. Jiang, and H. Deren. A semi-proximal-based strictly contractive Peaceman-Rachford splitting method. *Avaliable on http://www.arxiv.org.*

[14] K. Guo, D. R. Han, and T. T. Wu. Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. *To appear in Int. J. Comput. Math.*

[15] W. W. Hager, M. Yashtini, and H. Zhang. An $O(1/k)$ convergence rate for the variable stepsize Bregman operator splitting algorithm. *SIAM J. Numer. Anal.*, 54(3):1535–1556, 2016.

[16] B. He, F. Ma, and X. Yuan. On the step size of symmetric alternating directions method of multipliers. *Avaliable on http://www.optimization-online.org.*

[17] B. He and X. Yuan. On the $\mathcal{O}(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numer. Anal.*, 50(2):700–709, 2012.

[18] B. He and X. Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Numer. Math.*, 130(3):567–577, 2015.

[19] M. Hong, Z. Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM J. Optim.*, 26(1):337–364, 2016.

[20] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Avaliable on http://www.arxiv.org.*

[21] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM J. Optim.*, 25(4):2434–2460, 2015.

[22] A. P. Liavas and N. D. Sidiropoulos. Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers. *Avaliable on http://www.arxiv.org.*

[23] T. Lin, S. Ma, and S. Zhang. An extragradient-based alternating direction method for convex minimization. *Found. Comput. Math.*, pages 1–25, 2015.

[24] R. D. C. Monteiro and B. F Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM J. Optim.*, 23(1):475–507, 2013.

[25] B.S. Mordukhovich. *Variational analysis and generalized differentiation I: basic theory.* Grundlehren der mathematischen Wissenschaften. Springer, Berlin,, 2006.

[26] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM J. Imaging Sci.*, 8(1):644–681, 2015.

[27] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis.* Springer, Berlin, 1998.

[28] F. Wang, W. Cao, and Z. Xu. Convergence of multi-block bregman ADMM for nonconvex composite problems. *Avaliable on http://www.arxiv.org.*

[29] F. Wang, Z. Xu, and H. K. Xu. Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems. *Avaliable on http://www.arxiv.org.*

[30] W. Wang, Y. Yin and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Avaliable on http://www.arxiv.org.*

[31] X. Wang and X Yuan. The linearized alternating direction method of multipliers for dantzig selector. *SIAM J. Sci. Comput.*, 34(5):2792–2811, 2012.

[32] Z. Wen, X. Peng, X. Liu, X. Sun, and X. Bais. Asset allocation under the Basel accord risk measures. *Avaliable on http://www.arxiv.org.*

[33] Y. Xu, W. Yin, Z. Wen, and Y. Zhang. An alternating direction algorithm for matrix completion with nonnegative factors. *Frontiers Math. China*, 7(2):365–384, 2012.

[34] J. Yang and X. Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.*, 82(281):301–329, 2013.

[35] L. Yang, T. K. Pong, and X. Chen. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. *Avaliable on http://www.arxiv.org.*

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

[37] X. Zhang, M. Burger, X. Bresson, and S. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM J. Imaging Sci.*, 3(3):253–276, 2010.

[38] Y. Zhang. An alternating direction algorithm for nonnegative matrix factorization. *Rice Technical Report*, 2010.