On the Linear Convergence of the Alternating Direction Method of Multipliers*

Zhi-Quan Luo†

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

August 13, 2012

Abstract

We analyze the convergence of the alternating direction method of multipliers (ADMM) for solving the problem of minimizing a nonsmooth convex separable function 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 this paper, we consider using the ADMM to minimize the sum of two or more convex separable functions with a composite structure (subject to linear constraints) and establish its global convergence. Moreover, if the objective function is differentiable, we further show that the ADMM attains a global linear rate of convergence. These results settle a key question regarding the convergence of the ADMM when the number of blocks is more than two. Our proof is based on estimating the distance from a dual feasible solution to the optimal dual solution set by the norm of a certain residual.

KEY WORDS: Linear convergence, alternating directions of multipliers, error bound, relaxation, dual ascent.

AMS(MOS) Subject Classifications: 49, 90.

*The research is supported by the National Science Foundation, grant number DMS-1015346.

†Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: `luozq@umn.edu`

1 Introduction

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

$$\begin{array}{ll} \text{minimize} & f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K) \\ \text{subject to} & Ex = E_1x_1 + E_2x_2 + \cdots + E_Kx_K = q \end{array} \tag{1.1}$$

where each f_k is a nonsmooth convex function (possibly with extended values), $x = (x_1^T, ..., x_K^T)^T \in \Re^n$ is a partition of the optimization variable x, and $E = (E_1, E_2, ..., E_K) \in \Re^{m\times n}$ 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 $\Re_+^m$

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

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:

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

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. Similarly, in the stable version of the robust principal component analysis (PCA) [2], we are given an observation matrix $M \in \Re^{m\times n}$ which is a noise-corrupted sum of a low rank matrix L and a sparse matrix S. The goal is recover L and S by solving the following nonsmooth convex optimization problem

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \rho\|S\|_1 + \lambda\|Z\|_F^2 \\ \text{subject to} & L + S + Z = M \end{array}$$

where $\|\cdot\|_*$ denotes the matrix nuclear norm (defined as the sum of the matrix singular eigenvalues), while $\|\cdot\|_1$ and $\|\cdot\|_F$ denote, respectively, the ℓ_1 and the Frobenius norm of a matrix (equal to the standard ℓ_1 and ℓ_2 vector norms when the matrix is viewed as a vector). In the above formulation, Z denotes the noise matrix, and ρ, λ are some fixed penalty parameters. It is easily seen that the stable robust PCA problem corresponds to the three block case $K = 3$ in the problem (1.1) with $x = (L, S, Z)$ and

$$f_1(L) = \|L\|_*, \quad f_2(S) = \|S\|_1, \quad f_3(Z) = \|Z\|_F^2, \tag{1.2}$$

while the coupling linear constraint is given $L + S + Z = M$. In image processing applications where the low rank matrix L is additionally constrained to be nonnegative, then the above problem can be reformulated as

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \rho\|S\|_1 + \lambda\|Z\|_F^2 + i_{\Re_+^{mn}}(C) \\ \text{subject to} & L + S + Z = M, \; L - C = 0, \end{array}$$

where C is a slack matrix variable of the same size as L, and $i_{\Re_+^{mn}}(\cdot)$ is the indicator function for the nonnegative orthant $\Re_+^{mn}$. In this case, the stable robust PCA problem is again in the form of (1.1). In particular, it has 4 block variables (L, S, Z, C) and the first three convex functions are the same as in (1.2), while the fourth convex function is given by $f_4(C) = i_{\Re_+^{mn}}(C)$. The coupling linear constraints are $L + S + Z = M$, $L - C = 0$. Other applications of the form (1.1) include the latent variable Gaussian graphical model selection problem, see [3].

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

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

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

$$d(y) = \min_x \; f(x) + \langle y, q - Ex\rangle + \frac{\rho}{2}\|q - Ex\|^2 \tag{1.4}$$

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

$$\max_y d(y). \tag{1.5}$$

Moreover, if $\rho > 0$, then Ex is constant over the set of minimizers of (1.4) (see Lemma 2.1 in Section 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 (1.4). Given the differentiability of $d(y)$, it is natural to consider the following dual ascent method to solve the primal problem (1.1)

$$y := y + \alpha\nabla d(y) = y + \alpha(q - Ex(y)), \tag{1.6}$$

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 [9, 12, 13, 20, 25, 27, 34, 35, 39]), gradient ascent (see [20, 27, 37]) and gradient projection [15, 19]. (See [11, 20, 36] 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 (1.3), which makes the exact minimization step in (1.4) no longer decomposable across variable blocks even if f has a separable structure. In this case, it is more economical to minimize (1.4) 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 (1.4), and then update the multiplier y using an approximate optimal solution of (1.4) in a manner similar to (1.6). The resulting algorithm is called the Alternating Direction Method of Multipliers (ADMM) and is summarized as follows (see [40–43]). In the general context of sums of monotone operators, the work of [54] 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 (1.7) below.

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:

$$\begin{cases} x_k^{r+1} = \operatorname*{argmin}_{x_k} L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, ..., x_K^r; y^r), \;\; k = 1, 2, ..., K, \\ y^{r+1} = y^r + \alpha(q - Ex^{r+1}) = y^r + \alpha \left(q - \sum_{k=1}^{K} E_k x_k^{r+1} \right), \end{cases} \tag{1.7}$$

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

Notice that if there is only one block ($K = 1$), then the ADMM reduces to the standard augmented Lagrangian method of multipliers [1] for which the global convergence is well understood. In particular, it is known that, under mild assumptions on the problem, this type of dual gradient ascent methods generate a sequence of iterates whose limit points must be optimal solutions of the original problem (see [12,34,36]). 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 [9]). The rate of convergence of dual ascent methods has been studied in the reference [4] 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 dual optimality the '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 [21,22]; also see [8,14,17] 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 [26].)

When there are two blocks ($K = 2$), the convergence of the ADMM was studied in the context of Douglas-Rachford splitting method [5–7] 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 [44–46] have shown that the objective values generated by the ADMM algorithm and its accelerated version converge at a rate of $O(1/r)$ and $O(1/r^2)$ respectively. 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). [One notable exception to the strong convexity requirement is in the special case of linear programming for which [6] established the linear convergence of ADMM.] More recent convergence rate analysis of the ADMM still requires at least one of the component functions (f_1 or f_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 [50–52]. 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 [53] 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 [55] 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). The key requirement for the global (linear) convergence is the satisfaction of a certain error bound condition that is similar to that used in the analysis of [4]. This error bound estimates the distance from an iterate to the optimal solution set in terms of a certain proximity residual. The class of objective functions that are known to satisfy this error bound condition include many of the compressive sensing applications, such as LASSO [32], Group LASSO [38] or Sparse Group LASSO [23].

In our notation, all vectors are column vectors and $\Re^n$ denotes the n–dimensional Euclidean space. For any vector $x \in \Re^n$, we denote by x_i the ith coordinate of x and, for any $I \subseteq \{1, ..., n\}$, by x_I the vector obtained after removing from x those x_i with $i \notin I$. We also denote by $\|x\|$ the usual Euclidean norm of x, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$ with $\langle x, y\rangle = \sum_i x_i y_i$. For any $h \times k$ matrix A, we denote by A_i the ith row of A, and by A_I the submatrix of A obtained by removing all rows A_i with $i \notin I$. For any function h with gradient ∇h, the notations $\nabla_i h$ and $\nabla_I h$ carry analogous meaning. For any closed convex set X and any vector x in the same space, we denote by $[x]_X^+$ the orthogonal projection of x onto X.

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.

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

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

$$f_k(x_k) = g_k(A_k x_k) + \langle b_k, x_k\rangle + h_k(x_k)$$

where g_k and h_k are both convex and continuous over their domains, and A_k are some given matrices (not necessarily full column rank), while b_k's are given vectors of appropriate dimensions.

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

$$\|\nabla g_k(z_k) - \nabla g_k(z_k')\| \le L\|z_k - z_k'\|, \quad \forall\ z_k,\ z_k' \in \mathrm{int}(\mathrm{dom}\ g_k),$$

where $L > 0$ is a constant. Moreover, $\|\nabla g_k(z_k)\| \to \infty$ whenever z_k approaches the boundary of $\mathrm{int}(\mathrm{dom}\ f)$ or $\|z_k\| \to \infty$.

We remark that the inclusion of a linear term $\langle b_k, x_k\rangle$ in Assumption A(b) is needed to cover cases where b_k lies outside the row span of A_k. In the subsequent analysis, however, we shall assume $b_k = 0$ for all k to simplify the notations. All the ensuing proofs and results will remain true with trivial modifications for $b_k \neq 0$ case.

Under Assumption A, both the primal optimum and the dual optimum values of (1.1) are attained and are equal (i.e., the strong duality holds for (1.1)) so that

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

where d^* is the optimal value of the dual of (1.1).

Roughly speaking, Assumption A requires that the smooth part of f (i.e., the g_k's), in addition to satisfying certain regularity conditions, be a co–finite, strictly convex essentially smooth function or, in the terminology of Rockafellar [29, Sec. 26], a co–finite convex function of the Legendre type. In general, Assumption A(c) is satisfied whenever the smooth part of f is strongly convex twice differentiable or whenever its Hessian is positive definite everywhere on the interior of its effective domain.

Although Assumption A may seem restrictive, there are a number of important special cases that satisfy this assumption. These include (i) strictly convex quadratic programs (see [20]), (ii) certain problems of matrix balancing and image reconstruction, where $f(x)$ is the entropy function $\sum_{j=1}^n x_j \ln x_j$ (see [16, 18, 30]), (iii) a problem of optimal routing on data networks, where $f(x)$ is the inverse barrier function $\sum_{j=1}^n 1/(c_j - x_j)$ with $c_j > 0$ (see [4]), and (iv) the Hazen–Williams' model of flow through pipe networks, where $f(x)$ is the power function $\sum_{j=1}^n a_j (x_j)^c$ with $a_j > 0$ and $c \approx 2.85$ (see [31]).

Under Assumption A, 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

$$d(y) = \min_x L(x; y) = \min_x \; f(x) + \langle y, q - Ex \rangle + \frac{\rho}{2} \|q - Ex\|^2, \tag{2.1}$$

is differentiable everywhere. Let $X(y)$ denote the set of optimal solutions for (2.1).

Lemma 2.1 *For any $y \in \Re^m$, both Ex and $A_k x_k$, $k = 1, 2, ..., K$, 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 (2.1).

Proof. Fix $y \in \Re^m$. We first show that Ex is invariant over $X(y)$. Suppose the contrary, so that there exist two 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

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

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

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

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. **Q.E.D.**

To show the linear convergence of the ADMM method, we need a local error bound around the optimal solution set $X(y)$. To describe this local error bound, 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

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

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:

$$\|\text{prox}_h(x) - \text{prox}_h(x')\| \le \|x - x'\|, \quad \forall\ x, x'. \tag{2.2}$$

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

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

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 (2.1) and under Assumption A, the work of [4,23,33] suggests that the size of the proximal gradient

$$\begin{aligned} \tilde{\nabla}_x L(x; y) &:= x - \mathrm{prox}_h \left(x - \nabla_x (L(x; y) - h(x))\right) \\ &= x - \mathrm{prox}_h \left(x - A^T \nabla g(Ax) + E^T y - \rho E^T (Ex - q)\right) \end{aligned} \tag{2.3}$$

can be used to upper bound the distance to the optimal solution set $X(y)$ of (2.1). 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.

Assumption B. For any $\delta > 0$, there exists a positive scalar τ such that, for any (x, y) satisfying $\|x\| + \|y\| \le \delta$, the following error bounds hold

$$\mathrm{dist}\,(y, Y^*) = \|y - y^*\| \le \tau \|\nabla d(y)\|,$$

and

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

where Y^* denotes the dual optimal solution set and the proximal gradient $\tilde{\nabla}_x L(x; y)$ is given by (2.3). Moreover, the constant τ is *independent* of the choice of y.

The next lemma says if the nonsmooth part of f_k takes a certain form, then Assumption B holds.

Lemma 2.2 *Suppose Assumption A holds. Then the local error bounds in Assumption B hold if the nonsmooth component functions $h_k(x_k)$ of the objective function satisfy one of the following properties:*

(1) Either the epi-graph of $h_k(x_k)$ is a polyhedral set;

(2) Or $h_k(x_k) = \lambda_k \|x_k\|_1 + \sum_J \omega_J \|x_{k,J}\|_2$, where $x_k = (\cdots, x_{k,J}, \cdots)$ is a partition of x_k with J being the partition index.

The proof of Lemma 2.2 is identical to that of [4,23,33] for any fixed y. The only new ingredient in Lemma 2.2 is the additional claim that the constants δ, τ are both independent of the choice of y. This property follows directly from a similar property of Hoffman's error bound (on which the error bounds of [4, 23, 33] are based) for a feasible linear system $P := \{x \mid Ax \le b\}$:

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

where τ is independent of b. In fact, a careful checking of the proof of [4, 23, 33] shows that the corresponding error constants δ and τ for the augmented Lagrangian function $L(x; y)$ can be indeed made *independent* of y. We omit the proof of Lemma 2.2 for space consideration.

By using Lemma 2.1, we show below a Lipschitz continuity property of $\nabla f(x(y))$ and $\nabla d(y)$, for y over any level set of d.

Lemma 2.3 *Fix any scalar $\eta \le f(x^*)$ and let $\mathcal{U} = \{\ y \in \Re^m \mid d(y) \ge \eta\ \}$. Then there holds*

$$\|\nabla d(y') - \nabla d(y)\| \le \frac{1}{\rho}\|y' - y\|, \quad \forall\, y' \in \mathcal{U},\ y \in \mathcal{U}.$$

Proof. Fix any y and y' in $\mathcal{U}$. Let $x = x(y)$ and $x' = x(y')$ be two minimizers of $L(x; y)$ and $L(x; y')$ respectively. By convexity, we have

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

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

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

and

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

Adding the above two equalities yields

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

Upon rearranging terms and using the convexity property

$$\langle \nabla f(x') - \nabla f(x), x' - x\rangle \ge 0,$$

we get

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

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. **Q.E.D.**

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 (1.5) 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 (1.5). We show that this is true in the sense that, for all y such that the above residual is small and $d(y)$ is bounded below, the distance from y to Y^* (dual optimal solution set), defined by

$$\text{dist}\,(y, Y^*) = \min_{y^* \in Y^*} \|y - y^*\|,$$

is bounded above by the norm of this residual. Error bounds like this are similar to that in Assumption B and have been studied previously by Pang [28] and by Mangasarian and Shiau [24], though in different contexts. The above error bound is 'local' in that it holds only for those y that are bounded or near Y^* (as opposed to a 'global' error bound which would hold for all y in $\Re^m$). This local error bound was established in [22] (see Theorem 2.1 therein) which does not require f to be co–finite.

Corollary 2.1 *Let d^* denote the dual optimal value. Then, Assumptions A and B imply that for any scalar $\delta > 0$, there exists a positive scalar τ' such that*

$$d^* - d(y) \leq \tau'\|\nabla d(y)\|^2 = \tau'\|Ex(y) - q\|^2, \tag{2.4}$$

for any $y \in \Re^m$ with $\|y\| \leq \delta$. Moreover, if $h(x) = \sum_{k=1}^K h_k(x_k) = 0$ (i.e., the objective function $f(x)$ of (1.1) is smooth), then for any $x \in \Re^n$, $y \in \Re^m$ with $\|x\| + \|y\| \leq \delta$, there holds

$$L(x; y) - L(x(y); y) \leq \tau'\left(dist\,(x, X(y))\right)^2, \tag{2.5}$$

for any $x(y) \in X(y)$.

Proof. By Assumption A, the strong duality holds so that $f(x^*) = d(y^*)$ where y^* is any dual optimal solution for (1.5). Fix any y, and let y^* be the optimal dual solution closest to y. Then it follows from the mean value theorem that there exists some $\tilde{y}$ in the line segment joining y and y^* such that

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

where the second inequality follows from Lemma 2.3. Recall from Assumption B that there exists some τ such that

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

Combining the above two inequalities yields

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

where $\tau' \geq \tau^2/\rho$ is a constant. This establishes the bound on the size of dual gap (2.4).

It remains to prove the bound on the primal-gap (2.5). To this end, we fix any x and y with $\|x\| + \|y\| \leq \delta$. Let $\bar{x}$ be defined as

$$\bar{x} := \operatorname*{argmin}_{u \in X(y)} \|u - x\|$$

so that $\operatorname{dist}(x, X(y)) = \|x - \bar{x}\|$. Since the nonsmooth part of $f(x)$ is absent, that is,

$$h(x) = \sum_{k=1}^{K} h_k(x_k) = 0,$$

it follows that the augmented Lagrangian function is differentiable. Since $\bar{x}$ is an minimizer of $\min_u L(u; y)$, it follows from the optimality condition that

$$\langle \nabla_x L(\bar{x}; y), x - \bar{x} \rangle = 0.$$

Also, we have from the Mean Value theorem

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

Combining the above two expressions yields

$$\begin{aligned}
L(x;y) \;&-\; L(\bar{x};y) \\
&= \langle \nabla_x L(\tilde{x};y) - \nabla_x L(\bar{x};y), x - \bar{x}\rangle \\
&\leq \|\nabla_x L(\tilde{x};y) - \nabla_x L(\bar{x};y)\|\|x - \bar{x}\| \\
&= \left\| \sum_{k=1}^{K} A_k^T \nabla g_k(A_k \tilde{x}_k) + \rho E^T(E\tilde{x} - q) - \left(\sum_{k=1}^{K} A_k^T \nabla g_k(A_k \bar{x}_k) + \rho E^T(E\bar{x} - q) \right) \right\| \|x - \bar{x}\| \\
&\leq \left(\sum_{k=1}^{K} L\|A_k^T\|\|A_k\| + \rho\|E^T E\| \right) \|\tilde{x} - \bar{x}\|\|x - \bar{x}\| \\
&\leq \left(\sum_{k=1}^{K} L\|A_k^T\|\|A_k\| + \rho\|E^T E\| \right) \|x - \bar{x}\|^2,
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of ∇g_k and the last step is due to the fact that $\tilde{x}$ lies in the line segment joining x and $\bar{x}$ so that $\|\tilde{x} - \bar{x}\| \leq \|x - \bar{x}\|$. Setting $\tau' \geq \left(\sum_{k=1}^{K} L\|A_k^T\|\|A_k\| + \rho\|E^T E\| \right)$ proves (2.5). **Q.E.D.**

Assumption C. Each submatrix E_k has full column rank.

Under Assumption C, the augmented Lagrangian function $L(x;y)$ (cf. (1.3)) is strongly convex with respect to each subvector x_k. As a result, each alternating minimization iteration of ADMM (1.7)

$$x_k^{r+1} = \operatorname*{argmin}_{x_k} L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, ..., x_K^r; y^r), \quad k = 1, ..., K.$$

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 Assumption C holds. Then fix any index r, we have*

$$L(x^r; y^r) - L(x^{r+1}; y^r) \geq \gamma \|x^r - x^{r+1}\|^2, \tag{2.6}$$

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

Proof. By Assumption C, the augmented Lagrangian function

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

is strongly convex per 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,

$$L(x;y) - L(x_1, .., x_{k-1}, \bar{x}_k, x_{k+1}, ..., x_K; y) \geq \rho\lambda_{\min}(E_k^T E_k)\|x_k - \bar{x}_k\|^2, \tag{2.7}$$

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 $k \in \{1, ..., K\}$, by ADMM (1.7), x_{k+1}^{r+1} is the minimizer of $L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, x_{k+2}^r, ..., x_K^r; y^r)$. It follows from (2.7)

$$L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k^r, ..., x_K^r; y^r) - L(x_1^{r+1}, ..., x_k^{r+1}, x_{k+1}^r, ..., x_K^r; y^r) \geq \gamma\|x_k^r - x_k^{r+1}\|^2, \quad \forall\, k, \tag{2.8}$$

where

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

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.

Q.E.D.

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 Assumption A holds. Let $\{x^r\}$ be generated by the ADMM algorithm (1.7). Then there exists some constant $\sigma > 0$ (independent of y^r) such that*

$$\|\tilde{\nabla}L(x^r; y^r)\| \leq \sigma\|x^{r+1} - x^r\| \tag{2.9}$$

for all $r \geq 1$.

Proof. Fix any $r \geq 1$ and any $1 \leq k \leq K$. According to the ADMM procedure (1.7), the variable x_k is updated as follows

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

The corresponding optimality condition can be written as

$$x_k^{r+1} = \operatorname{prox}_{h_k}\left[x_k^{r+1} - A_k^T \nabla_{x_k} g_k(A_k x_k^{r+1}) + E_k^T y^r - \rho E_k^T \left(\sum_{j\leq k} E_j x_j^{r+1} + \sum_{j>k} E_j x_j^r - q \right) \right]. \tag{2.10}$$

Therefore, we have

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

where the first inequality follows from the nonexpansive property of the prox operator (2.2), and the second inequality is due to the Lipschitz property of the gradient vector ∇g_k (cf. Assumption A). Using this relation and the definition of the proximal gradient $\tilde{\nabla} L(x^r; y^r)$, we have

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

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

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

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

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 (1.7). Further we denote

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

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

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

Clearly, we have both $\Delta_d^r \ge 0$ and $\Delta_p^r \ge 0$ for all r. To establish the linear convergence of ADMM, we need several lemmas to estimate the size of primal and dual optimality gap. We first bound the decrease of the dual optimality gap.

Lemma 3.1 *For each $r \geq 1$, there holds*

$$\Delta_d^r - \Delta_d^{r-1} \leq -\alpha(Ex^r - q)^T(E\bar{x}^r - q). \tag{3.3}$$

Proof. 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

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

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}. **Q.E.D.**

Lemma 3.1 shows 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 3.2 *For each $r \geq 1$, we have*

$$\Delta_p^r - \Delta_p^{r-1} \leq \alpha\|Ex^r - q\|^2 - \gamma\|x^{r+1} - x^r\|^2 - \alpha(Ex^r - q)^T(E\bar{x}^r - q) \tag{3.4}$$

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. (1.7)), 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 3.1. **Q.E.D.**

Notice that when $\alpha = 0$ (i.e., no dual update in the ADMM algorithm), Lemma 3.2 reduces to the sufficient decrease estimate (2.6) 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 (3.4). 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 step size α is sufficiently small. This is used to establish the linear convergence of ADMM method.

Theorem 3.1 *Suppose Assumptions A, B, C hold and that the level set of $\Delta_d + \Delta_p$ is bounded, i.e.,*

$$\delta := \left\{\ \|x\| + \|y\| \mid [L(x; y) - d(y)] + [d^* - d(y)] \leq \Delta_p^0 + \Delta_d^0\ \right\} < \infty. \tag{3.5}$$

Then we have the following:

(a) *Every limit point of the sequence of iterates* $\{x^r, y^r\}$ *generated by the ADMM algorithm* (1.7) *is a primal-dual optimal solution pair for* (1.1), *provided the stepsize* α *is sufficiently small.*

(b) *If, additionally, the nonsmooth parts of* $f(x)$ *are zero (i.e.,* $h_k(x_k) = 0$*, for all* k*), then the sequence of iterates* $\{x^r, y^r\}$ *generated by the ADMM algorithm* (1.7) *converge linearly to an optimal primal-dual solution for* (1.1). *Moreover, the sequence of function values* $\{f(x^r)\}$ *also converges linearly.*

Proof. We first prove part (a). 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

$$\bar{x}^r = \operatorname*{argmin}_{\bar{x} \in X(y^r)} \|\bar{x} - x^r\|. \tag{3.6}$$

By induction, suppose $\Delta_d^{r-1} + \Delta_p^{r-1} \leq \Delta_d^0 + \Delta_p^0$ for some $r \geq 1$. Then, $\|x^r\| \leq \delta$ and it follows from Assumption B that

$$\|x^r - \bar{x}^r\| \leq \tau \|\tilde{\nabla} L(x^r; y^r)\| \tag{3.7}$$

for some $\tau > 0$ (independent of y^r). To prove Theorem 3.1, we combine the two estimates (3.3) and (3.4) to obtain

$$\begin{aligned}
[\Delta_p^r + \Delta_d^r] - [\Delta_p^{r-1} + \Delta_d^{r-1}] &= [\Delta_p^r - \Delta_p^{r-1}] + [\Delta_d^r - \Delta_d^{r-1}] \\
&\leq \alpha \|Ex^r - q\|^2 - \gamma \|x^{r+1} - x^r\|^2 - 2\alpha (Ex^r - q)^T (E\bar{x}^r - q) \\
&= \alpha \|Ex^r - E\bar{x}^r\|^2 - \alpha \|E\bar{x}^r - q\|^2 - \gamma \|x^{r+1} - x^r\|^2.
\end{aligned} \tag{3.8}$$

Now we invoke (3.7) and Lemma 2.5 to lower bound $\|x^{r+1} - x^r\|$:

$$\|x^r - \bar{x}^r\| \leq \tau \|\tilde{\nabla} L(x^r; y^r)\| \leq \tau\sigma \|x^{r+1} - x^r\|. \tag{3.9}$$

Substituting this bound into (3.8) yields

$$[\Delta_p^r + \Delta_d^r] - [\Delta_p^{r-1} + \Delta_d^{r-1}] \leq (\alpha \|E\|^2 \tau^2 \sigma^2 - \gamma) \|x^{r+1} - x^r\|^2 - \alpha \|E\bar{x}^r - q\|^2. \tag{3.10}$$

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

$$0 < \alpha < \gamma \tau^{-2} \sigma^{-2} \|E\|^{-2}, \tag{3.11}$$

then the above estimate shows that

$$[\Delta_p^r + \Delta_d^r] \leq [\Delta_p^{r-1} + \Delta_d^{r-1}],$$

which completes the induction. Moreover, the induction argument shows that if the stepsize α satisfies the condition (3.11), then the descent condition (3.10) holds for all $r \geq 1$.

Now we argue that the sequence $\{x^r, y^r\}$ converge to a primal-dual optimal solution pair for the original optimization problem (1.1). By the descent estimate (3.10), we have

$$\|x^{r+1} - x^r\| \to 0, \quad \|E\bar{x}^r - q\| \to 0. \tag{3.12}$$

By the assumption (3.5) and the above induction argument, the sequence $\{x^r, y^r\}$ is bounded. Let $(\bar{x}, \bar{y})$ be any limit point of the sequence $\{x^r, y^r\}$ so that

$$\lim_{r\in\mathcal{R}, r\to\infty} (x^r, y^r) = (\bar{x}, \bar{y})$$

for some subsequence $\mathcal{R}$. By (2.4), we obtain

$$d^* - d(y^r) \leq \tau' \|E\bar{x}^r - q\|^2, \quad r \in \mathcal{R}.$$

By taking limit along this subsequence $\mathcal{R}$, we have

$$0 \leq \lim_{r\in\mathcal{R}, r\to\infty} (d^* - d(y^r)) \leq \tau' \lim_{r\in\mathcal{R}, r\to\infty} \|E\bar{x}^r - q\|^2 = 0.$$

This further implies that

$$\lim_{r\in\mathcal{R}, r\to\infty} (d^* - d(y^r)) = d^* - d(\bar{y}) = 0.$$

Thus, each limit point of the sequence $\{y^r\}$ is a dual optimal solution. Similarly, by taking limit in (3.9) and using (3.12), we obtain

$$\lim_{r\in\mathcal{R}, r\to\infty} \|x^r - \bar{x}^r\| = 0.$$

Thus, $\|\bar{x} - \bar{x}^r\| \leq \|x^r - \bar{x}\| + \|x^r - \bar{x}^r\| \to 0$ as $r \to \infty$ and $r \in \mathcal{R}$. Since $\bar{x}^r \in X(y^r)$, it follows that the limit point of $\{\bar{x}^r\}_{r\in\mathcal{R}}$ must be an element of $X(\bar{y})$. Thus, $\bar{x} \in X(\bar{y})$, implying that $\bar{x}$ is a primal optimal solution for (1.1). This completes the proof of part (a).

Now we establish part (b) under the additional assumption that the nonsmooth part of f is absent (i.e., $\sum_{k=1}^K h_k(x_k) = 0$). We will show that the sum of optimality gaps $\Delta_d^r + \Delta_p^r$ in fact contracts geometrically after each ADMM iteration. To this end, we first notice that the inequalities (3.9) and (3.10) imply

$$[\Delta_p^r + \Delta_d^r] - [\Delta_p^{r-1} + \Delta_d^{r-1}] \leq (\alpha\|E\|^2 - \gamma\tau^{-2}\sigma^{-2})\|x^r - \bar{x}^r\|^2 - \alpha\|E\bar{x}^r - q\|^2. \tag{3.13}$$

Moreover, the condition

$$[L(x^{r+1}; y^r) - d(y^r)] + [d^* - d(y^r)] = \Delta_d^r + \Delta_p^r \leq \Delta_d^0 + \Delta_p^0$$

implies $\|x^{r+1}\| + \|y^r\| \leq \delta$ for all $r \geq 0$. This further implies $\|x^r\| \leq \delta$ and $\|y^r\| \leq \delta$ for all $r \geq 1$. Therefore, it follows from (2.4) of Corollary 2.1 that we have the following cost-to-go estimate

$$\Delta_d^r = d^* - d(y^r) \leq \tau' \|\nabla d(y^r)\|^2 = \tau' \|E\bar{x}^r - q\|^2, \tag{3.14}$$

for some $\tau' > 0$ and for all $r \geq 1$. Moreover, we can use Corollary 2.1 to bound $\|x^r - \bar{x}^r\|^2$ from below by Δ_p^r. In particular, we have from (2.5) that

$$\begin{aligned}\|x^r - \bar{x}^r\|^2 &\geq \frac{1}{\tau'}\left(L(x^r; y^r) - L(\bar{x}^r; y^r)\right)\\ &= \frac{1}{\tau'}\left(\Delta_p^r + L(x^r; y^r) - L(x^{r+1}; y^r)\right)\\ &\geq \frac{1}{\tau'}\left(\Delta_p^r + \gamma\|x^r - x^{r+1}\|^2\right)\\ &\geq \frac{1}{\tau'}\Delta_p^r,\end{aligned}$$

where the second step follows from the definition of Δ_p^r (cf. (3.2)), while the third step is due to Lemma 2.4. Rearranging the terms yields

$$\|x^r - \bar{x}^r\|^2 \geq (\tau')^{-1}\Delta_p^r, \quad \forall\, r.$$

Substituting this bound and (3.14) into (3.13), we obtain

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

Let us choose a sufficiently small $\alpha > 0$ such that (3.11) holds. Then we have

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

Consequently, we have

$$[\Delta_p^r + \Delta_d^r] - [\Delta_p^{r-1} + \Delta_d^{r-1}] \leq -\lambda[\Delta_p^r + \Delta_d^r]$$

which further implies

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

This shows that the sequence $\{\Delta_p^r + \Delta_d^r\}$ is Q-linearly convergent to zero. As a result, we conclude that both Δ_p^r and Δ_d^r converge to zero R-linearly. By the inequality (3.10), we can further conclude that

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

R-linearly. It then follows that the primal sequence $\{x^r\}$ converges R-linearly to a primal optimal solution.

We next show that the dual iterate sequence $\{y^r\}$ is also R-linearly convergent. First, by (3.13), we see that both $\|x^r - \bar{x}^r\| \to 0$ and $\|E\bar{x}^r - q\| \to 0$ R-linearly. This further implies that $Ex^r - q \to 0$ R-linearly. By the ADMM dual update formula (1.7), the linear convergence of $\|Ex^r - q\| \to 0$ implies that $\|y^{r+1} - y^r\| \to 0$ R-linearly. Thus, the dual iterate sequence $\{y^r\}$ is

R-linearly convergent. Furthermore, by part (a), the limit of the iterate sequence $\{x^r, y^r\}$ must be a primal-dual optimal solution pair for (1.1).

Finally, since

$$\begin{aligned} f(x^{r+1}) - d^* &= [f(x^{r+1}) - d(y^r)] + [d(y^r) - d^*] \\ &= [L(x^{r+1}; y^r) - d(y^r)] - [d^* - d(y^r)] - \langle y^r, q - Ex^{r+1}\rangle - \frac{\rho}{2}\|Ex^{r+1} - q\|^2 \\ &= \Delta_p^r - \Delta_d^r - \langle y^r, q - Ex^{r+1}\rangle - \frac{\rho}{2}\|Ex^{r+1} - q\|^2 \end{aligned}$$

and

$$\Delta_p^r \to 0, \quad \Delta_d^r \to 0, \quad \|q - Ex^r\| \to 0$$

linearly, it follows that $f(x^{r+1}) - d^* \to 0$ R-linearly (recall that y^r is bounded by δ). The proof is complete. **Q.E.D.**

An immediate corollary of Theorem 3.1 and Lemma 2.2 is that the ADMM algorithm is globally convergent if the objective function f satisfies any of the two conditions of Lemma 2.2. If f is additionally continuously differentiable, then the ADMM is globally linearly convergent.

4 Extensions to Other Variants of ADMM

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

4.1 Proximal ADMM

In the original ADMM (1.7), 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 $L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, ..., x_K^r)$. More specifically, we update each block of x_k according to the following procedure

$$\begin{aligned} x_k^{r+1} = \operatorname*{argmin}_{x_k} \Big\{ & h_k(x_k) + \langle y^r, q - E_k x_k\rangle + \big\langle A_k^T \nabla g_k(A_k x_k^r), x_k - x_k^r \big\rangle + \frac{\beta}{2}\|x_k - x_k^r\|^2 \\ & + \Big\langle \rho E_k^T \Big(\sum_{j<k} E_j x_j^{r+1} - \sum_{j \geq k} E_j x_j^r - q \Big), x_k - x_k^r \Big\rangle \Big\} \end{aligned} \tag{4.15}$$

in which the smooth part of the objective function in the k-th subproblem, namely,

$$g_k(A_kx_k) + \langle y^r, q - E_kx_k\rangle + \frac{\rho}{2}\Big\|E_kx_k + \sum_{j<k} E_jx_j^{r+1} + \sum_{j>k} E_jx_j^r - q\Big\|^2$$

is linearized locally at x_k^r, and a proximal term $\frac{\beta}{2}\|x_k - x_k^r\|^2$ 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 claim that Theorem 3.1 holds for the proximal ADMM algorithm. Indeed, to establish the (linear) convergence of the proximal ADMM (4.15), we can follow the same proof steps as that for Theorem 3.1, with the only changes in the proof Lemmas 2.4-2.5. To see why Lemma 2.4 holds, we just need to argue that there is a sufficient descent:

$$L(x^{r+1}; y^r) - L(x^r; y^r) \le -\gamma\|x^{r+1} - x^r\|^2, \quad \text{for some } \gamma > 0 \text{ independet of } y^r. \tag{4.16}$$

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

$$s_k(x_k) := g_k(A_kx_k) + \langle y^r, q - E_kx_k\rangle + \frac{\rho}{2}\Big\|\sum_{j<k} E_jx_j^{r+1} + \sum_{j>k} E_jx_j^r + E_kx_k - q\Big\|^2,$$

with the Taylor expansion at x_k^r:

$$s_k(x_k^{r+1}) \le s_k(x_k^r) + \langle \nabla s_k(x_k^r), x_k^{r+1} - x_k^r\rangle + \frac{\nu}{2}\|x_k^{r+1} - x_k^r\|^2$$

where

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

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

$$\begin{aligned}
&L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k^{r+1}, x_{k+1}^r, ..., x_K^r; y^r) - L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k^r, x_{k+1}^r, ..., x_K^r; y^r)\\
\le\ & h_k(x_k^{r+1}) - h_k(x_k^r) + \langle y^r, E_k(x_k^r - x_k^{r+1})\rangle + \left\langle A_k^T\nabla g_k(A_kx_k^r), x_k^{r+1} - x_k^r\right\rangle\\
&+ \left\langle \rho E_k^T\left(\sum_{j<k} E_jx_j^{r+1} + \sum_{j\ge k} E_jx_j^r - q\right), x_k^{r+1} - x_k^r\right\rangle + \frac{\nu}{2}\|x_k^{r+1} - x_k^r\|^2\\
\le\ & -\frac{\beta}{2}\|x_k^{r+1} - x_k^r\|^2 + \frac{1}{2}(L\|A_k\|\|A_k^T\| + \rho\|E_k^TE_k\|)\|x_k^{r+1} - x_k^r\|^2\\
=\ & -\gamma\|x_k^{r+1} - x_k^r\|^2, \quad \forall\, k,
\end{aligned} \tag{4.17}$$

provided the regularization parameter β satisfies

$$\gamma := \frac{1}{2}\left(\frac{\beta}{2} - (L\|A_k\|\|A_k^T\| + \rho\|E_k^T E_k\|)\right) > 0.$$

In the above derivation of (4.17), the first step is due to the fact that the gradient of the smooth part of $L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, ..., x_K^r; y^r)$ is Lipschitz continuous with a modulus of $L\|A_k\|\|A_k^T\| + \rho\|E_k^T E_k\|$, while the second inequality follows from the definition of x_k^{r+1} (cf. (4.15)). Summing (4.17) over all k yields the desired estimate of sufficient descent (4.16).

It remains to verify that Lemma 2.5 still holds for the proximal ADMM algorithm. To this end, we note from the corresponding optimality condition for (4.15)

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

Using this relation in place of (2.10) and following the same proof steps, we can easily prove that the bound (2.9) in Lemma 2.5 can be extended to the proximal ADMM algorithm. Thus, the convergence results in Theorem 3.1 remain true for the proximal ADMM algorithm (4.15).

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, we modify the ADMM iteration (1.7) as follows:

$$x_k^{r+1} = \underset{x_k}{\text{argmin}}\left(h_k(x_k) + g_k(A_k x_k) - \langle y^r, E_k x_k\rangle + \frac{\rho}{2}\left\|E_k x_k + \sum_{j\neq k} E_j x_j^r - q\right\|^2\right), \quad \forall\, k. \quad (4.18)$$

In this case, we claim both Lemma 2.4 and Lemma 2.5 still hold. In particular, the strong convexity of $L(x; y)$ with respect to the variable block x_k shows (2.7) still holds. The rest of the proof is the same as that of Lemma 2.4 except one obvious (and minor) modification in (2.8).

The proof of Lemma 2.5 also requires only minor modifications. In particular, we use the optimality condition for (4.18)

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

to replace (2.10). The remaining proof steps are quite similar to those in the proof of Lemma 2.5 and are omitted. Since both Lemma 2.4 and Lemma 2.5 hold for the ADMM algorithm with Jacobi updates, we conclude that the convergence results of Theorem 3.1 remain true in this case.

5 Concluding Remarks

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 in the absence of strong convexity. Our analysis is a departure of the conventional analysis of ADMM algorithm which relies on the descent of a weighted (semi-)norm of $(x^r - x^*, y^r - y^*)$, see [40–43, 46–49]. 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 the sum of the primal and the dual optimality gaps decreases after each ADMM iteration (even geometrically under smooth conditions), although separately they may individually increase. An interesting issue for further research is whether and how we can relax the smoothness condition in the linear convergence analysis. In particular, it would be interesting to extend the convergence rate analysis to allow nonsmooth components in the objective function $f(x)$ of (1.1). The latter is important in several emerging applications of ADMM (e.g., the group LASSO or matrix completion). Notice that, when the nonsmooth components are present in $f(x)$, our current analysis does establish the global convergence of the ADMM algorithm and allows an arbitrary number of variable blocks. However, the convergence rate analysis requires the objective function to be smooth.

Acknowledgement: The author is grateful to Dr. Min Tao of Nanjing University for her constructive comments.

References

[1] Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (2010).

[2] Zhou, Z., Li, X., Wright, J., Candes, E.J., and Ma, Y.: Stable principal component pursuit. Proceedings of 2010 IEEE International Symposium on Information Theory (2010)

[3] Chandrasekaran, V., Parrilo, P., and Willsky, A.: Latent variable graphical model via convex optimization. Preprint. (2010)

[4] Luo, Z.-Q. and Tseng, P.: On the convergence rate of dual ascent methods for strictly convex minimization. Mathematics of Operations Research 18, 846–867 (1993)

[5] Douglas, J. and Rachford, H.H.: On the numerical solution of the heat conduction problem in 2 and 3 space variables. Transactions of the American Mathematical Society. 82, 421-439 (1956)

[6] Eckstein, J.: Splitting methods for monotone operators with applications to parallel optimization. Ph.D Thesis, Operations Research Center, MIT. (1989)

[7] Eckstein, J. and Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Mathematical Programming.55, 293-318 (1992)

[8] Bertsekas, D. P. and Gafni, E.: Projection methods for variational inequalities with application to the traffic assignment problem. Math. Prog. Study. 17, 139–159 (1982)

[9] Bertsekas, D. P., Hosein, P. A., and Tseng, P.: Relaxation methods for network flow problems with convex arc costs. SIAM J. Control Optim. 25, 1219–1243 (1987)

[10] Bertsekas, D. P. and Gallager, R.: Data Networks. Prentice–Hall, Englewood Cliffs, New Jersey. (1987)

[11] Bertsekas, D. P. and Tsitsiklis, J. N.: Parallel and Distributed Computation: Numerical Methods. Prentice–Hall, Englewood Cliffs, New Jersey (1989)

[12] 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 Comp. Math. and Math. Physics 7, 200–217 (1967)

[13] Cottle, R. W. Duvall, S. G., and Zikan, K.: A Lagrangian relaxation algorithm for the constrained matrix problem. Naval Res. Logistics Quarterly. 33, 55–76 (1986)

[14] De Pierro, A. R. and Iusem, A. N.: On the convergence properties of Hildreth's quadratic programming algorithm. Math. Prog. 47, 37–51 (1990)

[15] Goldstein, A. A.: Convex programming in hilbert space. Bull. Am. Math. Soc. 70, 709–710 (1964)

[16] Herman, G. T.: Image Reconstruction from Projection: The Fundamentals of Computerized Tomography. Academic Press, New York. (1980)

[17] 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)

[18] Lamond, B., and Stewart, N. F.: Bregman's balancing method. Transportation Res. 15B, 239–248 (1981)

[19] Levitin, E. S. and 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)

[20] Lin, Y. Y., and Pang, J.-S.: Iterative methods for large convex quadratic programs: a survey. SIAM J. Control Optim. 18, 383-411 (1987)

[21] Luo, Z.-Q., and Tseng, P.: On the convergence of the coordinate descent method for convex differentiable minimization. J. Optim. Theory and Appl. 72, 7–35 (1992)

[22] Luo, Z.-Q., and Tseng, P.: On the Linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. 30, 408-425 (1992)

[23] Zhang, H., Jiang, J.J. and Luo, Z.-Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. Preprint. (2012)

[24] Mangasarian, O. L., and Shiau, T.-H.: Lipschitz Continuity of solutions of linear inequalities, programs and complementarity problems. SIAM J. Control Optim. 25 583–595 (1987)

[25] Ohuchi, A., and Kaji, I.: Lagrangian dual coordinatewise maximization algorithm for network transportation problems with quadratic costs. Networks 14, 515–530 (1984)

[26] Ortega, J. M., and Rheinboldt, W. C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York, New York. (1970)

[27] 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, Texas. (1984)

[28] Pang, J.-S.: A posteriori error bounds for the linearly–constrained variational inequality problem. Math. Oper. Res. 12, 474–484 (1987)

[29] Rockafellar, R. T.: Convex Analysis, Princeton University Press, Princeton, New Jersey. (1970)

[30] Schneider, M. H., and Zenios, S. A.: A comparative study of algorithms for matrix balancing. Oper. Res. 38, 439–455 (1990)

[31] Stephenson, D.: Pipeflow Analysis, Elsevier, Amsterdam. (1984)

[32] Tibshirani, R.: Regression shrinkage and selection via the Lasso. Jourbal of the Royal Statistical Society Ser. B. 58, 267-288 (1996)

[33] Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Technical report. (2009)

[34] Tseng, P., and Bertsekas D. P.: Relaxation methods for problems with strictly convex separable costs and linear constraints. Math. Prog. 38, 303–321 (1987)

[35] Tseng, P., and Bertsekas D. P.: Relaxation methods for problems with strictly convex costs and linear constraints. Math. Oper. Res. 16, 462–481 (1991)

[36] 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)

[37] Ventura, J. A., and Hearn, D. W.: Computational development of a Lagrangian dual approach for quadratic networks. University of Florida, Industrial and Systems Engineering Department Report 88–14 (1988)

[38] Yuan, M. and Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 68, 49-67 (2006)

[39] Zenios, S. A., and Mulvey, J. M.: Relaxation techniques for strictly convex network problems. Ann. Oper. Res. 5, 517–538 (1986)

[40] 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, North-Holland, Amsterdam, The Netherlands. 299-331 (1983)

[41] Gabay, D. and Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Computer and Mathematics with Applications. 2, 17-40 (1976)

[42] Glowinski, R.: Numerical methods for nonlinear variational problems. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo (1984)

[43] Glowinski, R. and Le Tallec, P.: Augmented Lagrangian and operator splitting methods in nonlinear mechanics. SIAM Studies in Applied Mathematics, Philadelphia, PA. (1989)

[44] Goldfarb, D. and Ma, S.: Fast multiple splitting algorithms for convex optimization. Arxiv preprint arXiv:0912.4570 (2009)

[45] Goldfarb, D., Ma, S. and Scheinberg, K.: Fast alternating linearization methods for minimizing the sum of two convex functions. Arxiv preprint arXiv:0912.4571 (2009)

[46] He, B. S., Tao, M. and Yuan, X.M.: Alternating direction method with gaussian back substitution for separable convex programming. SIAM J. Optim. 22, 313-340 (2012)

[47] He, B. S. and 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)

[48] Kontogiorgis, S. and Meyer, R. R.: A variable-penalty alternating directions method for convex optimization, Mathematical Programming. 83, 29-53 (1998)

[49] Tao, M. and Yuan, X.M.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. 21, 57-81 (2011)

[50] Goldstein, T., O'Donoghue, B. and Setzer, S.: Fast alternating direction optimization methods. CAM report 12-35, UCLA, (2012).

[51] Boley, D.: Linear convergence of ADMM on a model problem. TR 12-009, Department of Computer Science and Engineering, University of Minnesota (2012)

[52] Deng, W. and Yin, W.: On the global linear convergence of alternating direction methods. Preprint. (2012)

[53] Ma, S.: Alternating proximal gradient method for convex minimization. Preprint (2012)

[54] Eckstein, J. and Svaiter, B.F.: General projective splitting methods for sums of maximal monotone operators. SIAM Journal on Control and Optimization. 48, 787–811 (2010)

[55] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning, Michael Jordan, Editor in Chief, 3, 1–122 (2011)