Math. Program., Ser. A (2017) 162:165–199
DOI 10.1007/s10107-016-1034-2

FULL LENGTH PAPER

On the linear convergence of the alternating direction method of multipliers

Mingyi Hong[3] · **Zhi-Quan Luo**[1,2]

Received: 20 August 2012 / Accepted: 24 May 2016 / Published online: 6 July 2016

Abstract We analyze the convergence rate of the alternating direction method of multipliers (ADMM) for minimizing the sum of two or more nonsmooth convex separable functions subject to linear constraints. Previous analysis of the ADMM typically assumes that the objective function is the sum of only *two* convex functions defined on *two* separable blocks of variables even though the algorithm works well in numerical experiments for three or more blocks. Moreover, there has been no rate of convergence analysis for the ADMM without strong convexity in the objective function. In this paper we establish the global R-linear convergence of the ADMM for minimizing the sum of *any* number of convex separable functions, assuming that a certain error bound condition holds true and the dual stepsize is sufficiently small. Such an error bound condition is satisfied for example when the feasible set is a compact polyhedron and the objective function consists of a smooth strictly convex function composed with a linear mapping, and a nonsmooth ℓ_1 regularizer. This result implies

Dedicated to the fond memories of a close friend and collaborator, Paul Y. Tseng.

This research is also supported in part by Supported by NSFC, Grant No. 61571384, and by the Leading Talents of Guang Dong Province program, Grant No. 00201510.

✉ Zhi-Quan Luo
luozq@umn.edu

Mingyi Hong
mingyi@iastate.edu

1 Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

2 School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

3 Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011, USA

the linear convergence of the ADMM for contemporary applications such as LASSO without assuming strong convexity of the objective function.

Keywords Linear convergence · Alternating directions of multipliers · Error bound · Dual ascent

Mathematics Subject Classification 49 · 90

1 Introduction

Consider the problem of minimizing a separable potentially nonsmooth convex function subject to linear equality constraints:

$$\begin{array}{ll} \text{minimize} & f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K) \\ \text{subject to} & Ex = E_1x_1 + E_2x_2 + \cdots + E_Kx_K = q \\ & x_k \in X_k, \quad k = 1, 2, \ldots, K, \end{array} \tag{1.1}$$

where each f_k is a nonsmooth convex function (possibly with extended values), $x = (x_1^T, \ldots, x_K^T)^T \in \Re^n$ is a partition of the optimization variable x, $X = \prod_{k=1}^K X_k$ is the feasible set for x, and $E = (E_1, E_2, \ldots, E_K) \in \Re^{m\times n}$ is an appropriate partition of matrix E (consistent with the partition of x) and $q \in \Re^m$ is a vector. Notice that the model (1.1) can easily accommodate general linear inequality constraints $Ex \geq q$ by adding one extra block. In particular, we can introduce a slack variable $x_{K+1} \geq 0$ and rewrite the inequality constraint as $Ex - x_{K+1} = q$. The constraint $x_{K+1} \geq 0$ can be enforced by adding a new convex component function $f_{K+1}(x_{K+1}) = i_{\Re^m_+}(x_{K+1})$ to the objective function $f(x)$, where $i_{\Re^m_+}(x_{K+1})$ is the indicator function for the nonnegative orthant $\Re^m_+$

$$i_{\Re^m_+}(x_{K+1}) = \begin{cases} 0, & \text{if } x_{K+1} \geq 0 \quad \text{(entry wise)}, \\ \infty, & \text{otherwise.} \end{cases}$$

In this way, the inequality constrained problem with K blocks is reformulated as an equivalent equality constrained convex minimization problem with $K + 1$ blocks.

Optimization problems of the form (1.1) arise in many emerging applications involving structured convex optimization. For instance, in compressive sensing applications, we are given an observation matrix A and a noisy observation vector $b \approx Ax$. The goal is to estimate the sparse vector x by solving the following ℓ_1 regularized linear least squares problem:

$$\begin{array}{ll} \text{minimize} & \|z\|^2 + \lambda\|x\|_1 \\ \text{subject to} & Ax + z = b, \end{array}$$

where $\lambda > 0$ is a penalty parameter. Clearly, this is a structured convex optimization problem of the form (1.1) with $K = 2$. If the variable x is further constrained to be nonnegative, then the corresponding compressive sensing problem can be formulated as a three block ($K = 3$) convex separable optimization problem (1.1) by introducing a slack variable.

A popular approach to solving the separable convex optimization problem (1.1) is to attach a Lagrange multiplier vector y to the linear constraints $Ex = q$ and add a quadratic penalty, thus obtaining an augmented Lagrangian function of the form

$$L(x; y) = f(x) + \langle y, q - Ex\rangle + \frac{\rho}{2}\|q - Ex\|^2, \tag{1.2}$$

where $\rho \geq 0$ is a constant. The augmented dual function is given by

$$d(y) = \min_{x\in X} \; f(x) + \langle y, q - Ex\rangle + \frac{\rho}{2}\|q - Ex\|^2 \tag{1.3}$$

and the dual problem (equivalent to (1.1) under mild conditions) is

$$\max_y d(y). \tag{1.4}$$

Moreover, if $\rho > 0$, then Ex is constant over the set of minimizers of (1.3) (see Lemma 2.1 in Sect. 2). This implies that the dual function $d(y)$ is differentiable with

$$\nabla d(y) = q - Ex(y)$$

where $x(y)$ is a minimizer of (1.3); see Lemma 2.1 for a proof of this claim. Given the differentiability of $d(y)$, it is natural to consider the following dual ascent method to solve the primal problem (1.1)

$$y := y + \alpha\nabla d(y) = y + \alpha(q - Ex(y)), \tag{1.5}$$

where $\alpha > 0$ is a suitably chosen stepsize. Such a dual ascent strategy is well suited for structured convex optimization problems that are amenable to decomposition. For example, if the objective function f is separable (i.e., of the form given in (1.1)) and if we select $\rho = 0$, then the minimization in (1.3) decomposes into K independent minimizations whose solutions frequently can be obtained in a simple form. In addition, the iterations can be implemented in a manner that exploits the sparsity structure of the problem and, in certain network cases, achieve a high degree of parallelism. Popular choices for the ascent methods include (single) coordinate ascent (see [3,7,9,31,38,40,46,47,53]), gradient ascent (see [31,40,48]) and gradient projection [22,29]. (See [4,31,44] for additional references.)

For large scale optimization problems, it is numerically advantageous to select $\rho > 0$. Unfortunately, this also introduces variable coupling in the augmented Lagrangian (1.2), which makes the exact minimization step in (1.3) no longer decomposable across variable blocks even if f has a separable structure. In this case, it is more economical to minimize (1.3) inexactly by updating the components of x cyclically via the coordinate descent method. In particular, we can apply the Gauss–Seidel strategy to inexactly minimize (1.3), and then update the multiplier y using an approximate optimal solution of (1.3) in a manner similar to (1.5). The resulting algorithm is called the Alternating Direction Method of Multipliers (ADMM) and is summarized in the

following table (see [16–19]). In the general context of sums of monotone operators, the work of [15] describes a large family of splitting methods for $K \geq 3$ blocks which, when applied to the dual, result in similar but not identical methods to the ADMM algorithm (1.6).

Alternating Direction Method of Multipliers (ADMM)

At each iteration $r \geq 1$, we first update the primal variable blocks in the Gauss–Seidel fashion and then update the dual multiplier using the updated primal variables:

$$\begin{cases} x_k^{r+1} = \arg\min\limits_{x_k \in X_k} L\left(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, \ldots, x_K^r; y^r\right), \ k = 1, 2, \ldots, K, \\ y^{r+1} = y^r + \alpha(q - Ex^{r+1}) = y^r + \alpha\left(q - \sum\limits_{k=1}^K E_k x_k^{r+1}\right), \end{cases} \tag{1.6}$$

where $\alpha > 0$ is the stepsize for the dual update.

Notice that if there is only one block ($K = 1$) and $\rho = 0$, then the ADMM reduces to the standard dual gradient ascent method (see e.g., [1]). In particular, it is known that, under mild assumptions on the problem and with certain assumptions on the stepsize, this type of dual gradient ascent methods generate a sequence of iterates whose limit points must be optimal solutions of the original problem (see [7,44,46]). For the special case of ordinary network flow problems, it is further known that an associated sequence of dual iterates converges to an optimal solution of the dual (see [3]). The rate of convergence of dual ascent methods has been studied in the reference [34] which showed that, under mild assumptions on the problem, the distance to the optimal dual solution set from any $y \in \Re^m$ near the set is bounded above by the dual optimality 'residual' $\|\nabla d(y)\|$. By using this bound, it can be shown that a number of ascent methods, including coordinate ascent methods and a gradient projection method, converge at least linearly when applied to solve the dual problem (see [32,33]; also see [2,10,27] for related analysis). (Throughout this paper, by 'linear convergence' we mean root-linear convergence, denoted by R-linear convergence, in the sense of Ortega and Rheinboldt [39].)

When there are two blocks ($K = 2$), the convergence of the ADMM was studied in the context of Douglas–Rachford splitting method [12–14] for finding a zero of the sum of two maximal monotone operators. It is known that in this case every limit point of the iterates is an optimal solution of the problem. The recent work of [20,21,25,37] have shown that, under some additional assumptions, the objective values generated by the ADMM algorithm and its accelerated version (which performs some additional line search steps for the dual update) converge at a rate of $O(1/r)$ and $O(1/r^2)$ respectively, where r is the iteration index. Moreover, if the objective function $f(x)$ is strongly convex and the constraint matrix E is row independent, then the ADMM is known to converge linearly to the unique minimizer of (1.1) [30]. [One notable exception to the strong convexity requirement is in the special case of linear programming for which the ADMM is linearly convergent [13].] More recent convergence rate analysis of the ADMM still requires at least one of the component

functions ($f_1(x_1)$ or $f_2(x_2)$) to be strongly convex and have a Lipschitz continuous gradient. Under these and additional rank conditions on the constraint matrix E, some linear convergence rate results can be obtained for a subset of primal and dual variables in the ADMM algorithm (or its variant); see [5,11,23]. However, when there are more than two blocks involved ($K \geq 3$), the convergence (or the rate of convergence) of the ADMM method is unknown, and this has been a key open question for several decades. The recent work [35] describes a list of novel applications of the ADMM with $K \geq 3$ and motivates strongly for the need to analyze the convergence of the ADMM in the multi-block case. The recent monograph [6] contains more details of the history, convergence analysis and applications of the ADMM and related methods.

A main contribution of this paper is to establish the global (linear) convergence of the ADMM method for a class of convex objective functions involving any number of blocks (K is arbitrary). Two key requirements for the global (linear) convergence are the choice of a sufficiently small dual stepsize and the satisfaction of a certain error bound condition that is similar to that used in the analysis of [34]. This error bound estimates the distance from an iterate to the optimal solution set in terms of a certain proximity residual. The class of problems that are known to satisfy this error bound condition has a bounded polyhedral feasible set, and an objective that can be written as $f(x) = \ell(Ax)+h(x)$, where $\ell(\cdot)$ is a smooth and strictly convex function, A is a linear mapping not necessarily of full column rank, and $h(x)$ is the ℓ_1 regularization term. This family of problem includes many contemporary applications such as LASSO as special cases.

2 Technical preliminaries

Let f be a closed proper convex function in $\Re^n$, let E be an $m \times n$ matrix, let q be a vector in $\Re^m$. Let dom f denote the effective domain of f and let int(dom f) denote the interior of dom f. We make the following standing assumptions regarding f:

Assumption (a) The global minimum of (1.1) is attained and so is its dual optimal value. The intersection $X \cap \text{int}(\text{dom } f) \cap \{x \mid Ex = q\}$ is nonempty.

(b) $f = f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K)$, with each f_k further decomposable as

$$f_k(x_k) = g_k(A_k x_k) + h_k(x_k)$$

where g_k and h_k are both convex and continuous relative to their domains, and A_k's are some given matrices (not necessarily full column rank, and can be zero).

(c) Each g_k is strictly convex and continuously differentiable on int(dom g_k) with a uniform Lipschitz continuous gradient

$$\left\| A_k^T \nabla g_k(A_k x_k) - A_k^T \nabla g_k(A_k x_k') \right\| \leq L \|x_k - x_k'\|, \qquad \forall\, x_k, x_k' \in X_k$$

where $L > 0$ is a constant.

(d) The epigraph of each $h_k(x_k)$ is a polyhedral set.

(e) Each g_k and h_k is a proper convex function.

(f) Each submatrix E_k has full column rank.
(g) The feasible sets X_k, $k = 1, \ldots, K$ are compact polyhedral sets.

We have the following remarks regarding to the assumptions made.

1. Each f_k might only consist of the convex function h_k. That is, the strictly convex part g_k may effectively be absent by having $A_k = 0$. Also, since the matrices A_k's are not required to have full column rank, the overall objective function $f(x)$ is not necessarily strictly convex. In fact, under Assumption A, the optimization problem (1.1) can still have multiple primal or dual optimal solutions. This makes the convergence (and rate of convergence) analysis of ADMM difficult.
2. Assumption (d) allows h_k to be a simple linear function of the form $\langle b_k, x_k \rangle$, as its ephigraph is polyhedral. Moreover, from the assumption that X_k is polyhedral, the feasibility constraint $x_k \in X_k$ can be absorbed into h_k by adding to it an indicator function $i_{X_k}(x_k)$. To simplify notations, we will not explicitly write $x_k \in X_k$ in the ADMM update (1.6) from now on.
3. Assumption (f) is made to ensure that the subproblems for each x_k is strongly convex. This assumption will be relaxed later when the subproblems are solved inexactly; see Sect. 4.1.
4. Assumption (g) requires the feasible set of the variables to be compact. This condition is not needed in conventional analysis of ADMM, but is required here to ensure that certain error bounds of the primal and dual problems of (1.1) hold. This assumption is usually satisfied in practical applications (e.g. the consensus problems) whenever a priori knowledge on the variable domain is available. This assumption can be further relaxed; see the discussion at the end of Sect. 3. Additionally, with x being in a compact set, one can add a nonnegative *bounded* slack variable x_{K+1} and transform a general linear inequality constraint $Ex \geq q$ into a linear equality constraint.

Under Assumption (g), the feasible set of problem (1.1) is a polyhedral set. Therefore both the primal optimum and the dual optimum values of (1.1) are attained and are equal (i.e., strong duality holds for (1.1)) so that

$$\begin{aligned} d^* &= \max_y \min_{x \in X} L(x; y) = \max_y \left(f(x^*) + \langle y, q - Ex^* \rangle + \frac{\rho}{2} \|Ex^* - q\|^2 \right) \\ &= f(x^*) = \min_{x \in X,\ Ex = q} f(x), \end{aligned}$$

where d^* is the optimal value of the dual of (1.1), and x^* is the optimal solution for (1.1). To see why strong duality holds, we first note that the feasible sets are assumed to be polyhedral, so the feasible set of the entire problem can be written compactly as $Cx \geq b$ for some matrix C and vector b. Let x^* be the optimal solution. By Assumption (a), the intersection $X \cap \text{int}(\text{dom } f) \cap \{x \mid Ex = q\}$ is nonempty, this constraint qualification implies that the existence of multiplier vector y^* and z^* satisfying KKT condition:

$$\begin{aligned} &\xi^* - E^T y^* - C^T z^* = 0,\ \langle z^*, Cx^* - b \rangle = 0,\ z^* \geq 0 \\ &Ex^* - q = 0,\ Cx^* \geq b, \quad \text{for some } \xi^* \in \partial f(x^*). \end{aligned} \tag{2.1}$$

By the convexity of the objective function f, this KKT condition implies that

$$x^* = \arg\min_{Cx \geq b} \left(f(x) + \langle y^*, q - Ex \rangle + \frac{\rho}{2} \|Ex - q\|^2 \right). \tag{2.2}$$

Thus, we have from the definition of $d(y^*)$ that

$$d(y^*) = \arg\min_{Cx \geq b} \left(f(x) + \langle y^*, q - Ex \rangle + \frac{\rho}{2} \|Ex - q\|^2 \right) = f(x^*), \tag{2.3}$$

which is the desired strong duality.

Under Assumption (a)–(g), there may still be multiple optimal solutions for both the primal problem (1.1) and its dual problem. We first claim that the dual functional

$$d(y) = \min_{x \in X} L(x; y) = \min_{x \in X} \ f(x) + \langle y, q - Ex \rangle + \frac{\rho}{2} \|q - Ex\|^2, \tag{2.4}$$

is differentiable everywhere. Let $X(y)$ denote the set of optimal solutions for (2.4).

Lemma 2.1 *For any $y \in \Re^m$, both Ex and $A_k x_k$, $k = 1, 2, \ldots, K$, are constant over $X(y)$. Moreover, the dual function $d(y)$ is differentiable everywhere and*

$$\nabla d(y) = q - Ex(y),$$

where $x(y) \in X(y)$ is any minimizer of (2.4).

Proof Fix $y \in \Re^m$. We first show that Ex is invariant over $X(y)$. Suppose the contrary, so that there exist optimal solutions x and x' from $X(y)$ with the property that $Ex \neq Ex'$. Then, we have

$$d(y) = L(x; y) = L(x'; y).$$

Due to the convexity of $L(x; y)$ with respect to the variable x, the solution set $X(y)$ must be convex, implying $\bar{x} = (x + x')/2 \in X(y)$. By the convexity of $f(x)$, we have

$$\frac{1}{2} \left[(f(x) + \langle y, q - Ex \rangle) + (f(x') + \langle y, q - Ex' \rangle) \right] \geq f(\bar{x}) + \langle y, q - E\bar{x} \rangle.$$

Moreover, by the strict convexity of $\|\cdot\|^2$ and the assumption $Ex \neq Ex'$, we have

$$\frac{1}{2} \left(\|Ex - q\|^2 + \|Ex' - q\|^2 \right) > \|E\bar{x} - q\|^2.$$

Multiplying this inequality by $\rho/2$ and adding it to the previous inequality yields

$$\frac{1}{2} \left[L(x; y) + L(x'; y) \right] > L(\bar{x}; y),$$

which further implies

$$d(y) > L(\bar{x}; y).$$

This contradicts the definition $d(y) = \min_x L(x; y)$. Thus, Ex is invariant over $X(y)$. Notice that $d(y)$ is a concave function and its subdifferential is given by [1, Section 6.1]

$$\partial d(y) = \text{Closure of the convex hull } \{q - Ex(y) \mid x(y) \in X(y)\}.$$

Since $Ex(y)$ is invariant over $X(y)$, the subdifferential $\partial d(y)$ is a singleton. By Danskin's Theorem, this implies that $d(y)$ is differentiable and the gradient is given by $\nabla d(y) = q - Ex(y)$, for any $x(y) \in X(y)$.

A similar argument (and using the strict convexity of g_k) shows that $A_k x_k$ is also invariant over $X(y)$. The proof is complete. □

By using Lemma 2.1, we show below a Lipschitz continuity property of $\nabla d(y)$, for any y in dom(d).

Lemma 2.2 *For all $y, y' \in$ dom (d), there holds $\|\nabla d(y') - \nabla d(y)\| \leq \frac{1}{\rho}\|y' - y\|$.*

Proof Fix any y and y' in dom (d). Let $x = x(y)$ and $x' = x(y')$ be two minimizers of $L(x; y)$ and $L(x; y')$ respectively. By convexity, we have

$$z - E^T y + \rho E^T(Ex - q) = 0 \quad \text{and} \quad z' - E^T y' + \rho E^T(Ex' - q) = 0,$$

where z and z' are some subgradient vectors in the subdifferential $\partial f(x)$ and $\partial f(x')$ respectively. Thus, we have

$$\langle z - E^T y + \rho E^T(Ex - q), x' - x\rangle = 0$$

and

$$\langle z' - E^T y' + \rho E^T(Ex' - q), x - x'\rangle = 0.$$

Adding the above two equalities yields

$$\langle z - z' + E^T(y' - y) - \rho E^T E(x' - x), x' - x\rangle = 0.$$

Upon rearranging terms and using the convexity property

$$\langle z' - z, x' - x\rangle \geq 0,$$

we get

$$\langle y' - y, E(x' - x)\rangle = \langle z' - z, x' - x\rangle + \rho\|E(x' - x)\|^2 \geq \rho\|E(x' - x)\|^2.$$

Thus, $\rho\|E(x' - x)\| \leq \|y' - y\|$ which together with $\nabla d(y') - \nabla d(y) = E(x - x')$ (cf. Lemma 2.1) yields

$$\|\nabla d(y') - \nabla d(y)\| = \|E(x' - x)\| \leq \frac{1}{\rho}\|y - y'\|.$$

The proof is complete. □

To show the linear convergence of the ADMM method, we need certain local error bounds around the optimal solution set $X(y)$ as well as around the dual optimal solution set Y^*. To describe these local error bounds, we first define the notion of a proximity operator. Let $h : \text{dom}\ (h) \mapsto \Re$ be a (possibly nonsmooth) convex function. For every $x \in \text{dom}\ (h)$, the *proximity operator* of h is defined as [42, Section 31]

$$\text{prox}_h(x) = \underset{u \in \Re^n}{\text{argmin}}\ h(u) + \frac{1}{2}\|x - u\|^2.$$

Notice that if $h(x)$ is the indicator function of a closed convex set X, then

$$\text{prox}_h(x) = \text{proj}_X(x),$$

so the proximity operator is a generalization of the projection operator. In particular, it is known that the proximity operator satisfies the nonexpansiveness property:

$$\|\text{prox}_h(x) - \text{prox}_h(x')\| \leq \|x - x'\|, \quad \forall\, x, x' \in dom h. \tag{2.5}$$

The proximity operator can be used to characterize the optimality condition for a nonsmooth convex optimization problem. Suppose a convex function f is decomposed as $f(x) = g(Ax) + h(x)$ where g is strongly convex and differentiable, h is a convex (possibly nonsmooth) function, then we can define the *proximal gradient* of f with respect to h as

$$\tilde{\nabla} f(x) := x - \text{prox}_h(x - \nabla(f(x) - h(x))) = x - \text{prox}_h(x - A^T \nabla g(Ax)).$$

If $h \equiv 0$, then the proximal gradient $\tilde{\nabla} f(x) = \nabla f(x)$. In general, $\tilde{\nabla} f(x)$ can be used as the (standard) gradient of f for the nonsmooth minimization $\min_{x \in X} f(x)$. For example, $\tilde{\nabla} f(x^*) = 0$ iff x^* is a global minimizer.

For the Lagrangian minimization problem (2.4) and under assumptions (a)–(g), the work of [34,45,52] suggests that the size of the proximal gradient

$$\begin{aligned} \tilde{\nabla}_x L(x; y) &:= x - \text{prox}_h\,(x - \nabla_x(L(x; y) - h(x))) \\ &= x - \text{prox}_h\left(x - A^T \nabla g(Ax) + E^T y - \rho E^T(Ex - q)\right) \end{aligned} \tag{2.6}$$

can be used to upper bound the distance to the optimal solution set $X(y)$ of (2.4). Here

$$h(x) := \sum_{k=1}^{K} h_k(x_k), \quad g(Ax) := \sum_{k=1}^{K} g_k(A_k x_k)$$

represent the nonsmooth and the smooth parts of $f(x)$ respectively.

In our analysis of ADMM, we will also need an error bound for the dual function $d(y)$. Notice that a $y \in \Re^m$ solves (1.4) if and only if y satisfies the system of nonlinear equations

$$\nabla d(y) = 0.$$

This suggests that the norm of the 'residual' $\|\nabla d(y)\|$ may be a good estimate of how close y is from solving (1.4). The next lemma says if the nonsmooth part of f_k takes certain forms, then the distance to the primal and dual optimal solution sets can indeed be bounded.

Lemma 2.3 *Suppose assumptions (a)–(e) hold.*

(a) If in addition X is a polyhedral set, then for any y, there exists positive scalars τ and δ such that the following primal error bound holds

$$dist\,(x, X(y)) \leq \tau \|\tilde{\nabla}_x L(x; y)\|, \tag{2.7}$$

for all x such that $\|\tilde{\nabla}_x L(x; y)\| \leq \delta$, where the proximal gradient $\tilde{\nabla}_x L(x; y)$ is given by (2.6).

(b) If X is polyhedral and compact, then for any y, there exists some $\tau > 0$ such that the error bound (2.7) holds for all $x \in X \cap \mathrm{dom}(h)$.

(c) If assumption (g) also holds, and further if the epigraph of h_k is polyhedral (which includes ℓ_1 norm and indicator function for polyhedral sets), then for any scalar ζ, there exist positive scalars δ and τ such that the following dual error bound holds

$$dist\,(y, Y^*) = \|y - y^*\| \leq \tau \|\nabla d(y)\|, \text{ whenever } d(y) \geq \zeta \text{ and } \|\nabla d(y)\| \leq \delta. \tag{2.8}$$

(d) In all three cases stated above the constant τ is independent *of the choice of y and x.*

Proof For any fixed y, the proof for the first part of Lemma 2.3 is a simple extension to those of [34,45,52], each of which shows the error bound with different objective function f (e.g., smooth composite function in [34], smooth composite plus ℓ_1 penalty in [45]). In particular, it was shown that for any given y, (2.7) holds for all x with $\|\tilde{\nabla}_x L(x; y)\| \leq \delta$ (i.e., sufficiently close to $X(y)$).

The second part of the claim says that for any fixed y, the error bound holds over the compact set $X \cap \mathrm{dom}(h)$. This can be seen in two steps as follows: (1) for all $x \in X \cap \mathrm{dom}(h)$ such that $\|\tilde{\nabla}_x L(x; y)\| \leq \delta$, the error bound (2.7) is already known to hold; (2) for all $x \in X \cap \mathrm{dom}(h)$ such that $\|\tilde{\nabla}_x L(x; y)\| \geq \delta$, the ratio

$$\frac{\mathrm{dist}\,(x, X(y))}{\|\tilde{\nabla}_x L(x; y)\|}$$

is a continuous function and well defined over the compact set $X \cap \mathrm{dom}(h) \cap \left\{x \mid \|\tilde{\nabla}_x L(x; y)\| \geq \delta\right\}$. Thus, the above ratio must be bounded from above by a

constant τ' (independent of y). Combining (1) and (2) yields the desired error bound over the set $X \cap \mathrm{dom}(h)$.

Dual error bounds like the one stated in the third part of the lemma have been studied previously by Pang [41] and by Mangasarian and Shiau [36], though in different contexts. The above error bound is 'local' in that it holds only for those y that are bounded or near Y^* (i.e., when $\|\nabla d(y)\| \leq \delta$ as opposed to a 'global' error bound which would hold for all y in $\Re^m$). However if in addition y also lies in some compact set Y, then the dual error bound hold true for all $y \in Y$ (using the same argument as in the preceding paragraph). The proof of this claim is given in the "Appendix".

The last part of the Lemma 2.3 claims that the constants δ and τ are both independent of the choice of y. This property follows directly from a similar property of Hoffman's error bound [26] (on which the error bounds of [34,45,52] are based) for a feasible linear system $P := \{x \mid Ax \leq b\}$:

$$\mathrm{dist}\,(x, P) \leq \tau \|[Ax - b]_+\|, \quad \forall\, x \in \Re^n,$$

where τ is independent of b. In fact, this property is implicitly shown in the proofs of [34,45,52]. Interested readers are referred to these works for proof. From the proof of the third part of the lemma, it is clear that the value of τ in the dual error bound is indeed independent of x and y. □

As a remark, we mention that both the primal and dual error bounds described in Lemma 2.3 hold true for a wider class of nonsmooth functions than Assumption (d). For example, it can also include the group LASSO penalization $h_k(x_k) = \lambda_k \|x_k\|_1 + \sum_J w_J \|x_{k,J}\|_2$, where $x_k = (\ldots, x_{k,J}, \ldots)$ is a partition of x_k with $w_J \geq 0$ and J being the partition index. The proof of error bound for this type of functions follows [52], and we omit the proof of this more general case for space consideration.

Under Assumption (f), the augmented Lagrangian function $L(x; y)$ (cf. (1.2)) is strongly convex with respect to each subvector x_k. As a result, each alternating minimization iteration of ADMM (1.6)

$$x_k^{r+1} = \underset{x_k}{\mathrm{argmin}}\, L\left(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, \ldots, x_K^r; y^r\right), \quad k = 1, \ldots, K,$$

has a unique optimal solution. Thus the sequence of iterates $\{x^r\}$ of the ADMM are well defined. The following lemma shows that the alternating minimization of the Lagrangian function gives a sufficient descent of the Lagrangian function value.

Lemma 2.4 *Suppose assumptions (b) and (f) hold. Then fix any index r, we have*

$$L(x^r; y^r) - L(x^{r+1}; y^r) \geq \gamma \|x^r - x^{r+1}\|^2, \tag{2.9}$$

where the constant $\gamma > 0$ is independent of r and y^r.

Proof By assumptions (b) and (f), the augmented Lagrangian function

$$L(x; y) = \sum_{k=1}^K (f_k(x_k) + \langle y_k, q_k - E_k x_k\rangle) + \frac{\rho}{2}\left\|\sum_{k=1}^K E_k x_k - q\right\|^2$$

is strongly convex in each variable x_k and has a uniform modulus $\rho\lambda_{\min}(E_k^T E_k) > 0$. Here, the notation $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a symmetric matrix. This implies that, for each k, the following is true

$$L(x; y) - L(x_1, \ldots, x_{k-1}, \bar{x}_k, x_{k+1}, \ldots, x_K; y) \geq \rho\lambda_{\min}\left(E_k^T E_k\right) \|x_k - \bar{x}_k\|^2, \tag{2.10}$$

for all x, where $\bar{x}_k$ is the minimizer of $\min_{x_k} L(x; y)$ (when all other variables $\{x_j\}_{j\neq k}$ are fixed).

Fix any index r. For each $k \in \{1, \ldots, K\}$, by ADMM (1.6), x_k^{r+1} is the minimizer of $L(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, x_{k+2}^r, \ldots, x_K^r; y^r)$. It follows from (2.10)

$$\begin{aligned} &L\left(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, \ldots, x_K^r; y^r\right) - L\left(x_1^{r+1}, \ldots, x_k^{r+1}, x_{k+1}^r, \ldots, x_K^r; y^r\right) \\ &\quad \geq \gamma \left\|x_k^r - x_k^{r+1}\right\|^2, \quad \forall\, k, \end{aligned} \tag{2.11}$$

where

$$\gamma = \rho \min_k \lambda_{\min}\left(E_k^T E_k\right)$$

is independent of r and y^r. Summing this over k, we obtain the sufficient decrease condition

$$L(x^r; y^r) - L(x^{r+1}; y^r) \geq \gamma \|x^r - x^{r+1}\|^2.$$

This completes the proof of Lemma 2.4. □

To prove the linear convergence of the ADMM algorithm, we also need the following lemma which bounds the size of the proximal gradient $\tilde{\nabla} L(x^r; y^r)$ at an iterate x^r.

Lemma 2.5 *Suppose assumptions (b)–(c) hold. Let $\{x^r\}$ be generated by the ADMM algorithm (1.6). Then there exists some constant $\sigma > 0$ (independent of y^r) such that*

$$\|\tilde{\nabla} L(x^r; y^r)\| \leq \sigma \|x^{r+1} - x^r\| \tag{2.12}$$

for all $r \geq 1$.

Proof Fix any $r \geq 1$ and any $1 \leq k \leq K$. According to the ADMM procedure (1.6), the variable x_k is updated as follows

$$x_k^{r+1} = \operatorname*{argmin}_{x_k} \Bigg(h_k(x_k) + g_k(A_k x_k) - \langle y^r, E_k x_k \rangle$$

$$+\frac{\rho}{2}\left\|E_k x_k+\sum_{j<k}E_j x_j^{r+1}+\sum_{j>k}E_j x_j^r-q\right\|^2\Bigg).$$

The corresponding optimality condition can be written as

$$x_k^{r+1}=\operatorname{prox}_{h_k}\Bigg[x_k^{r+1}-A_k^T\nabla_{x_k}g_k\left(A_k x_k^{r+1}\right)+E_k^T y^r-\rho E_k^T\left(\sum_{j\le k}E_j x_j^{r+1}+\sum_{j>k}E_j x_j^r-q\right)\Bigg]. \tag{2.13}$$

Therefore, we have

$$\begin{aligned}
&\left\|x_k^{r+1}-\operatorname{prox}_{h_k}\left(x_k^r-A_k^T\nabla_{x_k}g_k(A_k x_k^r)+E_k^T y^r-\rho E_k^T\left(Ex^r-q\right)\right)\right\|\\
&\quad=\Bigg\|\operatorname{prox}_{h_k}\Bigg[x_k^{r+1}-A_k^T\nabla_{x_k}g_k\left(A_k x_k^{r+1}\right)+E_k^T y^r\\
&\qquad+\rho E_k^T\left(\sum_{j\le k}E_j x_j^{r+1}+\sum_{j>k}E_j x_j^r-q\right)\Bigg]\\
&\qquad-\operatorname{prox}_{h_k}\left(x_k^r-A_k^T\nabla_{x_k}g_k\left(A_k x_k^r\right)+E_k^T y^r+\rho E_k^T\left(Ex^r-q\right)\right)\Bigg\|\\
&\quad\le\Bigg\|\left(x_k^{r+1}-x_k^r\right)-A_k^T\left(\nabla_{x_k}g_k\left(A_k x_k^{r+1}\right)-\nabla_{x_k}g_k\left(A_k x_k^r\right)\right)\\
&\qquad+\rho E_k^T\sum_{j\le k}E_j\left(x_j^{r+1}-x_j^r\right)\Bigg\|\\
&\quad\le\left\|x_k^{r+1}-x_k^r\right\|+L\left\|A_k^T\right\|\|A_k\|\left\|x_k^{r+1}-x_k^r\right\|+\rho\left\|E_k^T\right\|\sum_{j\le k}\|E_j\|\left\|x_j^{r+1}-x_j^r\right\|\\
&\quad\le c\|x^{r+1}-x^r\|,\quad\text{for some } c>0 \text{ independent of } y^r,
\end{aligned}\tag{2.14}$$

where the first inequality follows from the nonexpansive property of the prox operator (2.5), and the second inequality is due to the Lipschitz property of the gradient vector ∇g_k (cf. Assumption (c)). Using this relation and the definition of the proximal gradient $\tilde{\nabla}L(x^r;y^r)$, we have

$$\begin{aligned}
\|\tilde{\nabla}_{x_k}L(x^r;y^r)\|&=\left\|x_k^r-\operatorname{prox}_{h_k}\left(x_k^r-A_k^T\nabla_{x_k}g_k\left(A_k x_k^r\right)+E_k^T y^r-\rho E_k^T\left(Ex^r-q\right)\right)\right\|\\
&\le\left\|x_k^r-x_k^{r+1}\right\|\\
&\quad+\left\|x_k^{r+1}-\operatorname{prox}_{h_k}\left(x_k^r-A_k^T\nabla_{x_k}g_k\left(A_k x_k^r\right)+E_k^T y^r-\rho E_k^T\left(Ex^r-q\right)\right)\right\|\\
&\le(c+1)\|x^{r+1}-x^r\|,\quad\forall\, k=1,2,\ldots,K.
\end{aligned}$$

This further implies that the entire proximal gradient vector can be bounded by $\|x^{r+1} - x^r\|$:

$$\|\tilde{\nabla} L(x^r; y^r)\| \leq (c+1)\sqrt{K}\|x^{r+1} - x^r\|.$$

Setting $\sigma = (c+1)\sqrt{K}$ (which is independent of y^r) completes the proof. □

3 Linear convergence of ADMM

Let d^* denote the dual optimal value and $\{x^r, y^r\}$ be the sequence generated by the ADMM method (1.6). Due to Assumption (a), d^* also equals to the primal optimal value. Further we denote

$$\Delta_d^r = d^* - d(y^r) \tag{3.1}$$

which represents the gap from dual optimality at the rth iteration. The primal gap to optimality at iteration r is defined as

$$\Delta_p^r = L(x^{r+1}; y^r) - d(y^r), \quad r \geq 1. \tag{3.2}$$

Clearly, we have both $\Delta_d^r \geq 0$ and $\Delta_p^r \geq 0$ for all r. See Fig. 1 for an illustration of these gaps.

To establish the linear convergence of ADMM, we need several lemmas to estimate the sizes of the primal and dual optimality gaps as well as their respective decrease.

Let $X(y^r)$ denote the set of optimal solutions for the following optimization problem

$$\min_x L(x; y^r) = \min_x f(x) + \langle y^r, q - Ex\rangle + \frac{\rho}{2}\|Ex - q\|^2.$$

We denote

$$\bar{x}^r = \operatorname*{argmin}_{\bar{x} \in X(y^r)} \|\bar{x} - x^r\|.$$

We first bound the sizes of the dual and primal optimality gaps.

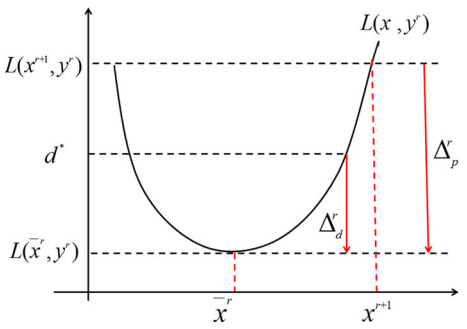

Fig. 1 Illustration of primal and dual gaps

Lemma 3.1 *Suppose assumptions (a)–(e) and (g) hold. Then for any scalar $\delta > 0$, there exists a positive scalar τ' such that*

$$\Delta_d^r \leq \tau' \|\nabla d(y^r)\|^2 = \tau' \|Ex(y^r) - q\|^2, \tag{3.3}$$

for any $y^r \in \Re^m$ with $\|\nabla d(y^r)\| \leq \delta$. Moreover, there exist positive scalars ζ and ζ' (independent of y^r) such that

$$\Delta_p^r \leq \zeta \|x^{r+1} - x^r\|^2 + \zeta' \|x^r - \bar{x}^r\|^2, \quad \textit{for all } r \geq 1. \tag{3.4}$$

Proof Fix any y^r, and let y^* be the optimal dual solution closest to y^r. Then it follows from the mean value theorem that there exists some $\tilde{y}$ in the line segment joining y^r and y^* such that

$$\begin{aligned}
\Delta_d^r &= d(y^*) - d(y^r) \\
&= \langle \nabla d(\tilde{y}), y^* - y^r \rangle \\
&= \langle \nabla d(\tilde{y}) - \nabla d(y^*), y^* - y^r \rangle \\
&\leq \|\nabla d(\tilde{y}) - \nabla d(y^*)\| \|y^* - y^r\| \\
&\leq \frac{1}{\rho} \|\tilde{y} - y^*\| \|y^* - y^r\| \\
&\leq \frac{1}{\rho} \|y^r - y^*\| \|y^* - y^r\| \\
&= \frac{1}{\rho} \|y^* - y^r\|^2
\end{aligned}$$

where the second inequality follows from Lemma 2.2. Recall from Lemma 2.3-(c) there exists some τ such that

$$\text{dist}\,(y^r, Y^*) = \|y^r - y^*\| \leq \tau \|\nabla d(y^r)\|.$$

Combining the above two inequalities yields

$$\Delta_d^r = d(y^*) - d(y^r) \leq \tau' \|\nabla d(y^r)\|^2,$$

where $\tau' = \tau^2/\rho$ is a constant. This establishes the bound on the size of dual gap (3.3).

It remains to prove the bound on the primal gap (3.4). For notational simplicity, let us separate the smooth and nonsmooth part of the augmented Lagrangian as follows

$$\begin{aligned}
L(x; y) &= g(x) + h(x) + \langle y, q - Ex \rangle + \frac{\rho}{2} \|q - Ex\|^2 \\
&:= \bar{L}(x; y) + h(x).
\end{aligned}$$

Let x_k^{r+1} denote the kth subvector of the primal vector x^{r+1}. From the way that the variables are updated (2.13), we have

$$\begin{aligned} x_k^{r+1} &= \text{prox}_{h_k}\left[x_k^{r+1} - \nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j>k\}}; y^r\right)\right] \\ &= \text{prox}_{h_k}\left[x_k^r - \nabla_{x_k}\bar{L}(x^r; y^r) - x_k^r + x_k^{r+1} - \nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{j>k}; y^r\right)\right. \\ &\quad \left. + \nabla_{x_k}\bar{L}(x^r; y^r)\right] \\ &:= \text{prox}_{h_k}\left[x_k^r - \nabla_{x_k}\bar{L}(x^r; y^r) - e_k^r\right] \end{aligned} \tag{3.5}$$

where the gradient vector $\nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j>k\}}; y^r\right)$ can be explicitly expressed as

$$\begin{aligned} \nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j>k\}}; y^r\right) &= A_k^T\nabla_{x_k}g_k\left(A_k x_k^{r+1}\right) - E_k^T y^r \\ &\quad + \rho E_k^T\left(\sum_{j\le k}E_j x_j^{r+1} + \sum_{j>k}E_j x_j^r - q\right) \end{aligned}$$

and the error vector e_k^r is defined by

$$e_k^r := x_k^r - x_k^{r+1} + \nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j>k\}}; y^r\right) - \nabla_{x_k}\bar{L}(x^r; y^r). \tag{3.6}$$

Note that we can bound the norm of e_k^r as follows

$$\begin{aligned} \left\|e_k^r\right\| &\le \left\|x_k^r - x_k^{r+1}\right\| + \left\|\nabla_{x_k}\bar{L}\left(\left\{x_{j\le k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j>k\}}; y^r\right) - \nabla_{x_k}\bar{L}(x^r; y^r)\right\| \\ &\le \left\|x_k^r - x_k^{r+1}\right\| + \left\|A_k^T\left(\nabla_{x_k}g_k\left(A_k x_k^{r+1}\right) - \nabla_{x_k}g_k\left(A_k x_k^r\right)\right)\right. \\ &\quad \left. + \rho E_k^T\left(\sum_{j\le k}E_j\left(x_j^{r+1} - x_j^r\right)\right)\right\| \\ &\le c\|x^r - x^{r+1}\|, \end{aligned} \tag{3.7}$$

where the constant $c > 0$ is independent of y^r, and can take the same value as in (2.14).

Using (3.5), and by the definition of the proximity operator, we have the following

$$\begin{aligned} &h_k\left(x_k^{r+1}\right) + \langle x_k^{r+1} - x_k^r, \nabla_{x_k}\bar{L}(x^r; y^r) + e_k^r\rangle + \frac{1}{2}\left\|x_k^{r+1} - x_k^r\right\|^2 \\ &\quad \le h_k\left(\bar{x}_k^r\right) + \langle \bar{x}_k^r - x_k^r, \nabla_{x_k}\bar{L}(x^r; y^r) + e_k^r\rangle + \frac{1}{2}\left\|\bar{x}_k^r - x_k^r\right\|^2. \end{aligned} \tag{3.8}$$

Summing over all $k = 1, \ldots, K$, we obtain

$$h(x^{r+1}) + \langle x^{r+1} - x^r, \nabla_x \bar{L}(x^r; y^r) + e^r \rangle + \frac{1}{2}\|x^{r+1} - x^r\|^2 \leq h(\bar{x}^r) + \langle \bar{x}^r - x^r, \nabla_x \bar{L}(x^r; y^r) + e^r \rangle + \frac{1}{2}\|\bar{x}^r - x^r\|^2.$$

Upon rearranging terms, we obtain

$$h(x^{r+1}) - h(\bar{x}^r) + \langle x^{r+1} - \bar{x}^r, \nabla_x \bar{L}(x^r; y^r) \rangle \leq \frac{1}{2}\|\bar{x}^r - x^r\|^2 - \langle x^{r+1} - \bar{x}^r, e^r \rangle. \tag{3.9}$$

Also, we have from the mean value theorem that there exists some $\tilde{x}$ in the line segment joining x^{r+1} and $\bar{x}^r$ such that

$$\bar{L}(x^{r+1}; y^r) - \bar{L}(\bar{x}^r; y^r) = \langle \nabla_x \bar{L}(\tilde{x}; y^r), x^{r+1} - \bar{x}^r \rangle.$$

Using the above results, we can bound Δ_p^r by

$$\begin{aligned}
\Delta_p^r &= L(x^{r+1}; y^r) - L(\bar{x}^r; y^r) \\
&= \bar{L}(x^{r+1}; y^r) - \bar{L}(\bar{x}^r; y^r) + h(x^{r+1}) - h(\bar{x}^r) \\
&= \langle \nabla_x \bar{L}(\tilde{x}; y^r), x^{r+1} - \bar{x}^r \rangle + h(x^{r+1}) - h(\bar{x}^r) \\
&= \langle \nabla_x \bar{L}(\tilde{x}; y^r) - \nabla_x \bar{L}(x^r; y^r), x^{r+1} - \bar{x}^r \rangle \\
&\quad + \langle \nabla_x \bar{L}(x^r; y^r), x^{r+1} - \bar{x}^r \rangle + h(x^{r+1}) - h(\bar{x}^r) \\
&\leq \langle \nabla_x \bar{L}(\tilde{x}; y^r) - \nabla_x \bar{L}(x^r; y^r), x^{r+1} - \bar{x}^r \rangle \\
&\quad + \frac{1}{2}\|\bar{x}^r - x^r\|^2 + c\sqrt{K}\|x^{r+1} - x^r\|\|x^{r+1} - \bar{x}^r\| \\
&\leq \left(\sum_{k=1}^{K} L\|A_k\|^T\|A_k\| + \rho\|E^T E\| \right) \|\tilde{x} - x^r\|\|x^{r+1} - \bar{x}^r\| \\
&\quad + \frac{1}{2}\|\bar{x}^r - x^r\|^2 + c\sqrt{K}\|x^{r+1} - x^r\|\|x^{r+1} - \bar{x}^r\| \\
&\leq \left(\sum_{k=1}^{K} L\|A_k\|^T\|A_k\| + \rho\|E^T E\| \right) \left(\|x^{r+1} - x^r\| + \|\bar{x}^r - x^r\| \right)^2 \\
&\quad + \frac{1}{2}\|\bar{x}^r - x^r\|^2 + c\sqrt{K}\|x^{r+1} - x^r\| \left(\|x^{r+1} - x^r\| + \|\bar{x}^r - x^r\| \right) \\
&\leq \zeta\|x^{r+1} - x^r\|^2 + \zeta'\|\bar{x}^r - x^r\|^2, \quad \text{for some } \zeta,\ \zeta' > 0,
\end{aligned}$$

where the first inequality follows from (3.9) and (3.7), the second inequality is due to the Cauchy–Schwartz inequality and the Lipschitz continuity of $\nabla \bar{L}_x(x; y^r)$, while the third inequality follows from the fact that $\tilde{x}$ lies in the line segment joining x^{r+1} and $\bar{x}^r$ so that $\|\tilde{x} - x^r\| \leq \|x^{r+1} - x^r\| + \|\bar{x}^r - x^r\|$. This completes the proof. □

We then bound the decrease of the dual optimality gap.

Lemma 3.2 *For each $r \geq 1$ and for any $\alpha \geq 0$, there holds*

$$\Delta_d^r - \Delta_d^{r-1} \leq -\alpha(Ex^r - q)^T(E\bar{x}^r - q). \tag{3.10}$$

Proof The reduction of the optimality gap in the dual space can be bounded as follows:

$$\begin{aligned}
\Delta_d^r - \Delta_d^{r-1} &= [d^* - d(y^r)] - [d^* - d(y^{r-1})] \\
&= d(y^{r-1}) - d(y^r) \\
&= L(\bar{x}^{r-1}; y^{r-1}) - L(\bar{x}^r; y^r) \\
&= [L(\bar{x}^r; y^{r-1}) - L(\bar{x}^r; y^r)] + [L(\bar{x}^{r-1}; y^{r-1}) - L(\bar{x}^r; y^{r-1})] \\
&= (y^{r-1} - y^r)^T(q - E\bar{x}^r) + [L(\bar{x}^{r-1}; y^{r-1}) - L(\bar{x}^r; y^{r-1})] \\
&= -\alpha(Ex^r - q)^T(E\bar{x}^r - q) + [L(\bar{x}^{r-1}; y^{r-1}) - L(\bar{x}^r; y^{r-1})] \\
&\leq -\alpha(Ex^r - q)^T(E\bar{x}^r - q), \quad \forall\, r \geq 1,
\end{aligned}$$

where the last equality follows from the update of the dual variable y^{r-1}, and the last inequality is from the fact that $\bar{x}^{r-1}$ minimizes $L(\cdot, y^{r-1})$. □

Lemma 3.2 implies that if $q - Ex^r$ is close to the true dual gradient $\nabla d(y^r) = q - E\bar{x}^r$, then the dual optimal gap is reduced after each ADMM iteration. However, since ADMM updates the primal variable by only one Gauss–Seidel sweep, the primal iterate x^r is not necessarily close the minimizer $\bar{x}^r$ of $L(x; y^r)$. Thus, unlike the method of multipliers (for which $x^r = \bar{x}^r$ for all r), there is no guarantee that the dual optimality gap Δ_d^r is indeed reduced after each iteration of ADMM.

Next we proceed to bound the decrease in the primal gap Δ_p^r.

Lemma 3.3 *Suppose assumptions (b) and (f) hold. Then for each $r \geq 1$ and any $\alpha \geq 0$, we have*

$$\Delta_p^r - \Delta_p^{r-1} \leq \alpha\|Ex^r - q\|^2 - \gamma\|x^{r+1} - x^r\|^2 - \alpha(Ex^r - q)^T(E\bar{x}^r - q) \tag{3.11}$$

for some γ independent of y^r.

Proof Fix any $r \geq 1$, we have

$$L(x^r; y^{r-1}) = f(x^r) + \langle y^{r-1}, q - Ex^r\rangle + \frac{\rho}{2}\|Ex^r - q\|^2$$

and

$$L(x^{r+1}; y^r) = f(x^{r+1}) + \langle y^r, q - Ex^{r+1}\rangle + \frac{\rho}{2}\|Ex^{r+1} - q\|^2.$$

By the update rule of y^r (cf. (1.6)), we have

$$L(x^r; y^r) = f(x^r) + \langle y^{r-1}, q - Ex^r\rangle + \frac{\rho}{2}\|Ex^r - q\|^2 + \alpha\|Ex^r - q\|^2.$$

This implies

$$L(x^r; y^r) = L(x^r; y^{r-1}) + \alpha \|Ex^r - q\|^2.$$

Recall from Lemma 2.4 that the alternating minimization of the Lagrangian function gives a sufficient descent. In particular, we have

$$L(x^{r+1}; y^r) - L(x^r; y^r) \leq -\gamma \|x^{r+1} - x^r\|^2,$$

for some $\gamma > 0$ that is independent of r and y^r. Therefore, we have

$$L(x^{r+1}; y^r) - L(x^r; y^{r-1}) \leq \alpha \|Ex^r - q\|^2 - \gamma \|x^{r+1} - x^r\|^2, \quad \forall\, r \geq 1.$$

Hence, we have the following bound on the reduction of primal optimality gap

$$\begin{aligned} \Delta_p^r - \Delta_p^{r-1} &= [L(x^{r+1}; y^r) - d(y^r)] - [L(x^r; y^{r-1}) - d(y^{r-1})] \\ &= [L(x^{r+1}; y^r) - L(x^r; y^{r-1})] - [d(y^r) - d(y^{r-1})] \\ &\leq \alpha \|Ex^r - q\|^2 - \gamma \|x^{r+1} - x^r\|^2 - \alpha (Ex^r - q)^T (E\bar{x}^r - q), \quad \forall\, r \geq 1, \end{aligned}$$

where the last step is due to Lemma 3.2. □

Notice that when $\alpha = 0$ (i.e., no dual update in the ADMM algorithm), Lemma 3.3 reduces to the sufficient decrease estimate (2.9) in Lemma 2.4. When $\alpha > 0$, the primal optimality gap is not necessarily reduced after each ADMM iteration due to the positive term $\alpha \|Ex^r - q\|^2$ in (3.11). Thus, in general, we cannot guarantee a consistent decrease of either the dual optimality gap Δ_d^r or the primal optimality gap Δ_p^r. However, somewhat surprisingly, the sum of the primal and dual optimality gaps decreases for all r, as long as the dual stepsize α is sufficiently small. This is used to establish the linear convergence of ADMM method.

Theorem 3.1 *Suppose assumptions (a)–(g) hold. Then the sequence of iterates* $\{(x^r, y^r)\}$ *generated by the ADMM algorithm* (1.6) *converges linearly to an optimal primal–dual solution for* (1.1)*, provided the stepsize* α *is sufficiently small. Moreover, the sequence of feasibility violation* $\{\|Ex^r - q\|\}$ *also converges linearly.*

Proof We show by induction that the sum of optimality gaps $\Delta_d^r + \Delta_p^r$ is reduced after each ADMM iteration, as long as the stepsize α is chosen sufficiently small. For any $r \geq 1$, we denote

$$\bar{x}^r = \operatorname*{argmin}_{\bar{x} \in X(y^r)} \|\bar{x} - x^r\|. \tag{3.12}$$

By induction, suppose $\Delta_d^{r-1} + \Delta_p^{r-1} \leq \Delta_d^0 + \Delta_p^0$ for some $r \geq 1$. Recall that each X_k is compact and that the indicator function $i_{X_k}(x_k)$ is included in $h_k(x_k)$ (see the discussion after Assumption (a)–(g)), it follows that $x^r \in X$, implying the boundedness of x^r. Thus, we obtain from Lemma 2.3 that

$$\|x^r - \bar{x}^r\| \leq \tau \|\tilde{\nabla} L(x^r; y^r)\| \tag{3.13}$$

for some $\tau > 0$ (independent of y^r). To prove Theorem 3.1, we combine the two estimates (3.10) and (3.11) to obtain

$$\begin{aligned}\left[\Delta_p^r + \Delta_d^r\right] - \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right] &= \left[\Delta_p^r - \Delta_p^{r-1}\right] + \left[\Delta_d^r - \Delta_d^{r-1}\right] \\ &\le \alpha\|Ex^r - q\|^2 - \gamma\|x^{r+1} - x^r\|^2 \\ &\quad - 2\alpha(Ex^r - q)^T(E\bar{x}^r - q) \\ &= \alpha\|Ex^r - E\bar{x}^r\|^2 - \alpha\|E\bar{x}^r - q\|^2 \\ &\quad - \gamma\|x^{r+1} - x^r\|^2. \end{aligned} \tag{3.14}$$

Now we invoke (3.13) and Lemma 2.5 to lower bound $\|x^{r+1} - x^r\|$:

$$\|x^r - \bar{x}^r\| \le \tau\|\tilde{\nabla}L(x^r; y^r)\| \le \tau\sigma\|x^{r+1} - x^r\|. \tag{3.15}$$

Substituting this bound into (3.14) yields

$$\left[\Delta_p^r + \Delta_d^r\right] - \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right] \le \left(\alpha\|E\|^2\tau^2\sigma^2 - \gamma\right)\|x^{r+1} - x^r\|^2 - \alpha\|E\bar{x}^r - q\|^2. \tag{3.16}$$

Thus, if we choose the stepsize α sufficiently small so that

$$0 < \alpha < \gamma\tau^{-2}\sigma^{-2}\|E\|^{-2}, \tag{3.17}$$

then the above estimate shows that

$$\left[\Delta_p^r + \Delta_d^r\right] \le \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right], \tag{3.18}$$

which completes the induction. Moreover, the induction argument shows that if the stepsize α satisfies the condition (3.17), then the descent condition (3.16) holds for all $r \ge 1$.

By the descent estimate (3.16), we have

$$\|x^{r+1} - x^r\| \to 0, \quad \|\nabla d(y^r)\| = \|E\bar{x}^r - q\| \to 0. \tag{3.19}$$

We now show that the sum of optimality gaps $\Delta_d^r + \Delta_p^r$ in fact contracts geometrically after a finite number of ADMM iterations. By (3.19), for any $\delta > 0$, there must exist a finite integer $\bar{r} > 0$ such that for all $r \ge \bar{r}$, $\|\nabla d(y^r)\| \le \delta$. Since Δ_d^r, Δ_p^r are nonnegative and bounded from above (see (3.18)), it follows that $d(y^r)$ is bounded from below by a constant ζ independent of r. Applying Lemma 2.3-(c), we have that for all $r \ge \bar{r}$, the dual error bound $\mathrm{dist}(y^r, Y^*) \le \tau\|\nabla d(y^r)\|$ holds true.

Therefore, it follows from Lemma 3.1 that we have the following cost-to-go estimate

$$\Delta_d^r = d^* - d(y^r) \le \tau'\|\nabla d(y^r)\|^2 = \tau'\|E\bar{x}^r - q\|^2, \tag{3.20}$$

for some $\tau' > 0$ and for all $r \ge \bar{r}$.

Moreover, we can use Lemma 3.1 to bound $\|x^{r+1} - x^r\|^2$ from below by Δ_p^r. In particular, we have from (3.15) and Lemma 3.1 that

$$\begin{aligned}\Delta_p^r &\le \zeta\|x^{r+1} - x^r\|^2 + \zeta'\|\bar{x}^r - x^r\|^2 \\ &\le \zeta\|x^{r+1} - x^r\|^2 + \zeta'\tau^2\sigma^2\|x^{r+1} - x^r\|^2 \\ &= \left(\zeta + \zeta'\tau^2\sigma^2\right)\|x^{r+1} - x^r\|^2.\end{aligned}$$

Substituting this bound and (3.20) into (3.16), and assuming that $\alpha > 0$ satisfies (3.17), we obtain

$$\begin{aligned}\left[\Delta_p^r + \Delta_d^r\right] - \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right] &\le \left(\alpha\|E\|^2\tau^2\sigma^2 - \gamma\right)\|x^{r+1} - x^r\|^2 - \alpha\|E\bar{x}^r - q\|^2 \\ &\le -\frac{(\gamma - \alpha\|E\|^2\tau^2\sigma^2)}{\zeta + \zeta'\tau^2\sigma^2}\Delta_p^r - \alpha(\tau')^{-1}\Delta_d^r \\ &\le -\min\left\{\frac{(\gamma - \alpha\|E\|^2\tau^2\sigma^2)}{\zeta + \zeta'\tau^2\sigma^2}, \alpha(\tau')^{-1}\right\}\left[\Delta_p^r + \Delta_d^r\right].\end{aligned}$$

Since $\alpha > 0$ is chosen small enough such that (3.17) holds, we have

$$\lambda := \min\left\{\frac{\gamma - \alpha\|E\|^2\tau^2\sigma^2}{\zeta + \zeta'\tau^2\sigma^2}, \alpha(\tau')^{-1}\right\} > 0.$$

Consequently, we have

$$\left[\Delta_p^r + \Delta_d^r\right] - \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right] \le -\lambda\left[\Delta_p^r + \Delta_d^r\right]$$

which further implies

$$0 \le \left[\Delta_p^r + \Delta_d^r\right] \le \frac{1}{1+\lambda}\left[\Delta_p^{r-1} + \Delta_d^{r-1}\right].$$

This shows that the sequence $\{\Delta_p^r + \Delta_d^r\}_{r\ge\bar{r}}$ converges to zero Q-linearly.[1] As a result, we conclude that $\{\Delta_p^r + \Delta_d^r\}$ and hence both Δ_p^r and Δ_d^r globally converge to zero R-linearly.[2]

We next show that the dual sequence $\{y^r\}$ as well as the dual objective values $\{d(y^r)\}$ are also R-linearly convergent. To this end, notice that the inequalities (3.15)

[1] A sequence $\{x^r\}$ is said to converge Q-linearly to some $\bar{x}$ if $\|x^{r+1} - \bar{x}\|/\|x^r - \bar{x}\| \le \mu$ for all r, where $\mu \in (0, 1)$ is some constant. A sequence $\{x^r\}$ is said to converge to $\bar{x}$ R-linearly if $\|x^r - \bar{x}\| \le c\mu^r$ for all r and for some $c > 0$.

[2] To see that such R-linear convergence is in fact global, note that $\bar{r} > 0$ is finite, and $\Delta_p^r + \Delta_d^r$ is Q-linearly convergent for $r \ge \bar{r}$. Then one can always find an appropriate constant c such that $\Delta_p^r + \Delta_d^r \le c(1+\lambda)^{-r}$ for all $r = 1, 2, \ldots$.

and (3.16) imply

$$\left[\Delta_p^r + \Delta_d^r\right] - \left[\Delta_p^{r-1} + \Delta_d^{r-1}\right] \leq \left(\alpha\|E\|^2 - \gamma\tau^{-2}\sigma^{-2}\right)\|x^r - \bar{x}^r\|^2 - \alpha\|E\bar{x}^r - q\|^2. \tag{3.21}$$

Then by (3.21), we see that both $\|x^r - \bar{x}^r\| \to 0$ and $\|E\bar{x}^r - q\| \to 0$ R-linearly. This implies that $Ex^r - q \to 0$ R-linearly and $\nabla d(y^r) \to 0$ R-linearly. Using the fact that $\text{dist}(y^r, Y^*) \leq \tau\|\nabla d(y^r)\|$, we conclude that y^r converges R-linearly to an optimal dual solution.

We now argue that the primal iterates $\{x^r\}$ converge to an optimal solution of (1.1). By the inequality (3.16), we can further conclude that

$$\|x^{r+1} - x^r\|^2 \to 0, \quad \|E\bar{x}^r - q\| \to 0$$

R-linearly. Notice that the R-linear convergence of $\|x^{r+1} - x^r\|^2 \to 0$ implies that $\|x^{r+1} - x^r\| \to 0$ R-linearly. This further shows that $x^r \to x^\infty$ R-linearly for some x^∞. Denote the limit of dual sequence $\{y^r\}$ by y^∞. By the preceding argument, we know y^∞ is a dual optimal solution of (1.1). To show that x^∞ is a primal optimal solution of (1.1), it suffices to prove that $x^\infty \in X(y^\infty)$. Using (3.15), and the fact that $\|x^r - \bar{x}^r\| \to 0$, we have

$$\|x^\infty - \bar{x}^r\| \leq \|x^r - x^\infty\| + \|x^r - \bar{x}^r\| \to 0.$$

Since $\bar{x}^r \in X(y^r)$, we have $L(\bar{x}^r, y^r) \leq L(x, y^r)$ for all $x \in X$. Passing limit, we obtain $L(x^\infty, y^\infty) \leq L(x, y^\infty)$ for all $x \in X$, that is, $x^\infty \in X(y^\infty)$. It then follows that the sequence $\{x^r\}$ converges R-linearly to a primal optimal solution. □

Remark Our analysis shows that the ADMM converges globally R-linearly, with a convergence rate explicitly depending upon the error bound constant τ. When the problem is strongly convex, the error bound holds globally, and the constant τ as well as the linear convergence rate can be computed explicitly. In general convex cases, such rate may not be explicitly known, but its existence offers some useful insights to the ADMM algorithm. □

We close this section by providing an example that satisfies the assumptions in Theorem 3.1. Consider the following ℓ_1 minimization problem

$$\min_x \|x\|_1, \quad \text{s.t. } Ex = b,\ a \leq x_k \leq b,\ k = 1, \ldots, K \tag{3.22}$$

which can be equivalently written as a K-block problem

$$\min_{\{x_k\}} \sum_{k=1}^K |x_k|, \quad \text{s.t. } \sum_{k=1}^K e_k x_k = b,\ a \leq x_k \leq b,\ k = 1, \ldots, K, \tag{3.23}$$

where e_k is the kth column of E, a and b are some scalars. It is easy to verify that this problem meets all the conditions listed in assumptions (a)–(g), hence the linear convergence result in Theorem 3.1 applies.

4 Variants of ADMM

The convergence analysis of Sect. 3 can be extended to some variants of the ADMM. We briefly describe two of them below.

4.1 Linearized proximal ADMM

In the original ADMM (1.6), each block x_k is updated by solving a convex optimization subproblem *exactly*. For large scale problems, this subproblem may not be easy to solve unless the matrix E_k is unitary (i.e., $E_k^T E_k = I$) in which case the variables in x_k can be further decoupled (assuming f_k is separable). If the matrix E_k is not unitary, we can still employ a simple proximal gradient step to *inexactly* minimize $L(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, \ldots, x_K^r)$. More specifically, we update each block of x_k according to the following procedure

$$x_k^{r+1} = \arg\min_{x_k} \Big\{ h_k(x_k) + \langle y^r, q - E_k x_k \rangle + \left\langle A_k^T \nabla g_k \left(A_k x_k^r \right), x_k - x_k^r \right\rangle + \frac{\beta}{2} \left\| x_k - x_k^r \right\|^2 + \left\langle \rho E_k^T \Big(\sum_{j<k} E_j x_j^{r+1} + \sum_{j \geq k} E_j x_j^r - q \Big), x_k - x_k^r \right\rangle \Big\} \tag{4.1}$$

in which the smooth part of the objective function in the kth subproblem, namely,

$$g_k(A_k x_k) + \langle y^r, q - E_k x_k \rangle + \frac{\rho}{2} \left\| E_k x_k + \sum_{j<k} E_j x_j^{r+1} + \sum_{j>k} E_j x_j^r - q \right\|^2$$

is linearized locally at x_k^r, and a proximal term $\frac{\beta}{2} \left\| x_k - x_k^r \right\|^2$ is added. Here, $\beta > 0$ is a positive constant. With this change, updating x_k is easy when h_k (the nonsmooth part of f_k) is separable. For example, this is the case for compressive sensing applications where $h_k(x_k) = \|x_k\|_1$, and the resulting subproblem admits a closed form solution given by the component-wise soft thresholding (also known as the shrinkage operator). We note that the linearized proximal ADMM algorithm described here is slightly more general than the proximal ADMM algorithm seen in the literature, in which only the penalization term $\frac{\rho}{2} \left\| E_k x_k + \sum_{j<k} E_j x_j^{r+1} + \sum_{j>k} E_j x_j^r - q \right\|^2$ is linearized locally at x_k^r; see e.g., [49,50].

We claim that Theorem 3.1 holds for the linearized proximal ADMM algorithm *without requiring assumption (f)* (the full rankness of E_k's). Indeed, to establish the (linear) convergence of the linearized proximal ADMM (4.1), we can follow the same proof steps as that for Theorem 3.1, with the only changes being in the proof of Lemmas 2.4–2.5 and Lemma 3.1. We first show that Lemma 2.4 holds without assumption (f). Clearly subproblem (4.1) is now *strongly convex* without the full column rank assumption of E_k's made in (f). In the following, we will show that as long as β is large enough, there is a sufficient descent:

$$L(x^{r+1}; y^r) - L(x^r; y^r) \leq -\gamma \|x^{r+1} - x^r\|^2, \quad \text{for some } \gamma > 0 \text{ independet of } y^r. \tag{4.2}$$

This property can be seen by bounding the smooth part of $L(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, \ldots, x_K^r)$, which is given by

$$\bar{L}_k(x_k) := g_k(A_k x_k) + \langle y^r, q - E_k x_k \rangle + \frac{\rho}{2} \left\| \sum_{j<k} E_j x_j^{r+1} + \sum_{j>k} E_j x_j^r + E_k x_k - q \right\|^2,$$

with the Taylor expansion at x_k^r:

$$\bar{L}_k\left(x_k^{r+1}\right) \le \bar{L}_k\left(x_k^r\right) + \langle \nabla \bar{L}_k\left(x_k^r\right), x_k^{r+1} - x_k^r \rangle + \frac{\nu}{2} \left\| x_k^{r+1} - x_k^r \right\|^2 \tag{4.3}$$

where

$$\nu := L\|A_k\| \left\|A_k^T\right\| + \rho \left\|E_k^T E_k\right\|$$

is the Lipschitz constant of $\bar{L}_k(\cdot)$ and L is the Lipschitz constant of $\nabla g_k(\cdot)$. Making the above inequality more explicit yields

$$\begin{aligned}
&L\left(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^{r+1}, x_{k+1}^r, \ldots, x_K^r; y^r\right) - L\left(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, x_{k+1}^r, \ldots, x_K^r; y^r\right) \\
&\quad \le h_k\left(x_k^{r+1}\right) - h_k\left(x_k^r\right) + \langle y^r, E_k\left(x_k^r - x_k^{r+1}\right)\rangle + \left\langle A_k^T \nabla g_k\left(A_k x_k^r\right), x_k^{r+1} - x_k^r \right\rangle \\
&\qquad + \left\langle \rho E_k^T \left(\sum_{j<k} E_j x_j^{r+1} + \sum_{j\ge k} E_j x_j^r - q \right), x_k^{r+1} - x_k^r \right\rangle + \frac{\nu}{2} \left\| x_k^{r+1} - x_k^r \right\|^2 \\
&\quad \le -\frac{\beta}{2} \left\| x_k^{r+1} - x_k^r \right\|^2 + \frac{\nu}{2} \left\| x_k^{r+1} - x_k^r \right\|^2 \\
&\quad = -\gamma \left\| x_k^{r+1} - x_k^r \right\|^2, \quad \forall\, k,
\end{aligned} \tag{4.4}$$

provided the regularization parameter β satisfies

$$\gamma := \frac{1}{2}(\beta - \nu) > 0.$$

In the above derivation of (4.4), the first step is due to (4.3), while the second inequality follows from the definition of x_k^{r+1} (cf. (4.1)). Summing (4.4) over all k yields the desired estimate of sufficient descent (4.2).

To verify that Lemma 2.5 still holds for the linearized proximal ADMM algorithm, we note from the corresponding optimality condition for (4.1)

$$\begin{aligned}
x_k^{r+1} = \operatorname{prox}_{h_k} \Bigg[& x_k^{r+1} - A_k^T \nabla_{x_k} g_k\left(A_k x_k^r\right) + E_k^T y^r \\
& - \rho E_k^T \left(\sum_{j<k} E_j x_j^{r+1} + \sum_{j\ge k} E_j x_j^r - q \right) - \beta\left(x_k^{r+1} - x_k^r\right) \Bigg].
\end{aligned}$$

Using this relation in place of (2.13) and following the same proof steps, we can easily prove that the bound (2.12) in Lemma 2.5 can be extended to the linearized proximal ADMM algorithm. Thus, the convergence results in Theorem 3.1 remain true for the linearized proximal ADMM algorithm (4.1).

It remains to verify that Lemma 3.1 still holds true. In fact the first part of Lemma 3.1 can be shown to be independent of the iterates, thus it trivially holds true for the linearized proximal ADMM algorithm. To show that the second part of Lemma 3.1 is true, note that the optimality condition of the linearized proximal ADMM algorithm implies that

$$\begin{aligned} x_k^{r+1} &= \text{prox}_{h_k}\left[x_k^{r+1} - \nabla_{x_k}\bar{L}\left(\left\{x_{j<k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j\geq k\}}; y^r\right) - \beta\left(x_k^{r+1} - x_k^r\right)\right] \\ &:= \text{prox}_{h_k}\left[x_k^r - \nabla_{x_k}\bar{L}(x^r; y^r) - e_k^r\right] \end{aligned}$$

where in this case e_k^r is given as

$$e_k^r := x_k^r - x_k^{r+1} + \nabla_{x_k}\bar{L}\left(\left\{x_{j<k}^{r+1}\right\}, \left\{x_j^r\right\}_{\{j\geq k\}}; y^r\right) - \nabla_{x_k}\bar{L}(x^r; y^r) + \beta\left(x_k^{r+1} - x_k^r\right).$$

It is then straightforward to show that the norm of e_k^r can be bounded by $c^{'}\|x^r - x^{r+1}\|$ for some constant $c^{'} > 0$. The rest of the proof follows the same steps as in Lemma 3.1.

4.2 Jacobi update

Another popular variant of the ADMM algorithm is to use a Jacobi iteration (instead of a Gauss–Seidel iteration) to update the primal variable blocks $\{x_k\}$. In particular, the ADMM iteration (1.6) is modified as follows:

$$\begin{aligned} x_k^{r+1} = \underset{x_k}{\text{argmin}} \Big(& h_k(x_k) + g_k(A_k x_k) - \langle y^r, E_k x_k\rangle \\ & + \frac{\rho}{2}\left\|E_k x_k \sum_{j\neq k} E_j x_j^r - q\right\|^2\Big), \quad \forall\, k. \end{aligned} \tag{4.5}$$

The convergence for this direct Jacobi scheme is unclear, as the augmented Lagrangian function may not decrease after each Jacobi update. In the following, we consider a modified Jacobi scheme with an explicit stepsize control. Specifically, let us introduce an intermediate variable $w = (w_1^T, \ldots, w_K^T)^T \in \Re^n$. The modified Jacobi update is given as follows:

$$w_k^{r+1} = \underset{x_k}{\text{argmin}} \Big(h_k(x_k) + g_k(A_k x_k) - \langle y^r, E_k x_k\rangle$$

$$+\frac{\rho}{2}\left\|E_k x_k+\sum_{j\neq k}E_j x_j^r-q\right\|^2\Bigg), \quad \forall\, k, \tag{4.6}$$

$$x_k^{r+1}=x_k^r+\frac{1}{K}\left(w_k^{r+1}-x_k^r\right), \quad \forall\, k. \tag{4.7}$$

where a stepsize of $1/K$ is used in the update of each variable block.

With this modification, we claim that Lemmas 2.4–2.5 and Lemma 3.1 still hold. In particular, Lemma 2.4 can be argued as follows. The strong convexity of $L(x;y)$ with respect to the variable block x_k implies that

$$\begin{aligned}&L\left(x_1^r,\ldots,x_{k-1}^r,x_k^r,x_{k+1}^r\ldots,x_K^r;y^r\right)-L\left(x_1^r,\ldots,x_{k-1}^r,w_k^r,x_{k+1}^r\ldots,x_K^r;y^r\right)\\&\quad\geq\gamma\left\|w_k^{r+1}-x_k^r\right\|^2,\ \forall\, k.\end{aligned}$$

Using this inequality we obtain

$$\begin{aligned}&L(x^r;y^r)-L(x^{r+1};y^r)\\&\quad=L(x^r;y^r)-L\left(\frac{K-1}{K}x^r+\frac{1}{K}w^{r+1};y^r\right)\\&\quad=L(x^r;y^r)-L\left(\frac{1}{K}\sum_{k=1}^K\left(x_1^r,\ldots,x_{k-1}^r,w_k^{r+1},x_{k+1}^r\ldots,x_K^r\right);y^r\right)\\&\quad\geq L(x^r;y^r)-\frac{1}{K}\sum_{k=1}^K L\left(x_1^r,\ldots,x_{k-1}^r,w_k^{r+1},x_{k+1}^r\ldots,x_K^r;y^r\right)\\&\quad=\frac{1}{K}\sum_{k=1}^K\left(L(x^r;y^r)-L\left(x_1^r,\ldots,x_{k-1}^r,w_k^{r+1},x_{k+1}^r\ldots,x_K^r;y^r\right)\right)\\&\quad\geq\frac{\gamma}{K}\sum_{k=1}^K\left\|w_k^{r+1}-x_k^r\right\|^2\\&\quad=\frac{\gamma}{K}\|w^{r+1}-x^r\|^2.\end{aligned}$$

where the first inequality comes from the convexity of the augmented Lagrangian function.

From the update rule (4.7) we have $K(x_k^{r+1}-x_k^r)=(w_k^{r+1}-x_k^r)$, which combined with the previous inequality yields

$$L(x^r;y^r)-L(x^{r+1};y^r)\geq\gamma K\|x^{r+1}-x^r\|^2.$$

The proof of Lemma 2.5 also requires only minor modifications. In particular, we have the following optimality condition for (4.5)

$$w_k^{r+1} = \text{prox}_{h_k}\Bigg[w_k^{r+1} - A_k^T \nabla_{x_k} g_k\left(A_k w_k^{r+1}\right) + E_k^T y^r - \rho E_k^T\left(\sum_{j\neq k} E_j x_j^r + E_k w_k^{r+1} - q\right)\Bigg]$$

Similar to the proof of Lemma 2.5, we have

$$\left\| w_k^{r+1} - \text{prox}_{h_k}\left[x_k^r - A_k^T \nabla_{x_k} g_k\left(A_k x_k^r\right) + E_k^T y^r - \rho E_k^T\left(Ex^r - q\right)\right]\right\| \leq c\|w^{r+1} - x^r\|.$$

Utilizing the relationship $K(x_k^{r+1} - x_k^r) = (w_k^{r+1} - x_k^r)$, we can establish Lemma 2.5 by following similar proof steps (which we omit due to space reason).

Lemma 3.1 can be shown as follows. We first express w_k^{r+1} as

$$\begin{aligned} w_k^{r+1} &= \text{prox}_{h_k}\left[w_k^{r+1} - \nabla_{x_k}\bar{L}\left(\left\{x_{j\neq k}^r\right\}, w_k^{r+1}; y^r\right)\right] \\ &= \text{prox}_{h_k}\left[x_k^r - \nabla_{x_k}\bar{L}\left(x^r; y^r\right) - e_k^r\right] \end{aligned}$$

where we have defined

$$e_k^r := \nabla_{x_k}\bar{L}\left(\left\{x_{j\neq k}^r\right\}, w_k^{r+1}; y^r\right) - \nabla_{x_k}\bar{L}\left(x^r; y^r\right) + x_k^r - w_k^{r+1}.$$

Again by using the relationship $K(x_k^{r+1} - x_k^r) = (w_k^{r+1} - x_k^r)$, we can bound the norm of e_k^r by $c'\|x^{r+1} - x^r\|$, for some $c' > 0$. The remaining proof steps are similar to those in Lemma 3.1.

Since Lemmas 2.4–2.5 and 3.1 hold for the Jacobi version of the ADMM algorithm with a stepsize control, we conclude that the convergence results of Theorem 3.1 remain true in this case.

5 Conclusion and discussion

In this paper we have established the convergence and the rate of convergence of the classical ADMM algorithm when the number of variable blocks are more than two and without requiring the objective function to be strongly convex. Our analysis is a departure of the conventional analysis of ADMM algorithm which relies on the descent of a weighted (semi-)norm of $(x^r - x^*, y^r - y^*)$ and a contraction argument, see [4,6,13,14,16–19,24,25,28,30,43]. In our analysis, we require neither the strong convexity

of the objective function nor the row independence assumption of the constrained matrix E. Instead, we use a local error bound to show that when the stepsize of dual update is made sufficiently small, the sum of the primal and the dual optimality gaps decreases after each ADMM iteration, although separately they may individually increase.

A key insight from our analysis is that proper dual stepsize control for the multi-block ADMM algorithm (with $K \geq 3$) is essential for its convergence. This point is further illustrated by the example given below. An interesting issue for further research is to identify good practical stepsize rules for dual update. As (3.17) suggests, the dual stepsize can be determined explicitly using error bound constants. Unfortunately it may be too conservative and is cumbersome to compute unless the objective function is strongly convex. One possible direction is to use an adaptive dual stepsize rule to guarantee the decrease of the sum of the primal and dual optimality gap.

Example 5.1 We show in this example that without proper dual stepsize control, the multiple block ADMM with $K \geq 3$ can diverge. Consider a slight modification of the example proposed in [8], with 3 block variables (x_1, x_2, x_3):

$$\begin{aligned} \min \quad & f(x_1, x_2, x_3) = 0 && (5.1)\\ \text{s.t.} \quad & E_1x_1 + E_2x_2 + E_3x_3 = 0, \\ & x_1 \in [-20,\ 20],\ x_2 \in [-20,\ 20],\ x_3 \in [-20,\ 20] && (5.2) \end{aligned}$$

where E is given by

$$[E_1\ E_2\ E_3] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix}. \tag{5.3}$$

It can be checked that $x_1 = x_2 = x_3 = 0$ is the unique solution.

In Fig. 2, we plot the iterates generated by the ADMM when setting $\rho = \alpha = 1$. Clearly the algorithm diverges, even when each primal variable is confined in the compact set $[-20,\ 20]$. In Fig. 3, we plot the iterates generated by the ADMM when setting $\rho = 1$ and $\alpha = 0.1$. We see that this time the algorithm converges to the globally optimal solution.

Acknowledgements The authors are grateful to Xiangfeng Wang and Dr. Min Tao of Nanjing University for their constructive comments.

6 Appendix

6.1 Proof of dual error bound (2.8)

The augmented Lagrangian dual function can be expressed as

$$d(y) = \min_{x \in X} \langle y, q - Ex \rangle + \frac{\rho}{2}\|q - Ex\|^2 + g(Ax) + h(x). \tag{6.1}$$

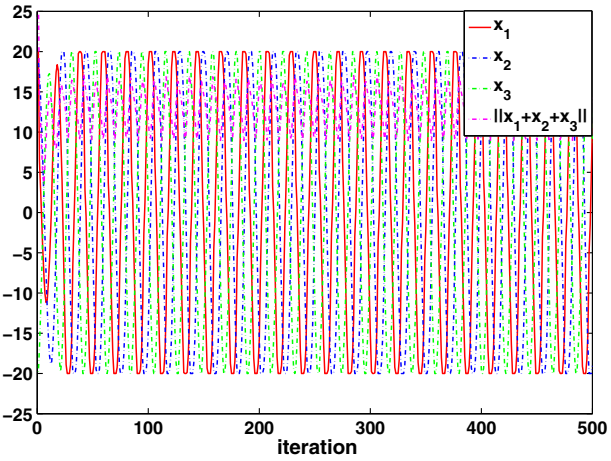

Fig. 2 Divergence of ADMM for Problem (5.2), with $\rho = \alpha = 1$

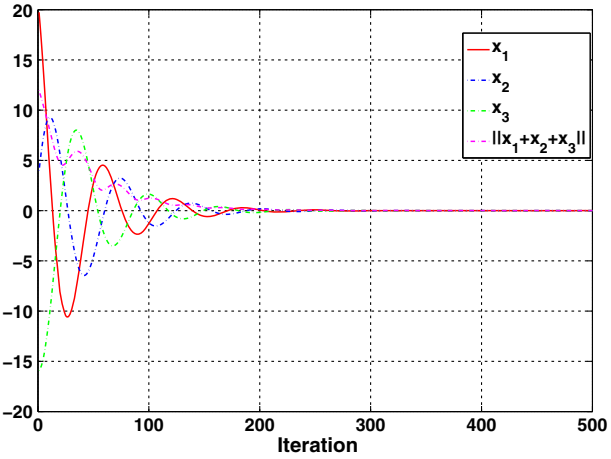

Fig. 3 Convergence of ADMM for Problem (5.2), with $\rho = 1, \alpha = 0.1$

For convenience, define $p(Ex) := \frac{\rho}{2}\|q - Ex\|^2$, and let $\ell(x) := p(Ex) + g(Ax) + h(x)$. For simplicity, in this proof we further restrict ourselves to the case where the nonsmooth part has polyhedral level sets, i.e., $\{x : h(x) \leq \xi\}$ is polyhedral for each ξ. More general cases can be shown along similar lines, but the arguments become more involved.

Let us define

$$x(y) \in \arg\min_{x \in X} \ell(x) + \langle y, q - Ex \rangle.$$

Let (x^*, y^*) denote a primal and dual optimal solution pair. Let X^* and Y^* denote the primal and dual optimal solution set. The following properties will be useful in our subsequent analysis.

(a) There exist positive scalars σ_g, L_g such that $\forall\, x(y), x(y') \in X$
  a-1) $\langle A^T\nabla g(Ax(y')) - A^T\nabla g(Ax(y)), x(y') - x(y)\rangle \geq \sigma_g \|Ax(y') - Ax(y)\|^2$.
  a-2) $g(Ax(y')) - g(Ax(y)) - \langle A^T\nabla g(Ax(y)), x(y') - x(y)\rangle \geq \frac{\sigma_g}{2}\|Ax(y') - Ax(y)\|^2$.
  a-3) $\|A^T\nabla g(Ax(y')) - A^T\nabla g(Ax(y))\| \leq L_g\|Ax(y') - Ax(y)\|$.
(b) All a-1)–a-3) are true for $p(\cdot)$ as well, with some constants σ_p and L_p.
(c) $\nabla d(y) = q - Ex(y)$, and $\|\nabla d(y') - \nabla d(y)\| \leq \frac{1}{\rho}\|y' - y\|$.

Part (a) is true due to the assumed Lipchitz continuity and strong convexity of the function $g(\cdot)$. Part (b) is from the Lipchitz continuity and strong convexity of the quadratic penalization $p(\cdot)$. Part (c) has been shown in Lemmas 2.1 and 2.2.

To proceed, let us rewrite the primal problem equivalently as

$$d(y) = \min_{(x,s):x\in X, h(x)\leq s} \langle y, q - Ex\rangle + p(Ex) + g(Ax) + s. \tag{6.2}$$

Let us write the polyhedral set $\{(x, s) : x \in X, h(x) \leq s\}$ compactly as $C_x x + C_s s \geq c$ for some matrices $C_x \in \mathbb{R}^{j\times n}$, $C_s \in \mathbb{R}^{j\times 1}$ and $c \in \mathbb{R}^{j\times 1}$, where $j \geq 0$ is some integer. For any fixed y, let $(x(y), s(y))$ denote one optimal solution for (6.2), note we must have $h(x(y)) = s(y)$. Due to equivalence, if $y^* \in Y^*$, we must also have $x(y^*) \in X^*$.

Define a set-valued function $\mathcal{M}$ that assigns the vector $(d, e) \in \mathbb{R}^n \times \mathbb{R}^m$ to the set of vectors $(x, s, y, \lambda) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^j$ that satisfy the following system of equations

$$\begin{aligned}
&E^T y + C_x^T\lambda = d,\\
&C_s^T\lambda = 1,\\
&q - Ex = e,\\
&\lambda \geq 0,\ (C_x x + C_s s) \geq c,\ \langle C_x x + C_s s - c, \lambda\rangle = 0.
\end{aligned}$$

It is easy to verify by using the optimality condition for problem (6.2) that

$$\begin{aligned}
(x, s, y, \lambda) \in \mathcal{M}(E^T\nabla p(Ex) + A^T\nabla g(Ax), e) \text{ for some } \lambda\\
\text{if and only if } x = x(y),\ e = \nabla d(y).
\end{aligned} \tag{6.3}$$

We can take $e = 0$, and use the fact that $x(y^*) \in X^*$, we see that $(x, s, y, \lambda) \in \mathcal{M}(E^T\nabla p(Ex) + A^T\nabla g(Ax), 0)$ if and only if $x \in X^*$ and $y \in Y^*$.

The following result states a well-known local upper Lipschitzian continuity property for the polyhedral multifunction $\mathcal{M}$; see [26,33,34].

Proposition 6.1 *There exists a positive scalar θ that depends on A, E, C_x, C_s only, such that for each $(\bar{d}, \bar{e})$ there is a positive scalar δ' satisfying*

$$\mathcal{M}(d, e) \subseteq \mathcal{M}(\bar{d}, \bar{e}) + \theta\|(d, e) - (\bar{d}, \bar{e})\|\mathcal{B}, \tag{6.4}$$

$$\text{whenever } \|(d, e) - (\bar{d}, \bar{e})\| \le \delta'. \tag{6.5}$$

where $\mathcal{B}$ denotes the unit Euclidean ball in $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^j$.

The following is the main result for this appendix. Note that the scalar τ in the claim is independent the choice of y, x, s, and is independent on the coefficients of the linear term s.

Claim 6.1 Suppose all the assumptions in Assumption A are satisfied. Then there exist positive scalars δ, τ such that $\text{dist}(y, Y^*) \le \tau\|\nabla d(y)\|$ for all $y \in \mathcal{U}$ with $\|\nabla d(y)\| \le \delta$.

Proof By the previous claim, $\mathcal{M}$ is locally Lipschitzian with modulus θ at $(\nabla \ell(x^*), 0) = (E^T \nabla p(Ex^*) + A^T \nabla g(Ax^*), 0)$.

Let $\delta \le \delta'/2$. We first show that if $\|\nabla d(y)\| \le \delta$, then we must have $\|\nabla \ell(x(y)) - \nabla \ell(x^*)\| \le \delta'/2$. To this end, take a sequence $y^1, y^2, \ldots$, such that $e^r := \nabla d(y^r) \to 0$. By Assumption (g) $\{x(y^r)\}$ lies in a compact set. Due to the fact that $s(y^r) = h(x(y^r))$, so the sequence $\{s(y^r)\}$ also lies in a compact set (cf. Assumption s(e)). By passing to a subsequence if necessary, let (x^∞, s^∞) be a cluster point of $\{x(y^r), s(y^r)\}$. In light of the continuity of $\nabla \ell(\cdot)$, we have $(\nabla \ell(x(y^r)), e^r) \to (\nabla \ell(x^\infty), 0)$. Now for all r, $\{(x(y^r), s(y^r), \nabla \ell(x(y^r)), e^r)\}$ lies in the set

$$\{(x, s, d, e) \mid (x, s, y, \lambda) \in \mathcal{M}(d, e) \text{ for some } (y, \lambda)\}$$

which is polyhedral and thus is closed. Then we can pass limit to it and conclude (cf. Proposition 6.1)

$$(x^\infty, s^\infty, y^\infty, \lambda^\infty) \in \mathcal{M}(\nabla \ell(x^\infty), 0)$$

for some $(y^\infty, \lambda^\infty) \in \mathbb{R}^m \times \mathbb{R}^j$. Thus by (6.3) and the discussions that follow, we have $x^\infty \in X^*$ and $y^\infty \in Y^*$. By Lemma 2.1, we have $\nabla \ell(x^*) = \nabla \ell(x^\infty)$, which further implies that $\nabla \ell(x(y^r)) \to \nabla \ell(x^*)$. This shows that the desired δ exists.

Then we let $e = \nabla d(y)$, and suppose $\|e\| \le \delta$. From the previous argument we have

$$\|\nabla \ell(x(y)) - \nabla \ell(x^*)\| + \|e\| \le \delta'/2 + \delta'/2 = \delta'.$$

Using the results in Proposition 6.1, we have that there exists $(x^*, s^*, y^*, \lambda^*) \in \mathcal{M}(\nabla \ell(x^*), 0)$ satisfying

$$\|(x(y), s, y, \lambda) - (x^*, s^*, y^*, \lambda^*)\| \le \theta \left(\|\nabla \ell(x^*) - \nabla \ell(x(y))\| + \|e\|\right).$$

Since $(x(y), s, y, \lambda) \in \mathcal{M}(\nabla \ell(x(y)), e)$, it follows from the definition of $\mathcal{M}$ that

$$E^T y + C_x^T \lambda = \nabla \ell(x(y)), \tag{6.6}$$

$$C_s^T \lambda = 1, \tag{6.7}$$

$$q - Ex(y) = e, \tag{6.8}$$
$$\lambda \geq 0,\ (C_x x(y) + C_s s(y)) \geq c,\ \langle C_x x(y) + C_s s(y) - c, \lambda \rangle = 0. \tag{6.9}$$

Since $(x^*, s^*, y^*, \lambda^*) \in \mathcal{M}(\nabla \ell(x^*), 0)$, we have from the definition of $\mathcal{M}$

$$E^T y^* + C_x^T \lambda^* = \nabla \ell(x^*), \tag{6.10}$$
$$C_s^T \lambda^* = 1, \tag{6.11}$$
$$q - Ex^* = 0, \tag{6.12}$$
$$\lambda^* \geq 0,\ (C_x x^* + C_s s^*) \geq c,\ \langle C_x x^* + C_s s^* - c, \lambda^* \rangle = 0. \tag{6.13}$$

Moreover, we have

$$\begin{aligned}
&\sigma_g \|A(x(y) - x^*)\|^2 + \sigma_p \|E(x(y) - x^*)\|^2 \\
&\quad \leq \langle A^T \nabla g(Ax(y)) - A^T \nabla g(Ax(y^*)), x(y) - x(y^*) \rangle \\
&\qquad + \langle E^T \nabla p(Ex(y)) - E^T \nabla p(Ex(y^*)), x(y) - x(y^*) \rangle \\
&\quad = \langle \nabla \ell(x(y)) - \nabla \ell(x(y^*)), x(y) - x(y^*) \rangle \\
&\quad = \langle \lambda - \lambda^*, C_x x(y) - C_x x^* \rangle + \langle y - y^*, Ex(y) - Ex^* \rangle
\end{aligned}$$

where the first inequality comes from the strong convexity of $g(\cdot)$ and $p(\cdot)$; the last equality is from (6.6) and (6.10). Moreover, we have

$$\begin{aligned}
&\langle \lambda - \lambda^*, C_x x(y) - C_x x^* \rangle \\
&\quad = \langle \lambda - \lambda^*, C_x x(y) - C_x x^* \rangle + \langle \lambda - \lambda^*, C_s s - C_s s^* \rangle \\
&\quad = \langle \lambda - \lambda^*, (C_x x(y) + C_s s) - (C_x x^* + C_s s^*) \rangle \\
&\quad = -\langle \lambda^*, C_x x(y) + C_s s - c \rangle - \langle \lambda, C_x x^* + C_s s^* - c \rangle \leq 0
\end{aligned} \tag{6.14}$$

where in the first equality we have used the fact that $C_s^T \lambda - C_s^T \lambda^* = 0$; see (6.7) (6.11); in the third equality and in the last inequality we have used the complementary conditions (6.13) and (6.9). As a result, we have

$$\begin{aligned}
&\sigma_g \|A(x(y) - x^*)\|^2 + \sigma_p \|E(x(y) - x^*)\|^2 \\
&\quad \leq \langle y - y^*, (Ex(y) - q) - (Ex^* - q) \rangle \leq \|y - y^*\| \|e\|,
\end{aligned} \tag{6.15}$$

where the last step is due to $\nabla d(y) = Ex(y) - q$ and $\nabla d(y^*) = Ex^* - q = 0$. Finally we have from Proposition 6.1

$$\begin{aligned}
&\|(x(y), s, y, \lambda) - (x^*, s^*, y^*, \lambda^*)\|^2 \\
&\quad \leq \theta^2 \left(\|\nabla \ell(x^*) - \nabla \ell(x(y))\| + \|e\| \right)^2 \\
&\quad \leq \theta^2 \left(2\|\nabla \ell(x^*) - \nabla \ell(x(y))\|^2 + 2\|e\|^2 \right) \\
&\quad \leq 2\theta^2 \left(2\|\nabla g(x^*) - \nabla g(x(y))\|^2 + 2\|\nabla p(x^*) - \nabla p(x(y))\|^2 + \|e\|^2 \right) \\
&\quad \leq 2\theta^2 \left(L_g^2 \|A^T(x(y) - x^*)\|^2 + L_p^2 \|E^T(x(y) - x^*)\|^2 + \|e\|^2 \right)
\end{aligned}$$

$$\leq 2\theta^2 \max\left(\frac{2L_g^2}{\sigma_g}, \frac{2L_p^2}{\sigma_p}, 1\right)\left(\sigma_g\|A^T(x(y)-x^*)\|^2+\sigma_p\|E^T(x(y)-x^*)\|^2+\|e\|^2\right)$$
$$\leq 2\theta^2 \max\left(\frac{2L_g^2}{\sigma_g}, \frac{2L_p^2}{\sigma_p}, 1\right)\left(\|e\|\|y-y^*\|+\|e\|^2\right)$$
$$\leq 2\theta^2 \max\left(\frac{2L_g^2}{\sigma_g}, \frac{2L_p^2}{\sigma_p}, 1\right)\left(\|e\|\|(x(y),s,y,\lambda)-(x^*,s^*,y^*,\lambda^*)\|+\|e\|^2\right),$$

where the second inequality is due to $\nabla\ell(x) = \nabla g(x) + \nabla p(x)$ and the fourth step follows from properties a-3) and b).

We see that the above inequality is quadratic in $\|(x(y), s, y, \lambda)-(x^*, s^*, y^*, \lambda^*)\|/\|e\|$, so we have

$$\|(x(y), s, y, \lambda) - (x^*, s^*, y^*, \lambda^*)\|/\|e\| \leq \tau$$

for some scalar τ depending on θ, L_g, L_p, σ_g, σ_p. It is worth noting that τ does not depend on the choice of the coefficients of the linear term s. We conclude $\text{dist}(y, Y^*) \leq \tau\|\nabla d(y)\|$. □

References

1. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1999)
2. Bertsekas, D.P., Gafni, E.: Projection methods for variational inequalities with application to the traffic assignment problem. Math. Prog. Study **17**, 139–159 (1982)
3. Bertsekas, D.P., Hosein, P.A., Tseng, P.: Relaxation methods for network flow problems with convex arc costs. SIAM J. Control Optim. **25**, 1219–1243 (1987)
4. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice-Hall, Englewood Cliffs (1989)
5. Boley, D.: Local Linear Convergence of the Alternating Direction Method of Multipliers on Quadratic or Linear Programs. SIAM J. Optim. **23**(4), 2183–2207 (2013)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2011). (Michael Jordan, Editor in Chief)
7. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**, 200–217 (1967)
8. Chen, C., He, B. , Yuan, X., Ye, Y.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Prog. **155**(1), 57–79 (2013)
9. Cottle, R.W., Duvall, S.G., Zikan, K.: A Lagrangian relaxation algorithm for the constrained matrix problem. Nav. Res. Logist. Q. **33**, 55–76 (1986)
10. De Pierro, A.R., Iusem, A.N.: On the convergence properties of Hildreth's quadratic programming algorithm. Math. Prog. **47**, 37–51 (1990)
11. Deng, W., Yin, W.: On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers. J. Sci. Comput. **66**(3), 889–916 (2012)
12. Douglas, J., Rachford, H.H.: On the numerical solution of the heat conduction problem in 2 and 3 space variables. Trans. Am. Math. Soc. **82**, 421–439 (1956)
13. Eckstein, J.: Splitting Methods for Monotone Operators with Applications to Parallel Optimization. Ph.D. Thesis, Operations Research Center, MIT (1989)
14. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)

15. Eckstein, J., Svaiter, B.F.: General projective splitting methods for sums of maximal monotone operators. SIAM J. Control Optim. **48**, 787–811 (2010)
16. Gabay, D.: Application of the method of multipliers to varuational inequalities. In: Fortin, M., Glowinski, R. (eds.) Augmented Lagrangian Methods: Application to the Numerical Solution of Boundary-Value Problem, pp. 299–331. North-Holland, Amsterdam (1983)
17. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
18. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
19. Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics. SIAM Studies in Applied Mathematics, Philadelphia (1989)
20. Goldfarb, D., Ma, S.: Fast multiple splitting algorithms for convex optimization. SIAM J. Optim. 22(2), 533–556 (2012)
21. Goldfarb, D., Ma, S., Scheinberg, K.: Fast alternating linearization methods for minimizing the sum of two convex functions. Math. Prog. A. **141**(1,2), 349–382 (2013)
22. Goldstein, A.A.: Convex programming in hilbert space. Bull. Am. Math. Soc. **70**, 709–710 (1964)
23. Goldstein, T., O'Donoghue, B., Setzer, S.: Fast alternating direction optimization methods. SIAM J. Imaging Sci, **7**(3), 1588–1623 (2014)
24. He, B.S., Tao, M., Yuan, X.M.: Alternating direction method with gaussian back substitution for separable convex programming. SIAM J. Optim. **22**, 313–340 (2012)
25. He, B.S., Yuan, X.M.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
26. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. J. Res. Nat. Bur. Stand. **49**, 263–265 (1952)
27. Iusem, A.N.: On dual convergence and the rate of primal convergence of bregman's convex programming method. SIAM J. Control Optim. **1**, 401–423 (1991)
28. Kontogiorgis, S., Meyer, R.R.: A variable-penalty alternating directions method for convex optimization. Math. Program. **83**, 29–53 (1998)
29. Levitin, E.S., Poljak, B.T.: Constrained minimization methods. Z. Vycisl. Mat. i Mat. Fiz. **6**, 787–823 (1965). English translation in USSR Comput. Math. Phys. **6**, 1–50 (1965)
30. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
31. Lin, Y.Y., Pang, J.-S.: Iterative methods for large convex quadratic programs: a survey. SIAM J. Control Optim. **18**, 383–411 (1987)
32. Luo, Z.-Q., Tseng, P.: On the convergence of the coordinate descent method for convex differentiable minimization. J. Optim. Theory Appl. **72**, 7–35 (1992)
33. Luo, Z.-Q., Tseng, P.: On the Linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**, 408–425 (1992)
34. Luo, Z.-Q., Tseng, P.: On the convergence rate of dual ascent methods for strictly convex minimization. Math. Oper. Res. **18**, 846–867 (1993)
35. Ma, S.: Alternating proximal gradient method for convex minimization. J. Sci. Comput. (2015). doi:10.1007/s10915-015-0150-0
36. Mangasarian, O.L., Shiau, T.-H.: Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. SIAM J. Control Optim. **25**, 583–595 (1987)
37. Monteiro, R., Svaiter, B.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**, 475–507 (2013)
38. Ohuchi, A., Kaji, I.: Lagrangian dual coordinatewise maximization algorithm for network transportation problems with quadratic costs. Networks **14**, 515–530 (1984)
39. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York (1970)
40. Pang, J.-S.: On the Convergence of Dual Ascent Methods for Large Scale Linearly Constrained Optimization Problems. The University of Texas, School of Management, Dallas (1984)
41. Pang, J.-S.: A posteriori error bounds for the linearly-constrained variational inequality problem. Math. Oper. Res. **12**, 474–484 (1987)
42. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
43. Tao, M., Yuan, X.M.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**, 57–81 (2011)

44. Tseng, P.: Dual ascent methods for problems with strictly convex costs and linear constraints: a unified approach. SIAM J. Control Optim. **28**, 214–242 (1990)
45. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Prog. **125**(2), 263–295 (2010)
46. Tseng, P., Bertsekas, D.P.: Relaxation methods for problems with strictly convex separable costs and linear constraints. Math. Prog. **38**, 303–321 (1987)
47. Tseng, P., Bertsekas, D.P.: Relaxation methods for problems with strictly convex costs and linear constraints. Math. Oper. Res. **16**, 462–481 (1991)
48. Ventura, J.A., Hearn, D.W.: Computational Development of a Lagrangian Dual Approach for Quadratic Networks. **21**(4), 469–485 (1991)
49. Wang, X.F., Yuan, X.M.: The linearized alternating direction method of multipliers for dantzig selector. SIAM J. Sci. Comput. **34**, 2792–2811 (2012)
50. Yang, J.F., Zhang, Y.: Alternating direction algorithms for l_1-problems in compressive sensing. SIAM J. Sci. Comput. **33**, 250–278 (2011)
51. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B (Statistical Methodology) **68**, 49–67 (2006)
52. Zhang, H., Jiang, J.J., Luo, Z.-Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. J. Oper. Res. Soc. Chin. 1(2), 163–186 (2013)
53. Zenios, S.A., Mulvey, J.M.: Relaxation techniques for strictly convex network problems. Ann. Oper. Res. **5**, 517–538 (1986)
54. Zhou, Z., Li, X., Wright, J., Candes, E.J., Ma, Y.: Stable principal component pursuit. In: Proceedings of 2010 IEEE International Symposium on Information Theory (2010)