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

M.L.N. GONÇALVES, [1] J.G. MELO [1] AND R.D.C. MONTEIRO [2]

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, this is the first time that stepsize parameter larger than $(1 + \sqrt{5})/2$ is considered in the ADMM literature.

Keywords: alternating direction method of multipliers (ADMM), nonconvex program, pointwise iteration-complexity, first-order methods.
Mathematics Subject Classification: 47J22, 49M27, 90C25, 90C26, 90C30, 90C60, 65K10.

1 Introduction

We consider the following linearly constrained problem

$$\min\{f(x) + g(y) : Ax + By = b,\ x \in \mathbb{R}^n, y \in \mathbb{R}^p\} \tag{1.1}$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ and $g : \mathbb{R}^p \to (-\infty, \infty]$ are proper lower semicontinuous functions, $A \in \mathbb{R}^{l\times n}$, $B \in \mathbb{R}^{l\times p}$ and $b \in \mathbb{R}^l$. Optimization problems such as (1.1) appear in many important applications such as nonnegative matrix factorization, distributed matrix factorization, distributed clustering, sparse zero variance discriminant analysis, tensor decomposition, and matrix completion, asset allocation (see, e.g., [1, 8, 25, 34, 35, 39, 41]). 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 (1.1), namely, the proximal ADMM, recursively computes a sequence $\{(x_k, y_k, \lambda_k)\}$ as

$$\begin{aligned} x_k &= \operatorname{argmin}_x \left\{ \mathcal{L}_\beta(x, y_{k-1}, \lambda_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|_G^2 \right\}, \\ y_k &= \operatorname{argmin}_y \left\{ \mathcal{L}_\beta(x_k, y, \lambda_{k-1}) + \frac{1}{2}\|y - y_{k-1}\|_H^2 \right\}, \\ \lambda_k &= \lambda_{k-1} - \theta\beta\,[Ax_k + By_k - b] \end{aligned} \tag{1.2}$$

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 (1.1). 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 (1.1) (see [5, 19, 33, 37] for more details).

For the case in which f and g in (1.1) are both convex (e.g., see [12, 18, 19, 27]), the complexity results for the proximal ADMM (1.2) can be conveniently stated in terms of the following simple termination criterion

[1] 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 406975/2016-7 and 302666/2017-6.

[2] 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.

associated with the optimality condition for (1.1), namely: for given $\rho, \varepsilon > 0$, terminate with a quintuple $(x, y, \lambda, r_1, r_2) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^l \times \mathbb{R}^n \times \mathbb{R}^p$ satisfying

$$\max\{\|Ax + By - b\|, \|r_1\|, \|r_2\|\} \leq \rho, \quad r_1 \in \partial_\varepsilon f(x) - A^*\lambda, \quad r_2 \in \partial_\varepsilon g(y) - B^*\lambda \tag{1.3}$$

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 (1.1) in which f is as stated immediately following (1.1) (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 [28, 30]), this paper establishes an $\mathcal{O}(\rho^{-2})$-pointwise iteration-complexity bound to obtain a quadruple $(x, y, \lambda, r_1) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^l \times \mathbb{R}^n$ satisfying

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

for an important subclass of the proximal ADMM (1.2). The latter subclass has the following properties: the penalty parameter β is sufficiently large (see (2.3)), 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, this is the first time that iteration-complexity is established in the literature for a variant of the ADMM with stepsize $\theta > (\sqrt{5} + 1)/2$, even for the case in which (1.1) is assumed to be a convex problem. It is worth pointing out that [7, 10] show that larger choice of θ usually improves the practical performance of the proximal ADMM. Finally, asymptotic convergence of the proposed method is also analyzed under Kurdyka-Łojasiewicz property.

Previous related works. The ADMM was introduced in [9, 11] and is thoroughly discussed in [3, 10]. 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 (1.3). The first iteration-complexity bound for the ADMM was established in [27] under the assumptions that C is injective. More specifically, the ergodic iteration-complexity for the standard ADMM is derived in [27] for any $\theta \in (0, 1]$ while a weak pointwise iteration-complexity easily follows from the approach in [27] for any $\theta \in (0, 1)$. Subsequently, without assuming that C is injective, [19] established the ergodic iteration-complexity of the proximal ADMM (1.2) with $G = 0$ and $\theta = 1$ and, as a consequence, of the split inexact Uzawa method [40]. Paper [18] 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 [18, 19] 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 (1.2) with $G = 0$ and $\theta = 1$ is derived in [20]. Pointwise and ergodic iteration-complexity results for the whole proximal ADMM (1.2) and for any $\theta \in (0, (1 + \sqrt{5})/2)$ are given in [4, 14]. In addition to providing alternative proofs for these latter results, paper [12] obtains an ergodic iteration complexity bound for the proximal ADMM with $\theta = (1 + \sqrt{5})/2$. A number of papers (see for example [5, 6, 13, 16, 26, 29] and references therein) have obtained similar complexity results in the context of other ADMM classes. Finally, it should be mentioned that, subsequently to this paper, [17] studied complexity results for the proximal ADMM with stepsize parameter $\theta \in (0, 2)$ in the convex case.

Iteration-complexity analysis of the ADMM has also been established for possibly nonconvex instances of (1.1) 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., [15, 21, 22, 23, 24, 31, 32, 36, 38]). The results developed in [15, 24, 31, 32, 36, 38] establish convergence of the generated sequence to a stationary point of (1.1) under the assumption that the objective function of (1.1) satisfies the so-called Kurdyka-Łojasiewicz (K-Ł) 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 [22] analyzes the convergence of ADMM for solving nonconvex consensus and sharing problems and establishes the iteration complexity of ADMM for the consensus problem. Paper [23] 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 (1.2) is replaced by its linear approximation, G is positive definite and H is chosen as LI where L is the Lipschitz constant of $\nabla g(\cdot)$. Finally, [21] studies the iteration-complexity of a proximal variant of the augmented Lagrangian method for solving the 1-block special form of (1.1), i.e., with $f = 0$ and $A = 0$.

Organization of the paper. Subsection 1.1 presents some notations and basic results. Section 2 describes the proximal ADMM and presents corresponding convergence rate bounds whose proofs are given in Section 3. The asymptotic convergence of the proposed method under Kurdyka-Łojasiewicz property are discussed in Section 4

1.1 Notations and basic results

This subsection presents some definitions, notations 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 $\mathbb{R}^{l\times n}$ to denote the set of all $l \times n$ matrices. The image space of a matrix $Q \in \mathbb{R}^{l\times n}$ 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 $\{z_k : k \geq 0\}$, let $\{\Delta z_k\}$ be the 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 $h(x) < \infty$ for some $x \in \mathbb{R}^n$.

We next recall some definitions and results of subdifferential calculus [28, 30].

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 $x \in \mathit{dom}\, h$, written by $\partial h(x)$, is defined as

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

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

We next recall the definition of critical points of (1.1).

Definition 1.3. A triple $(x^*, y^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^l$ is a critical point of problem (1.1) if

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

Under some mild conditions, it can be shown that if (x^*, y^*) is a local minimum of (1.1), then there exists λ^* such that (x^*, y^*, λ^*) is a critical point of (1.1).

We end this section by presenting an auxiliary result which is used in our presentation.

Lemma 1.4. *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

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

Moreover, we have

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

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

2 Proximal ADMM and its convergence rate

This section describes the assumptions made on problem (1.1) and states the variant of the proximal ADMM considered in this paper. It also states the main result of this paper (Theorem 2.2), and a special case of it (Corollary 2.3), both of them describing convergence rate bounds for the aforementioned proximal ADMM variant. The proof of Theorem 2.2 is however postponed to Section 3.

The augmented Lagrangian associated with problem (1.1) is defined as

$$\mathcal{L}_\beta(x,y,\lambda) := f(x) + g(y) - \langle \lambda, Ax+By-b\rangle + \frac{\beta}{2}\|Ax+By-b\|^2. \tag{2.1}$$

This paper considers problem (1.1) 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 $\operatorname{Im}(B) \supset \{b\} \cup \operatorname{Im}(A)$;

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

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

(A3) there exists $m \ge 0$ such that the function $g(\cdot) + m\|\cdot\|^2/2$ is convex, or equivalently,

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

(A4) there exists $\bar\beta \ge 0$ 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 (1.1) 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 (1.1). The extra condition that $B \neq 0$ is very mild since otherwise (1.1) would be much simpler to solve. Third, if $\nabla g(\cdot)$ is L-Lipschitz continuous, then (**A2**) and (**A3**) with $m = L$ 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 (1.1).

Proximal ADMM

---

(0) Let an initial point $(x_0, y_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^l$ 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

$$\gamma := \frac{\theta}{(1-|\theta-1|)^2}. \tag{2.2}$$

Choose scalars $\beta \ge \bar\beta$ (see **(A4)**) and $\tau \ge 0$ such that

$$\delta_1 := \left(\frac{\beta\sigma_B + 2\tau - m}{4} - \frac{3\gamma(L^2+\tau^2)}{\beta\sigma_B^+} \right) > 0, \tag{2.3}$$

where σ_B (resp., σ_B^+) denotes the smallest eigenvalue (resp., positive eigenvalue) of B^*B, and set $k = 1$;

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

$$\min_{x\in\mathbb{R}^n} \left\{ \mathcal{L}_\beta(x, y_{k-1}, \lambda_{k-1}) + \frac{1}{2}\|x - x_{k-1}\|_G^2 \right\} \tag{2.4}$$

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

$$\min_{y\in\mathbb{R}^p} \left\{ \mathcal{L}_\beta(x_k, y, \lambda_{k-1}) + \frac{\tau}{2}\|y - y_{k-1}\|^2 \right\}; \tag{2.5}$$

(2) set

$$\lambda_k = \lambda_{k-1} - \theta\beta\,[Ax_k + By_k - b] \tag{2.6}$$

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 (2.2) 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, note that if $\beta B^*B + \tau I - mI \succ 0$, then the objective function of subproblem (2.5) is strongly convex and hence y_k is uniquely determined. Fourth, the subproblems (2.4) and (2.5) are of the form

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

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 $\tau \geq 0$, $\beta > 0$ and G appropriately. Fifth, condition (2.3) imposed on the different data constants and parameters of the proximal ADMM method is needed to establish convergence rate bounds for it (see Theorem 2.2). Note that, if either $\sigma_\beta > 0$ or $\tau > m/2$, then it is always possible to choose a sufficiently large penalty parameter β satisfying this condition. Hence, if $\sigma_\beta > 0$, convergence rate bounds for the standard ADMM (i.e., the special case of the above method with $G = 0$ and $\tau = 0$) can be derived for β sufficiently large (see Corollary 2.3).

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

$$\begin{aligned} \eta_0(y_0, \lambda_0; \theta) := \min_{(\Delta y_0, \Delta\lambda_0)} \ & \frac{c_1}{2}\|B^*\Delta\lambda_0\|^2 + \left(\frac{\beta\sigma_B + 2\tau - m}{4}\right)\|\Delta y_0\|^2 \\ \text{s.t.} \ & \tau\Delta y_0 + (1 - 1/\theta)B^*\Delta\lambda_0 = B^*\lambda_0 - \nabla g(y_0) \end{aligned} \tag{2.7}$$

where

$$c_1 := \frac{2|\theta - 1|}{\beta\theta(1 - |\theta - 1|)\sigma_B^+} \geq 0. \tag{2.8}$$

Theorem 2.2 below expresses the complexity of the proximal ADMM in terms of the quantity η_0, which depends on the initial iterate pair (y_0, λ_0) as well as the constant m and the parameters θ, β and τ used by the method. This contrasts with the analysis of the papers [21, 22, 23] which derive iteration-complexity for variants of the augmented Lagrangian and the proximal ADMM expressed in terms of both (x_0, y_0, λ_0) and (x_1, y_1, λ_1). We believe that the one derived in this paper is more convenient since quantities expressed only in terms of (x_0, y_0, λ_0) are easier to compute and/or estimate. Definition (2.7) of η_0 is somewhat complicated but, under some conditions, it simplifies or an upper bound on η_0 can easily be obtained. The following trivial result elaborates on this point and gives sufficient conditions for the quantity η_0 to be finite.

Lemma 2.1. *Let $(y_0, \lambda_0) \in \mathbb{R}^p \times \mathbb{R}^l$ and $\theta \in (0,2)$ be given. Then, problem (2.7) is feasible, and hence the quantity $\eta_0 := \eta_0(y_0, \lambda_0; \theta)$ is finite, under either one of the following conditions:*

(i) $B^*\lambda_0 = \nabla g(y_0)$*, in which case* $\eta_0 = 0$*;*

(ii) $\tau = 0$, $\theta \neq 1$ *and* B^*B *invertible;*

(iii) $\tau > 0$.

The following result derives convergence rate bounds for the proximal ADMM for solving the nonconvex optimization problem (1.1) satisfying assumptions **(A0)-(A4)** for any $\theta \in (0,2)$ and β sufficiently large.

Theorem 2.2. *Assume that the stepsize $\theta \in (0,2)$ and the initial pair $(y_0, \lambda_0) \in \mathbb{R}^p \times \mathbb{R}^l$ is such that the quantity $\eta_0 := \eta_0(y_0, \lambda_0; \theta)$ defined in (2.7) is finite and define*

$$\Delta^0_\beta := \mathcal{L}_\beta(x_0, y_0, \lambda_0) - \bar{\mathcal{L}} \tag{2.9}$$

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

$$\hat{\lambda}_k := \lambda_{k-1} - \beta\left(Ax_k + By_{k-1} - b\right), \tag{2.10}$$

then we have

$$-G\Delta x_k \in \partial f(x_k) - A^*\hat{\lambda}_k, \tag{2.11}$$

and there exists $j \leq k$ such that

$$\|\Delta x_j\|_G \leq \sqrt{\frac{6\max\{\eta_0, \Delta^0_\beta\}}{k}}, \qquad \|\nabla g(y_j) - B^*\hat{\lambda}_j\| \leq (\beta\|B^*B\| + \tau)\sqrt{\frac{3\max\{\eta_0, \Delta^0_\beta\}}{\delta_1 k}},$$

$$\|Ax_j + By_j - b\| \leq \frac{1}{\beta\theta}\sqrt{\frac{3\max\{\eta_0, \Delta^0_\beta\}}{\delta_2 k}}$$

where δ_1 is as in (2.3)*, and δ_2 is defined as*

$$\delta_2 := \left(\beta\theta + \frac{6\theta\gamma(L^2+\tau^2)}{\sigma_B^+\delta_1}\right)^{-1}. \tag{2.12}$$

As a consequence of the previous result, the following corollary establishes convergence rate bounds for the standard ADMM for solving (1.1) with B^*B invertible for any stepsize $\theta \in (0,2)$ and sufficiently large penalty parameter β.

Corollary 2.3. *Consider the standard ADMM, i.e., the special case of the proximal ADMM with $G = 0$ and $\tau = 0$, applied to problem* (1.1) *with B^*B invertible. Assume that the initial pair (y_0, λ_0) satisfies $B^*\lambda_0 = \nabla g(y_0)$ and $\beta \geq \bar{\beta}$ is chosen in a such a way that*

$$\frac{\beta\sigma_B - 2m}{8} \geq \frac{3\gamma L^2}{\beta\sigma_B}. \tag{2.13}$$

Then, $\Delta_\beta^0 \geq 0$ where Δ_β^0 is as in (2.9)*, and for every $k \geq 1$,*

$$0 \in \partial f(x_k) - A^*\hat{\lambda}_k$$

and there exists $j \leq k$ such that

$$\|\nabla g(y_j) - B^*\hat{\lambda}_j\| \leq \mathcal{O}\left(\sqrt{\beta}\|B^*B\|\sqrt{\frac{\Delta_\beta^0}{\sigma_B k}}\right), \qquad \|Ax_j + By_j - b\| \leq \mathcal{O}\left(\sqrt{\frac{\Delta_\beta^0}{\beta\theta k}}\right).$$

Proof. Since $B^*\lambda_0 = \nabla g(y_0)$, it follows from Lemma 2.1(i) that $\eta_0 = 0$. Moreover, since B^*B is invertible, we have $\sigma_B = \sigma_B^+$. The conclusion that $\Delta_\beta^0 \geq 0$ follows from Lemma 3.8 with $k = 0$, and the fact that $\eta_0 = 0$. Moreover, inequality (2.13) yields $\gamma L^2 \leq (\sigma_B\beta)^2/24$. Hence, since $\tau = 0$, it follows from the definitions of δ_1 and δ_2 in (2.3) and (2.12), respectively, and inequality (2.13) that

$$\frac{\beta\sigma_B}{8} \leq \delta_1 \leq \frac{\beta\sigma_B}{4}, \qquad \beta\theta \leq \frac{1}{\delta_2} \leq 3\beta\theta.$$

Hence, $\delta_1 = \mathcal{O}(\beta\sigma_B)$ and $1/\delta_2 = \mathcal{O}(\beta\theta)$. Therefore, the desired result trivially follows from the facts that $G = 0$, $\tau = 0$ and $\eta_0 = 0$, and Theorem 2.2. □

3 Proof of Theorem 2.2

This section gives the proof of Theorem 2.2.

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* (2.10)*. Then, for every $k \geq 1$, the following inclusions hold:*

$$0 \in \left[\partial f(x_k) - A^*\hat{\lambda}_k\right] + G(x_k - x_{k-1}), \tag{3.1}$$

$$0 = \left[\nabla g(y_k) - B^*\hat{\lambda}_k\right] + \beta B^*B(y_k - y_{k-1}) + \tau(y_k - y_{k-1}), \tag{3.2}$$

$$0 = [Ax_k + By_k - b] + \frac{1}{\theta\beta}(\lambda_k - \lambda_{k-1}). \tag{3.3}$$

Proof. The optimality conditions for (2.4) and (2.5) 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 (2.10) immediately yield (3.1) and (3.2). Relation (3.3) follows immediately from (2.6). □

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 $\Delta \lambda_0 \in \mathbb{R}^l$ be such that*

$$\tau \Delta y_0 + (1 - 1/\theta) B^* \Delta \lambda_0 = B^* \lambda_0 - \nabla g(y_0). \tag{3.4}$$

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

$$B^* \Delta \lambda_k = (1 - \theta) B^* \Delta \lambda_{k-1} + \theta u_k \tag{3.5}$$

where

$$u_k = \nabla g(y_k) - \nabla g(y_{k-1}) + \tau (\Delta y_k - \Delta y_{k-1}). \tag{3.6}$$

Proof. Using (2.10) and (3.3) 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 (3.2) then imply that

$$B^* \lambda_k = (1 - \theta) B^* \lambda_{k-1} + \theta [\nabla g(y_k) + \tau \Delta y_k] \quad \forall k \geq 1. \tag{3.7}$$

Hence, in view of (3.6), relation (3.5) holds for every $k \geq 2$. Also, (3.6) and (3.7) 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 (3.4), shows that (3.5) 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 (2.1).

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

(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 (2.4), 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) Using the definition of $\mathcal{L}_\beta$ in (2.1), we have

$$\begin{aligned}
&\mathcal{L}_\beta(x_k, y_{k-1}, \lambda_{k-1}) - \mathcal{L}_\beta(x_k, y_k, \lambda_{k-1}) \\
&= g(y_{k-1}) - g(y_k) - \langle \lambda_{k-1}, B(y_{k_1} - y_k) \rangle + \frac{\beta}{2} \|Ax_k + By_{k-1} - b\|^2 - \frac{\beta}{2} \|Ax_k + By_k - b\|^2 \\
&= g(y_{k-1}) - g(y_k) - \langle \lambda_{k-1} - \beta(Ax_k + By_k - b), B(y_{k_1} - y_k) \rangle + \frac{\beta}{2} \|B(y_{k-1} - y_k)\|^2 \\
&= g(y_{k-1}) - g(y_k) - \langle \hat{\lambda}_k, B(y_{k-1} - y_k) \rangle
\end{aligned} \tag{3.8}$$

where the last inequality is due to (2.10). On the other hand, it follows from **(A3)** and (3.2) that

$$g(y_{k-1}) - g(y_k) - \langle \hat{\lambda}_k, B(y_{k-1} - y_k) \rangle \geq -\frac{m}{2} \|y_{k-1} - y_k\|^2 + \frac{\beta}{2} \|B(y_{k-1} - y_k)\|^2 + \tau \|y_{k-1} - y_k\|^2,$$

which, combined with (3.8) and the fact that $\|B \Delta y_k\|^2 \geq \sigma_B \|\Delta y_k\|^2$, proves the desired inequality.

(c) This statement follows from (2.6), the identity $\Delta \lambda_k = \lambda_k - \lambda_{k-1}$ and the fact that (2.1) implies that

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

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 $\{\Delta_\beta^k + \eta_k\}$ where

$$\Delta_\beta^k := \mathcal{L}_\beta(x_k, y_k, \lambda_k) - \bar{\mathcal{L}} \qquad \forall k \geq 0, \tag{3.9}$$

$$\eta_k := \frac{c_1}{2} \|B^* \Delta \lambda_k\|^2 + \left(\frac{\beta \sigma_B + 2\tau - m}{4} \right) \|\Delta y_k\|^2 \qquad \forall k \geq 1, \tag{3.10}$$

and $\bar{\mathcal{L}}$, $\eta_0 = \eta_0(y_0, \lambda_0; \theta)$ and c_1 are as defined in **(A4)**, (2.7) and (2.8), respectively.

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

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

$$\Delta_\beta^k + \eta_k - (\Delta_\beta^{k-1} + \eta_{k-1}) \leq -\frac{1}{2}\|\Delta x_k\|_G^2 + \Theta_k^1 + \Theta_k^2 \tag{3.11}$$

where

$$\Theta_k^1 := \frac{1}{\beta\theta}\|\Delta\lambda_k\|^2 + \frac{c_1}{2}\left(\|B^*\Delta\lambda_k\|^2 - \|B^*\Delta\lambda_{k-1}\|^2\right) \tag{3.12}$$

and

$$\Theta_k^2 := -\left(\frac{\beta\sigma_B + 2\tau - m}{4}\right)\left(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2\right) \tag{3.13}$$

where c_1 is defined in (2.8)*.*

Proof. The proof of the lemma follows by adding the three inequalities given in statements (a), (b) and (c) of Lemma 3.3 and using the definitions of Δ_β^k and η_k in (3.9) and (3.10), respectively. □

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

Lemma 3.5. *Let u_k and Θ_k^1 be as in* (3.6) *and* (3.12)*, respectively. Then,*

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

where γ is defined in (2.2)*.*

Proof. Assumption **(A1)** clearly implies that $\Delta\lambda_k = -\beta\theta(Ax_k + By_k - b) \in \operatorname{Im}(B)$. Hence, it follows from Lemma 1.4 that

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

where $\mathcal{P}_B(\cdot)$ is defined in Subsection 1.1. Hence, in view of (3.5) and (3.12), we have

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

Note that if $\theta = 1$, then (2.8) 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 $\theta \neq 1$. The previous inequality together with the relation $\|s_1 + s_2\|^2 \leq (1+t)\|s_1\|^2 + (1 + 1/t)\|s_2\|^2$ which holds for every $s_1, s_2 \in \mathbb{R}^l$ and $t > 0$ yield

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

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

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

where the last equality is due to (2.8). Hence, in view of (2.2), the conclusion of the lemma follows. □

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

$$\|u_k\|^2 \leq 3(L^2 + \tau^2)(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2).$$

Proof. Noting that (3.5) implies that $u_k \in \operatorname{Im} B^*$ and using assumption **(A2)** and non-expansiveness of the projection operator, we obtain

$$\begin{aligned}\|u_k\|^2 = \|\mathcal{P}_{B^*}(u_k)\|^2 &= \|\mathcal{P}_{B^*}\left(\nabla g(y_k) - \nabla g(y_{k-1})\right) + \tau \mathcal{P}_{B^*}\left(\Delta y_k - \Delta y_{k-1}\right)\|^2 \\ &\leq [L\|\Delta y_k\| + \tau\|\Delta y_k - \Delta y_{k-1}\|]^2 \\ &\leq 3L^2\|\Delta y_k\|^2 + 3\tau^2(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2) \end{aligned} \tag{3.14}$$

where the last inequality follows from the triangle inequality and the relation $(s_1+s_2+s_3)^2 \leq 3s_1^2+3s_2^2+3s_3^2$ for $s_1, s_2, s_3 \in \mathbb{R}$. Therefore, the desired inequality follows trivially from (3.14). □

Finally, the next proposition shows that the sequence $\{\Delta_\beta^k\}$ decreases.

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

$$\Delta_\beta^k + \eta_k - (\Delta_\beta^{k-1} + \eta_{k-1}) \leq -\frac{1}{2}\|\Delta x_k\|_G^2 - \delta_1(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2) \quad \forall k \geq 1$$

where δ_1, Δ_β^k and η_k are as in (2.3), (3.9) *and* (3.10), *respectively.*

Proof. It follows from Lemmas 3.5 and 3.6 that

$$\Theta_k^1 \leq \frac{3\gamma(L^2+\tau^2)}{\beta\sigma_B^+}(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2)$$

and hence, in view of (2.3) and (3.13), we have

$$\begin{aligned}\Theta_k^1 + \Theta_k^2 &\leq \left(\frac{3\gamma(L^2+\tau^2)}{\beta\sigma_B^+} + \frac{m - \beta\sigma_B - 2\tau}{4}\right)(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2) \\ &= -\delta_1(\|\Delta y_k\|^2 + \|\Delta y_{k-1}\|^2)\end{aligned}$$

where the last inequality is due to the definition of δ_1 in (2.3). Hence, the result follows due to (3.11). □

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 $\{\Delta_\beta^k + \eta_k\}$ is nonnegative.

Lemma 3.8. *Let Δ_β^k and η_k be as in* (3.9) *and* (3.10), *respectively. Then,*

$$\Delta_\beta^k + \eta_k \geq 0 \quad \forall k \geq 0. \tag{3.15}$$

Proof. Let us first consider that case $k \geq 1$. Assume for contradiction that there exists an index $k_0 \geq 0$ such that $\Delta_\beta^{k_0+1} + \eta_{k_0+1} < 0$. Since $\{\Delta_\beta^k + \eta_k\}$ is decreasing (see Proposition 3.7), we obtain

$$\sum_{k=1}^{j}(\Delta_\beta^k + \eta_k) \leq \sum_{k=1}^{k_0}(\Delta_\beta^k + \eta_k) + (j - k_0)(\Delta_\beta^{k_0+1} + \eta_{k_0+1}) \quad \forall j > k_0$$

and hence

$$\lim_{j\to\infty}\sum_{k=1}^{j}(\Delta_\beta^k + \eta_k) = -\infty.$$

On the other hand, since $\beta \geq \bar{\beta}$, it follows from (2.1), (2.6), (3.9), (3.10) and assumption **(A4)** that

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

and hence that

$$\sum_{k=1}^{j}(\Delta_\beta^k + \eta_k) \geq \frac{1}{2\beta\theta}\left(\|\lambda_j\|^2 - \|\lambda_0\|^2\right) \geq -\frac{1}{2\beta\theta}\|\lambda_0\|^2 \quad \forall j \geq 1,$$

which yields the desired contradiction. Therefore, (3.15) holds for $k \geq 1$. Now, for the case $k = 0$, the desired inequality follows from the last conclusion and Proposition 3.7 with $k = 1$. □

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

$$\sum_{j=1}^{k} \left(\frac{1}{2}\|\Delta x_j\|_G^2 + \delta_1 \|\Delta y_j\|^2 + \delta_2 \|\Delta \lambda_j\|^2 \right) \leq 3 \max\{\Delta_\beta^0, \eta_0\} \tag{3.16}$$

where δ_1, Δ_β^0 and δ_2 are as defined in (2.3), (2.9) *and* (2.12), *respectively.*

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

$$\sum_{j=1}^{k} \left(\frac{1}{2}\|\Delta x_j\|_G^2 + \delta_1 (\|\Delta y_j\|^2 + \|\Delta y_{j-1}\|^2) \right) \leq \Delta_\beta^0 + \eta_0 \leq 2 \max\{\Delta_\beta^0, \eta_0\} \tag{3.17}$$

which, in particular, implies that

$$\sum_{j=1}^{k} (\|\Delta y_j\|^2 + \|\Delta y_{j-1}\|^2) \leq \frac{2 \max\{\Delta_\beta^0, \eta_0\}}{\delta_1}. \tag{3.18}$$

Due to (3.17), in order to prove (3.16), it suffices to show that

$$\sum_{j=1}^{k} \|\Delta \lambda_j\|^2 \leq \frac{\max\{\Delta_\beta^0, \eta_0\}}{\delta_2}. \tag{3.19}$$

Then, in the remaining part of the proof we will show that (3.19) holds. By rewriting (3.12), 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 such that the pair $(\Delta y_0, \Delta \lambda_0)$ is a solution of (2.7). Hence, using (2.7) and Lemmas 3.5 and 3.6, we obtain

$$\begin{aligned} \sum_{j=1}^{k} \|\Delta \lambda_j\|^2 \leq \beta\theta \left[\frac{c_1}{2} \|B^* \Delta \lambda_0\|^2 + \sum_{j=1}^{k} \Theta_j^1 \right] &\leq \beta\theta\eta_0 + \frac{\theta\gamma}{\sigma_B^+} \sum_{j=1}^{k} \|u_j\|^2 \\ &\leq \beta\theta\eta_0 + \frac{3\theta\gamma(L^2+\tau^2)}{\sigma_B^+} \sum_{j=1}^{k} (\|\Delta y_j\|^2 + \|\Delta y_{j-1}\|^2) \\ &\leq \beta\theta\eta_0 + \frac{6\theta\gamma(L^2+\tau^2)\max\{\Delta_\beta^0, \eta_0\}}{\sigma_B^+ \delta_1} \end{aligned}$$

where the last inequality is due to (3.18). Hence, (3.19) follows from the last inequality and the definition of δ_2 in (2.12). □

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

$$\|\Delta x_j\|_G \leq \sqrt{\frac{6 \max\{\eta_0, \Delta_\beta^0\}}{k}}, \quad \|\Delta y_j\| \leq \sqrt{\frac{3 \max\{\eta_0, \Delta_\beta^0\}}{\delta_1 k}}, \quad \|\Delta \lambda_j\| \leq \sqrt{\frac{3 \max\{\eta_0, \Delta_\beta^0\}}{\delta_2 k}}$$

where δ_1, η_0, Δ_β^0 and δ_2 are as defined in (2.3), (2.7), (2.9) *and* (2.12), *respectively.*

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

We are now ready to prove Theorem 2.2.

Proof of Theorem 2.2: First note that the inclusion (2.11) follows immediately from (3.1). Also, we obtain from (3.2) and (3.3) that

$$\nabla g(y_k) - B^* \hat{\lambda}_k = -(\beta B^* B + \tau)\Delta y_k, \quad A x_k + B y_k - b = -\frac{1}{\beta\theta}\Delta\lambda_k, \quad \forall k \geq 1.$$

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

4 Convergence analysis of the proximal ADMM under Kurdyka-Łojasiewicz property

This section analyzes the convergence of the proximal ADMM under the assumption that a specific potencial function is Kurdyka-Łojasiewicz (K-Ł). The K-Ł property and K-Ł function can be described as follows.

Definition 4.1. Let $T : \mathbb{R}^n \to (-\infty, \infty]$ be a proper lower semicontinuous function.

(a) T is said to have the Kurdyka-Łojasiewicz property at $z^* \in \operatorname{dom} \partial T$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of z^* and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$ such that: i) $\phi(0) = 0$; ii) ϕ is C^1 on $(0, \eta)$; iii) for all $s \in (0, \eta)$, $\phi'(s) > 0$; iv) for all $z \in U \cap \{z \in \mathbb{R}^n : T(z^*) < T(z) < T(z^*) + \eta\}$, the Kurdyka-Łojasiewicz inequality holds

$$\phi'(T(z) - T(z^*)) \operatorname{dist}(0, \partial T(z)) \geq 1.$$

(b) If T has the KŁ property at each point of $\operatorname{dom} \partial T$, then T is called a K-Ł function.

We refer the reader to [2] and references therein for examples of K-Ł functions. We first show the convergence of the sequence $\{(x_k, y_k, \lambda_k)\}$ generated by the proximal ADMM assuming that it is bounded. Subsequently, we discuss a case in which this boundedness can be ensured.

Proposition 4.2. *Let $\{(x_k, y_k, \lambda_k)\}$ be generated by the proximal ADMM with $\tau = 0$. Assume that G is positive definite and define*

$$T(x, y, \lambda) := \mathcal{L}_\beta(x, y, \lambda) + 3c_1\theta^2\beta^2\|B^*(Ax + By + b)\|^2/2, \tag{4.1}$$

where c_1 is as in (2.8). Then there exist $\kappa_1, \kappa_2 > 0$ such that, for every $k \geq 1$,

$$T(x_k, y_k, \lambda_k) + \kappa_1 \left(\|\Delta x_k\|^2 + \|\Delta y_k\|^2 + \|\Delta\lambda_k\|^2\right) \leq T(x_{k-1}, y_{k-1}, \lambda_{k-1}) \tag{4.2}$$

and there exists $(w_k^x, w_k^y, w_k^\lambda) \in \partial T(x_k, y_k, \lambda_k)$ satisfying $\|(w_k^x, w_k^y, w_k^\lambda)\| \leq \kappa_2\|(\Delta x_k, \Delta y_k, \Delta\lambda_k)\|$. Additionally, if T is a KŁ function and $\{(x_k, y_k, \lambda_k)\}$ is bounded, then $\{(x_k, y_k, \lambda_k)\}$ converges to a critical point of problem (1.1).

Proof. From (4.1), Lemma 3.3 with $\tau = 0$ and (2.6), we have

$$\begin{aligned} T(x_k, y_k, \lambda_k) - T(x_{k-1}, y_{k-1}, \lambda_{k-1}) \leq & -\frac{\|\Delta x_k\|_G^2}{2} + \frac{(m - \beta\sigma_B)}{2}\|\Delta y_k\|^2 - 2\frac{\|\Delta\lambda_k\|^2}{\theta\beta} \\ & + 3\frac{\|\Delta\lambda_k\|^2}{\theta\beta} + \frac{3c_1}{2}\left(\|B^*\Delta\lambda_k\|^2 - \|B^*\Delta\lambda_{k-1}\|^2\right). \end{aligned} \tag{4.3}$$

Using Lemma 3.5 and the facts that $\tau = 0$ and $\|\nabla g(y_k) - \nabla g(y_{k-1})\|^2 \leq L^2\|\Delta y_k\|^2$ (see the first inequality in (3.14) with $\tau = 0$), we obtain

$$\frac{\|\Delta\lambda_k\|^2}{\theta\beta} + \frac{c_1}{2}\left(\|B^*\Delta\lambda_k\|^2 - \|B^*\Delta\lambda_{k-1}\|^2\right) \leq \frac{\gamma}{\beta\sigma_B^+}\|\nabla g(y_k) - \nabla g(y_{k-1})\|^2 \leq \frac{\gamma L^2}{\beta\sigma_B^+}\|\Delta y_k\|^2,$$

which, combined with (4.3), yields

$$T(x_k, y_k, \lambda_k) - T(x_{k-1}, y_{k-1}, \lambda_{k-1}) \leq -\frac{\|\Delta x_k\|_G^2}{2} - \left[\frac{(\beta\sigma_B - m)}{2} - \frac{3\gamma L^2}{\beta\sigma_B^+}\right]\|\Delta y_k\|^2 - 2\frac{\|\Delta\lambda_k\|^2}{\theta\beta}.$$

Since G is positive definite, the last inequality and (2.3) with $\tau = 0$ imply that there exists $\kappa_1 > 0$ such that (4.2) holds. Now, it follows from (3.1)–(3.3), (2.10) and some algebraic manipulations that

$$\begin{aligned} w_k^x &:= -3c_1\theta\beta A^*(BB^* + I)\Delta\lambda_k - G\Delta x_k \in \partial_x T(x_k, y_k, \lambda_k) \\ w_k^y &:= -3c_1\theta\beta B^*(BB^* + I)\Delta\lambda_k - \beta B^*B\Delta y_k = \nabla_y T(x_k, y_k, \lambda_k) \\ w_k^\lambda &:= \Delta\lambda_k/\theta\beta = \nabla_\lambda T(x_k, y_k, \lambda_k), \end{aligned}$$

and hence the second statement of the proposition easily follows. In order to prove the last statement of the proposition, note the boundedness of $\{(x_k, y_k, \lambda_k)\}$ implies that there exists a subsequence $\{(x_{k_j}, y_{k_j}, \lambda_{k_j})\}$ converging to some $(\bar{x}, \bar{y}, \bar{\lambda})$. Now, in view of (2.4), we have

$$\mathcal{L}_\beta(x_{k_j}, y_{k_j-1}, \lambda_{k_j-1}) \leq \mathcal{L}_\beta(\bar{x}, y_{k_j-1}, \lambda_{k_j-1}) + \|\bar{x} - x_{k_j-1}\|_G^2/2,$$

which, combined with the fact that $(\|\Delta x_k\|, \|\Delta y_k\|, \|\Delta\lambda_k\|) \to (0, 0, 0)$ (see (4.2)), yields $\limsup_{j\to\infty} f(x_{k_j}) \leq f(\bar{x})$. Hence, since f is lower semi continuous, we obtain $\lim_{j\to\infty} f(x_{k_j}) = f(\bar{x})$. Thus, using (4.1) and the fact that g is continuous, we conclude that $T(x_{k_j}, y_{k_j}, \lambda_{k_j}) \to T(\bar{x}, \bar{y}, \bar{\lambda})$ as $j \to \infty$. Hence, the desired result follows from the first part of the proposition and [2, Theorem 2.9]. □

Proposition 4.3. *Assume that $B = I$, f is coercive, i.e., $\lim_{\|x\| \to +\infty} f(x) = +\infty$, and $\bar{g} := \inf_y g(y) > -\infty$. Then the sequence $\{(x_k, y_k, \lambda_k)\}$ generated by the proximal ADMM with $\tau = 0$ is bounded.*

Proof. Since ∇g is L-Lipschitz continuous (see **(A2)** with $B = I$) we have

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

Hence, using $y = y_k$ and $y' = y_k - (1/L)\nabla g(y_k)$, and the definition of $\bar{g}$, we obtain

$$\bar{g} \leq g(y_k - (1/L)\nabla g(y_k)) \leq g(y_k) - \frac{1}{2L}\|\nabla g(y_k)\|^2. \tag{4.4}$$

On the other hand, the optimality condition for (2.5) yields

$$\lambda_k = \nabla g(y_k) + (1 - \theta)\beta(Ax_k + y_k - b).$$

Since (3.16) implies that $\sum_{j=1}^{k} \|\Delta\lambda_j\|^2 < \infty$, we obtain, from (2.6), that $\{Ax_k + By_k - b\}$ is bounded. Thus, we conclude that there exists $\kappa > 0$ such that

$$\|\lambda_k\|^2 \leq 2\|\nabla g(y_k)\|^2 + 2\kappa\beta. \tag{4.5}$$

Hence, using the decreasing property of T (see (4.2)) together with the latter inequality, we obtain

$$\begin{aligned} T(x_0, y_0, \lambda_0) &\geq T(x_k, y_k, \lambda_k) \geq \mathcal{L}_\beta(x_k, y_k, \lambda_k) \\ &= f(x_k) + \frac{\beta}{2}\left\|[Ax_k + y_k - b] - \frac{\lambda_k}{\beta}\right\|^2 + g(y_k) - \frac{\|\lambda_k\|^2}{2\beta} \\ &\geq f(x_k) + \frac{\beta}{2}\left\|[Ax_k + y_k - b] - \frac{\lambda_k}{\beta}\right\|^2 + g(y_k) - \frac{\|\nabla g(y_k)\|^2}{\beta} - \kappa. \end{aligned}$$

Since $\gamma \geq 1$ (see (2.2)), it follows from (2.3) that $\beta > 2L$. So, the last inequality and (4.4) imply that

$$T(x_0, y_0, \lambda_0) \geq f(x_k) + \frac{\beta}{2}\left\|[Ax_k + y_k - b] - \frac{\lambda_k}{\beta}\right\|^2 + \bar{g} - \kappa.$$

Therefore, the last inequality together with the coerciveness of f and boundedness of $\{Ax_k + y_k - b\}$ imply that $\{(x_k, y_k, \lambda_k)\}$ is bounded. □

5 Concluding remark

In this paper, we have established convergence rate bounds for the proximal ADMM for solving nonconvex linearly constrained optimization problems. In this study, the stepsize parameter included in the Lagrange multiplier updating can be chosen in the interval $(0, 2)$ instead of the classical one $(0, (\sqrt{5} + 1)/2)$.

Due to the possible nonconvexity of the objective function of (1.1), a sufficiently large penalty parameter was required in the analysis of the method. This kind of assumption is quite common in the nonconvex ADMM literature. Although a large penalty parameter may compromise the performance of the method, some proximal ADMM variants have been demonstrated to be efficient for solving some nonconvex problems; see, for instance, [23, 24, 32]. We also mention that some assumptions related to the matrix B (see Corollary 2.3 and Propositio 4.3) may imply that (1.1) can be reduced to an unconstrained problem and then proximal gradient type methods can be used to solve it. Even in this case, the use of the proximal ADMM may be interesting; see, for example, [23].

Acknowledgments. The authors would like to thank the two anonymous referees for their careful reading and insightful comments on earlier drafts of this paper.

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] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Math. Program.*, 137(1):91–129, 2013.

[3] 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.

[4] 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.

[5] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *J. Sci. Comput.*, 66(3):889–916, 2016.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] M. L. N. Gonçalves, J. G. Melo, and R. D. C. Monteiro. Extending the ergodic convergence rate of the proximal ADMM. *Avaliable on https://arxiv.org/abs/1611.02903.*

[13] 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. *SIAM J. Optim.*, 27(1):379–407, 2017.

[14] Y. Gu, B. Jiang, and H. Deren. A semi-proximal-based strictly contractive Peaceman-Rachford splitting method. *Avaliable on https://arxiv.org/abs/1506.02221.*

[15] K. Guo, D. R. Han, and T. T. Wu. Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. *Int. J. Comput. Math.*, 94(8):1653–1669, 2017.

[16] 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.

[17] B. He and F. Ma. Convergence study on the proximal alternating direction method with larger step size. *Avaliable on http://www.optimization-online.org/DB_FILE/2017/02/5856.pdf.*

[18] B. He, F. Ma, and X. Yuan. Convergence study on the symmetric version of ADMM with larger step sizes. *SIAM J. Imaging Sci.*, 9(3):1467–1501, 2016.

[19] 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.

[20] 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.

[21] M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: algorithms, convergence, and applications. *Avaliable on https://arxiv.org/abs/1604.00543.*

[22] 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.

[23] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Avaliable on https://arxiv.org/abs/1605.02408.*

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

[25] A. P. Liavas and N. D. Sidiropoulos. Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers. *IEEE Trans. Signal Process.*, 63(20):5450–5463, 2015.

[26] T. Lin, S. Ma, and S. Zhang. An extragradient-based alternating direction method for convex minimization. *Found. Comput. Math.*, 17(1):17–35, 2017.

[27] 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.

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

[29] 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.

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

[31] F. Wang, W. Cao, and Z. Xu. Convergence of multi-block Bregman ADMM for nonconvex composite problems. *Sci. China Inf. Sci.*, 61(12):122101, 2018.

[32] F. Wang, Z. Xu, and H. K. Xu. Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems. *Avaliable on https://arxiv.org/abs/1410.8625*.

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

[34] Z. Wen, X. Peng, X. Liu, X. Sun, and X. Bais. Asset allocation under the Basel accord risk measures. *Avaliable on https://arxiv.org/abs/1308.1321*.

[35] 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.

[36] W. Yin Y. Wang and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Avaliable on https://arxiv.org/abs/1511.06324*.

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

[38] 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. *SIAM J. Imaging Sci.*, 10(1):74–110, 2017.

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

[40] 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.

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