ON THE GLOBAL AND LINEAR CONVERGENCE OF THE GENERALIZED ALTERNATING DIRECTION METHOD OF MULTIPLIERS

WEI DENG* AND WOTAO YIN*

Abstract. The formulation $\min_{x,y} f(x)+g(y)$ subject to $Ax+By=b$, where f and g are extended-value convex functions, arises in many application areas such as signal processing, imaging and image processing, statistics, and machine learning either naturally or after variable splitting. In many common problems, one of the two objective functions is strictly convex and has Lipschitz continuous gradient. On this kind of problem, a very effective approach is the alternating direction method of multipliers (ADM or ADMM), which solves a sequence of f/g-decoupled subproblems. However, its effectiveness has not been matched by a provably fast rate of convergence; only sublinear rates such as $O(1/k)$ and $O(1/k^2)$ were recently established in the literature, though these rates do not require strict convexity. This paper shows that global linear convergence can be guaranteed under the above assumptions on strict convexity and Lipschitz gradient on one of the two functions, along with certain rank assumptions on A and B. The result applies to the generalized ADM that allows the subproblems to be solved faster and less exactly in certain manners. The derived rate of convergence also provides some theoretical guidance for optimizing the ADM parameters. In addition, this paper makes meaningful extensions to the existing global convergence theory of the generalized ADM.

Key words. alternating direction method of multipliers, ADMM, augmented Lagrangian, global convergence, linear convergence, strictly convex, distributed computing

1. Introduction.

The alternating direction method of multipliers (ADM or ADMM) is very effective at solving many practical optimization problems and has wide applications in areas such as signal and image processing, machine learning, statistics, compressive sensing, and operations research. We refer to [1–10] for a few examples of applications. The ADM is applied to constrained convex optimization problems with separable objective functions in the following form

$$\begin{aligned}\min_{x,y}\ & f(x)+g(y)\\ \text{s.t. }\ & Ax+By=b,\end{aligned} \tag{1.1}$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are unknown variables, $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{p\times m}$ are given matrices, and $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ are closed proper convex functions. Some original problems are not in the form of (1.1), but after introducing variables and constraints, they become the form of (1.1). For example, introducing $y = Ax$, problem $\min_x f(x) + g(Ax)$ is transformed to (1.1) with $B = -I$ and $b = \mathbf{0}$. Constraints $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are closed convex sets, can be included as the (extended-value) indicator functions $I_{\mathcal{X}}(x)$ and $I_{\mathcal{Y}}(y)$ in the objective functions f and g. Here the indicator function of a convex set $\mathcal{C}$ is defined by

$$I_{\mathcal{C}}(z) := \begin{cases} 0 & \text{if } z \in \mathcal{C},\\ \infty & \text{if } z \notin \mathcal{C}.\end{cases} \tag{1.2}$$

The main goal of this paper is to show that the ADM applied to (1.1) has global linear convergence provided that f is strictly convex, ∇f is Lipschitz continuous, and matrices A and B have certain rank conditions. The convergence analysis is performed under a general framework that allows the ADM subproblems to be solved inexactly and faster.

*Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005 ({wei.deng, wotao.yin}@rice.edu)

1.1. Background. The classic ADM was first introduced in [11, 12]. Consider the augmented Lagrangian function of (1.1):

$$\mathcal{L}_{\mathcal{A}}(x,y,\lambda)=f(x)+g(y)-\lambda^T(Ax+By-b)+\frac{\beta}{2}\|Ax+By-b\|_2^2, \tag{1.3}$$

where $\lambda\in\mathbb{R}^p$ is the Lagrangian multiplier vector and $\beta>0$ is a penalty parameter. The classic augmented Lagrangian method (ALM) minimizes $\mathcal{L}_{\mathcal{A}}(x,y,\lambda)$ over x and y jointly and then updates λ. However, the ADM replaces the joint minimization by minimization over x and y, one after another, as described in Algorithm 1. Compared to the ALM, though the ADM may take more iterations, it often runs faster due to the easier subproblems.

Algorithm 1: Classic ADM

```
Initialize $x^0,\ \lambda^0,\ \beta>0$;
for $k=0,\ 1,\dots$ do
    $y^{k+1}=\arg\min_y \mathcal{L}_{\mathcal{A}}(x^k,y,\lambda^k)$;
    $x^{k+1}=\arg\min_x \mathcal{L}_{\mathcal{A}}(x,y^{k+1},\lambda^k)$;
    $\lambda^{k+1}=\lambda^k-\beta(Ax^{k+1}+By^{k+1}-b)$.
```

Although there is extensive literature on the ADM and its applications, there are very few results on its rate of convergence until the very recent past. Work [13] shows that for a Jacobi version of the ADM applied to smooth functions with Lipschitz continuous gradients, the objective value descends at the rate of $O(1/k)$ and that of an accelerated version descends at $O(1/k^2)$. Then, work [14] establishes the same rates on a Gauss-Seidel version and requires only one of the two objective functions to be smooth with Lipschitz continuous gradient. Lately, work [15] shows that $\|u^k-u^{k+1}\|$, where $u^k:=(x^k,y^k,\lambda^k)$, of the ADM converges at $O(1/k)$ assuming at least one of the subproblems is exactly solved. Work [16] proves that the dual objective value of a modification to the ADM descends at $O(1/k^2)$ under the assumption that the objective functions are strongly convex (one of them being quadratic) and both subproblems are solved exactly. We show the linear rate of convergence $O(1/c^k)$ for some $c>1$ under a variety of scenarios in which at least one of the two objective functions is strictly convex and has Lipschitz continuous gradient. This rate is stronger than the sublinear rates such as $O(1/k)$ and $O(1/k^2)$ and is given in terms of the solution error, which is stronger than those given in terms of the objective error in [13, 14, 16] and the solution relative change in [15]. On the other hand, [13–15] do not require any strictly convex functions. The fact that a wide range of applications give rise to model (1.1) with at least one strictly convex functions has motivated this work.

During the final polishing of this paper, we learned the work [17] through private communication, which proves the linear convergence of ADM in a different approach. The linear convergence in [17] requires that the objective function is smooth and the step size for updating the multipliers is sufficiently small, while no explicit linear rate is given. On the other hand, it allows more than two blocks of separable variables and it does not require strict convexity; instead, it requires the objective function to include $f(Ex)$, where f is strictly convex and E is a possibly rank-deficient matrix.

It is worth mentioning that the ADM applied to linear programming is known to converge at a global linear rate [18]. For quadratic programming, work [19] presents an analysis leading to a conjecture that the ADM should converge linearly near the optimal solution. Our analysis in this paper is different from those in [18, 19].

Note that when restricted to a compact set, a strictly convex function is strongly convex. So, as long as an algorithm generates a bounded sequence, strict convexity is effectively strong convexity. For simplicity, we use "strong convexity" or "strongly convex" in the remaining of the paper.

Variants of Algorithm 1 that allow $\mathcal{L}_\mathcal{A}$ to be inexactly minimized over x or y are very important to the applications in which it is expensive to exactly solve either the x-subproblem or the y-subproblem, or both of them. For this reason, we present Algorithm 2 below, which is more general than Algorithm 1 by allowing easier subproblems. Our results are established for Algorithm 2.

Algorithm 2: Generalized ADM

```
1 Choose $Q \succeq 0$ and a symmetric matrix $P$. Initialize $x^0,\ \lambda^0,\ \beta > 0, \gamma > 0$;
2 for $k = 0,\ 1, \ldots$ do
3     $y^{k+1} = \arg\min_y \mathcal{L}_\mathcal{A}(x^k, y, \lambda^k) + \frac{1}{2}(y - y^k)^T Q (y - y^k)$;
4     $x^{k+1} = \arg\min_x \mathcal{L}_\mathcal{A}(x, y^{k+1}, \lambda^k) + \frac{1}{2}(x - x^k)^T P (x - x^k)$;
5     $\lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b)$.
```

Compared to Algorithm 1, Algorithm 2 adds $\frac{1}{2}\|y - y^k\|_Q^2$ and $\frac{1}{2}\|x - x^k\|_P^2$ to the y- and x-subproblems, respectively, and assigns γ as the step size for the update of λ. Here, we use the notion $\|x\|_M^2 := x^T M x$. If $M \succ \mathbf{0}$, $\|x\|_M$ is a norm, but we abuse the notation by allowing any *symmetric matrix* M. Different choices of P and Q are overviewed in the next subsection. They can make steps 3 and 4 of Algorithm 2 easier than those of Algorithm 1.

We do not fix $\gamma = 1$ like in most of the ADM literature since γ plays an important role in convergence and speed. For example, when $P = \mathbf{0}$ and $Q = \mathbf{0}$, any $\gamma \in (0, (\sqrt{5}+1)/2)$ guarantees the convergence of Algorithm 2 [20], but $\gamma = 1.618$ makes it converge noticeably faster than $\gamma = 1$. The range of γ depends on P and Q, as well as β. When P is indefinite, γ must be smaller than 1 or the iteration may diverge.

Let us overview two works very related to Algorithm 2. Work [21] considers (1.3) where the quadratic penalty term is generalized to $\|Ax + By - b\|_{H_k}^2$ for a sequence of bounded positive definite matrices $\{H_k\}$, and the work proves the convergence of Algorithm 2 restricted to $\gamma = 1$ and differential functions f and g. Work [22] replaces γ by a general positive definite matrix C and establishes convergence assuming that $A = I$ and the smallest eigenvalue of C is no greater than 1, which corresponds to $\gamma \le 1$ when $C = \gamma I$. In these works, both P and Q are restricted to positive semi-definite matrices, and there is no rate of convergence given.

In addition to deriving linear convergence rates, this paper makes meaningful extensions to the existing convergence theory of Algorithm 2. Specifically, the step size γ is less restrictive, and P is allowed to be indefinite. These extensions translate to faster convergence and more options of solving the x-subproblem efficiently.

1.2. Inexact ADM subproblems.

By "inexact", we mean that the ADM subproblems in Algorithm 1 are replaced by their approximations that are easier to solve. We do not consider the errors in the subproblem solutions due to finite-precision arithmetics or early-stopping of a subproblem solver.

Let us give a few examples of matrix P in step 4 of Algorithm 2. These examples also apply to Q in step 3. Note that P and Q can be different.

Prox-linear [23]. Setting

$$P = \frac{\beta}{\tau} I - \beta A^T A$$

gives rise to a prox-linear problem at step 4 of Algorithm 2:

$$\min_x f(x) + \beta\left((h^k)^T(x - x^k) + \frac{1}{2\tau}\|x - x^k\|_2^2\right), \tag{1.4}$$

where $\tau > 0$ is a proximal parameter and $h^k := A^T(Ax^k + By^{k+1} - b - \lambda^k/\beta)$ is the gradient of the last two terms of (1.3) at $x = x^k$.

This P makes step 4 much easier to compute in various applications. For example, if f is a separable function, problem (1.4) reduces to a set of independent one-dimensional problems. In particular, if f is ℓ_1 norm, the solution is given in the closed form by so-called soft-thresholding. If f is the matrix nuclear norm, then singular-value soft-thresholding is used. If $f(x) = \|\Phi x\|_1$ where Φ is an orthogonal operator or a tight frame, (1.4) also has a closed-form solution. If f is total variation, (1.4) can be solved by graph-cut [24,25]. There are a large number of such examples in signal processing, imaging, statistics, machine learning, etc.

Gradient descent. When function f is *quadratic*, letting

$$P = \frac{1}{\alpha}I - H_f - \beta A^T A, \qquad H_f := \nabla^2 f(x) \succeq 0,$$

gives rise to a gradient descent step for step 4 of Algorithm 2 since the problem becomes

$$\min_x \, (g^k)^T(x - x^k) + \frac{1}{2\alpha}\|x - x^k\|_2^2, \tag{1.5}$$

where $g^k := \nabla f(x^k) + \beta A^T(Ax^k + By^{k+1} - b - \lambda^k/\beta)$ is the gradient of $\mathcal{L}_{\mathcal{A}}(x, y^{k+1}, \lambda^k)$ at $x = x^k$. The solution is

$$x^{k+1} = x^k - \alpha g^k, \tag{1.6}$$

where $\alpha > 0$ is obviously the step size. When step 4 of Algorithm 1 must solve a large, nontrivial linear system, taking the gradient step has a clear advantage.

Approximating A^TA. The term $\frac{\beta}{2}\|Ax+By-b\|_2^2$ in $\mathcal{L}_A(x,y,\lambda)$ contains the quadratic term $\frac{\beta}{2}x^TA^TAx$. Sometimes, replacing A^TA by a certain $D \approx A^TA$ makes step 4 (much) easier to compute. Then one can let

$$P = \beta(D - A^TA).$$

The choice of P effectively turns $\frac{\beta}{2}x^TA^TAx$ into $\frac{\beta}{2}x^TDx$ since

$$\frac{\beta}{2}\|Ax + By - b\|_2^2 + \frac{1}{2}\|x - x^k\|_P^2 = \frac{\beta}{2}x^TDx + [\text{terms linear in } x] + [\text{terms independent of } x].$$

This approach is useful when A^TA is nearly diagonal (D is the diagonal matrix), or is an orthogonal matrix plus error (D is the orthogonal matrix), as well as when an off-the-grid operator A can be approximated by its on-the-grid counterpart that has very fast implementations (e.g., the discrete Fourier transforms and FFT). Note that P can be indefinite.

Goals of P and Q. The general goal is to wisely choose P so that step 4 of Algorithm 2 becomes much easier to carry out and the entire algorithm runs in less time. The same applies to Q of step 3 of Algorithm 2 except we require $Q \succeq 0$ for provable convergence. In the ADM, the two subproblems can be solved in either order (but fixed throughout the iterations). However, when one subproblem is solved less exactly than the other, Algorithm 2 tends to run faster if the *less* exact one is solved *later* — assigned as step 4 of Algorithm 2 — because at each iteration, the ADM updates the variables in the Gauss-Seidel fashion. If the less exact one runs first, its relatively inaccurate solution will then affect the more exact step, making its solution also inaccurate. Since the less exact subproblem should be assigned as the later step 4, more choices of P are needed than Q, which is the case in this paper.

Table 1.1

Four scenarios leading to linear convergence

scenario	strongly convex	Lipschitz continuous	full row rank	remark
1	f	∇f	A	*if* $Q \succ 0$, B has full column rank
2	f, g	∇f	A	
3	f	$\nabla f, \nabla g$	-	B has full column rank
4	f, g	$\nabla f, \nabla g$	-	

Table 1.2

Summary of linear convergence results

case	$P, \hat{P}$	Q	any scenario $1-4$	
			Q-linear convergence	R-linear convergence*
1	$P = 0$	$= 0$	(Ax^k, λ^k)	x^k, y^k, λ^k
2	$\hat{P} \succ 0$	$= 0$	(x^k, λ^k)	
3	$P = 0$	$\succ 0$	(Ax^k, y^k, λ^k)	
4	$\hat{P} \succ 0$	$\succ 0$	(x^k, y^k, λ^k)	

* In cases 1 and 2, scenario 1, R-linear convergence y^k requires full column rank of B; otherwise, only By^k has R-linear convergence

1.3. Summary of results. Table 1.1 summarizes the four scenarios under which we study the linear convergence of Algorithm 2, and Table 1.2 specifies the linear convergent quantities for different types of matrices $\hat{P}$, P, and Q, where

$$\hat{P} := P + \beta A^T A$$

is defined for the convenience of convergence analysis. $P = 0$ and $Q = 0$ correspond to exactly solving the x- and y-subproblems, respectively. Although $P = 0$ and $\hat{P} \succ 0$ are different cases in Table 1.2, they may happen at the same time if A has full column rank; if so, apply the result under $\hat{P} \succ 0$, which is stronger.

The conclusions in Table 1.2 are the quantities that converge either Q-linearly or R-linearly[1]. Q-linear convergent quantities are the *entireties* of multiple variables whereas R-linear convergent quantities are the individual variables x^k, y^k, and λ^k.

Four scenarios. In scenario 1, only function f needs to be strongly convex and having Lipschitz continuous gradient; there is no assumption on g besides convexity. On the other hand, matrix A must have full row rank. Roughly speaking, the full row rank of A makes sure that the error of λ^k can be bounded just from the x-side by applying the Lipschitz continuity of ∇f. One cannot remove this condition or relax it to the full row rank of $[A, B]$ without additional assumptions. Consider the example of $A = [1; 0]$ and $B = [0; 1]$, where $[A, B] = I$ has full rank. Since λ_2^k is not affected by f or $\{x^k\}$ at all, there is no way to take advantages of the Lipschitz continuity of ∇f to bound the error of λ_2^k. In general, without the full row rank of A, a part of λ^k needs to be controlled from the y-side using properties of g.

Scenario 2 adds the strong convexity assumption on g. As a result, the remark in case 1 regarding the full column rank of B is no longer needed.

[1] Suppose a sequence $\{u^k\}$ converges to u^*. We say the convergence is (in some norm $\|\cdot\|$)

- *Q-linear*, if there exists $\mu \in (0, 1)$ such that $\frac{\|u^{k+1} - u^*\|}{\|u^k - u^*\|} \leq \mu$;
- *R-linear*, if there exists a sequence $\{\sigma^k\}$ such that $\|u^k - u^*\| \leq \sigma^k$ and $\sigma^k \to 0$ Q-linearly.

Both scenarios 3 and 4 assume that g is differentiable and ∇g is Lipschitz continuous. As a result, the error of λ^k can be controlled by taking advantages of the Lipschitz continuity of both ∇f and ∇g, and the full row rank assumption on A is no longer needed. On the other hand, scenarios 3 and 4 exclude the problems with non-differentiable g. Compared to scenario 3, scenario 4 adds the strong convexity assumption on g and drops the remark on the full column rank of B.

Under scenario 1 with $Q \succ 0$ and scenario 3, the remarks in Table 1.1 are needed essentially because y^k gets coupled with x^k and λ^k in certain inequalities in our convergence analysis. The full column rank of B helps bound the error of y^k by those of x^k and λ^k.

Four cases. When $P = 0$ (corresponds to exactly solving the ADM x-subproblem), we have $\hat{P} \succeq 0$ and only obtain linear convergence in Ax. However, when $\hat{P} \succ 0$, linear convergence in x is obtained.

When $Q = 0$ (corresponds to exactly solving the ADM y-subproblem), y is not part of the Q-linear convergent joint variable. But, when $Q \succ 0$, y becomes part of it.

1.4. The penalty parameter β. It is well known that the penalty parameter β can significantly affect the speed of the ADM. Since the rate of convergence developed in this paper is a function of β, the rate can be optimized over β. We give some examples in Section 3.2 below, which shows the rate of convergence is positively related to the strong convexity constant of f and g, while being negatively related to the Lipschitz constant of ∇f and ∇g as well as the condition number of A, B and $[A, B]$. More analysis and numerical simulations are left as future research.

1.5. Notation. We let $\langle \cdot, \cdot \rangle$ denote the standard inner product and $\|\cdot\|$ denote the ℓ_2-norm $\|\cdot\|_2$ (the Euclidean norm of a vector or the spectral norm of a matrix). In addition, we use $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ for the smallest and largest eigenvalues of a symmetric matrix M, respectively.

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called *strongly convex* with constant $\nu > 0$ if for all $x_1, x_2 \in \mathbb{R}^n$ and all $t \in [0, 1]$,

$$f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2) - \frac{1}{2}\nu t(1-t)\|x_1 - x_2\|^2. \tag{1.7}$$

The gradient ∇f is called *Lipschitz continuous* with constant L_f if for all $x_1, x_2 \in \mathbb{R}^n$,

$$\|\nabla f(x_1) - \nabla f(x_2)\| \le L_f \|x_1 - x_2\|. \tag{1.8}$$

1.6. Assumptions. Throughout the paper, we make the following standard assumptions.

Assumption 1. *There exists a saddle point $u^* := (x^*, y^*, \lambda^*)$ to problem (1.1), namely, x^*, y^*, and λ^* satisfy the KKT conditions:*

$$B^T\lambda^* \in \partial g(y^*), \tag{1.9}$$

$$A^T\lambda^* \in \partial f(x^*), \tag{1.10}$$

$$Ax^* + By^* - b = 0. \tag{1.11}$$

When assumption 1 fails to hold, the ADM method has either unsolvable or unbounded subproblems or a diverging sequence of λ^k. The optimality conditions of the subproblems of Algorithm 2 are

$$B^T\hat{\lambda} - \beta B^T A(x^k - x^{k+1}) + Q(y^k - y^{k+1}) \in \partial g(y^{k+1}), \tag{1.12}$$

$$A^T\hat{\lambda} + P(x^k - x^{k+1}) \in \partial f(x^{k+1}). \tag{1.13}$$

For notation simplicity, we introduce

$$\hat{\lambda} := \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \tag{1.14}$$

If $\gamma = 1$, then $\hat{\lambda} = \lambda^{k+1}$; otherwise,

$$\hat{\lambda} - \lambda^{k+1} = (\gamma - 1)\beta(Ax^{k+1} + By^{k+1} - b) = (1 - \frac{1}{\lambda})(\lambda^k - \lambda^{k+1}). \tag{1.15}$$

This relation between $\hat{\lambda}$ and λ^{k+1} is used frequently in our analysis.

Assumption 2. *Functions f and g are convex.* We define scalars ν_f and ν_g as the modulus of f and g, respectively. Following from (1.7), they satisfy

$$\langle s_1 - s_2, x_1 - x_2 \rangle \geq \nu_f \|x_1 - x_2\|^2, \quad \forall x_1,\ x_2,\ s_1 \in \partial f(x_1),\ s_2 \in \partial f(x_2), \tag{1.16}$$

$$\langle t_1 - t_2, y_1 - y_2 \rangle \geq \nu_g \|y_1 - y_2\|^2, \quad \forall y_1,\ y_2,\ t_1 \in \partial g(y_1),\ t_2 \in \partial g(y_2). \tag{1.17}$$

From the convexity of f and g, it follows that $\nu_f, \nu_g \geq 0$, which are used throughout Section 2. They are strictly positive if the functions are strongly convex. To show *linear* convergence, Section 3 uses $\nu_f > 0$ and, for scenarios 3 and 4, $\nu_g > 0$ as well. Indeed, we only f and g's properties over the compact sets including $\{x^k\}$ and $\{y^k\}$, not globally; in particular, strict convexity can replace strong convexity.

1.7. Organization. The rest of the paper is organized as follows. Section 2 shows the global convergence of the generalized ADM under mild assumptions. Then Section 3, under the assumptions in Table 1.1, further proves the global linear convergence. Section 4 discusses several interesting applications that are covered by our linear convergence theory. In Section 5, we present some preliminary numerical results to demonstrate the linear convergence behavior of ADM. Finally, Section 6 concludes the paper.

2. Global convergence.

In this section, we show the global convergence of Algorithm 2. The proof steps are similar to the existing ADM convergence theory in [21,22] but are adapted to Algorithm 2. Several inequalities in the section are used in the linear convergence analysis in the next section.

2.1. Convergence analysis.

For notation simplicity, we introduce

$$u^* := \begin{pmatrix} x^* \\ y^* \\ \lambda^* \end{pmatrix}, \quad u^k := \begin{pmatrix} x^k \\ y^k \\ \lambda^k \end{pmatrix}, \quad \hat{u} := \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \hat{\lambda} \end{pmatrix}, \text{ for } k = 0, 1, \ldots,$$

where u^* is a KKT point, u^k is the current point, and $\hat{u}$ is the next point *as if* $\gamma = 1$, and

$$G_0 := \begin{pmatrix} I_n & & \\ & I_m & \\ & & \gamma I_p \end{pmatrix}, \quad G_1 := \begin{pmatrix} \hat{P} & & \\ & Q & \\ & & \frac{1}{\beta} I_p \end{pmatrix}, \quad G := G_0^{-1} G_1 = \begin{pmatrix} \hat{P} & & \\ & Q & \\ & & \frac{1}{\beta\gamma} I_p \end{pmatrix}, \tag{2.1}$$

where we recall $\hat{P} = P + \beta A^T A$. From these definitions it follows

$$u^{k+1} = u^k - G_0(u^k - \hat{u}). \tag{2.2}$$

We choose P, Q and β such that $\hat{P} \succeq 0$ and $Q \succeq 0$. Hence $G \succeq 0$ and $\|\cdot\|_G$ is a (semi-)norm. The analysis is based on bounding the error $\|u^k - u^*\|_G$ and estimate its decrease.

Lemma 2.1. *Under Assumptions 1 and 2, the sequence $\{u^k\}$ of algorithm 2 obeys*

$$\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 \geq h(u^k - \hat{u}) + 2\nu_f \|x^{k+1} - x^*\|^2 + 2\nu_g \|y^{k+1} - y^*\|^2, \tag{2.3}$$

where

$$\begin{aligned} h(u^k - \hat{u}) &= h(x^k - x^{k+1}, y^k - y^{k+1}, \lambda^k - \hat{\lambda}) \\ &= \|x^k - x^{k+1}\|_{\hat{P}}^2 + \|y^k - y^{k+1}\|_Q^2 + \frac{2-\gamma}{\beta}\|\lambda^k - \hat{\lambda}\|^2 + 2(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1}). \end{aligned}$$

Proof. By the convexity of g and the optimality conditions (1.9) and (1.12), it follows that

$$\langle y^{k+1} - y^*,\ B^T \left(\hat{\lambda} - \lambda^* - \beta A(x^k - x^{k+1})\right) + Q(y^k - y^{k+1})\rangle \geq \nu_g \|y^{k+1} - y^*\|^2. \tag{2.4}$$

Similarly, by the convexity of f and the optimality conditions (1.10) and (1.13), we have

$$\langle x^{k+1} - x^*,\ A^T \left(\hat{\lambda} - \lambda^* - \beta A(x^k - x^{k+1})\right) + \hat{P}(x^k - x^{k+1})\rangle \geq \nu_f \|x^{k+1} - x^*\|^2. \tag{2.5}$$

In addition, it follows from (1.11) and (1.14) that

$$A(x^{k+1} - x^*) + B(y^{k+1} - y^*) = \frac{1}{\beta}(\lambda^k - \hat{\lambda}). \tag{2.6}$$

Then, adding (2.4) and (2.5) and using (2.6) give

$$\begin{aligned}&\frac{1}{\beta}\langle \lambda^k - \hat{\lambda},\ \hat{\lambda} - \lambda^* - \beta A(x^k - x^{k+1})\rangle + \langle x^{k+1} - x^*,\ \hat{P}(x^k - x^{k+1})\rangle + \langle y^{k+1} - y^*,\ Q(y^k - y^{k+1})\rangle \\ &\geq \nu_f \|x^{k+1} - x^*\|^2 + \nu_g \|y^{k+1} - y^*\|^2,\end{aligned} \tag{2.7}$$

which can be simplified as

$$(\hat{u} - u^*)^T G_1 (u^k - \hat{u}) \geq \langle A(x^k - x^{k+1}),\ \lambda^k - \hat{\lambda}\rangle + \nu_f \|x^{k+1} - x^*\|^2 + \nu_g \|y^{k+1} - y^*\|^2. \tag{2.8}$$

By rearranging the terms, we have

$$(u^k - u^*)^T G_1 (u^k - \hat{u}) \geq \|u^k - \hat{u}\|_{G_1}^2 + \langle A(x^k - x^{k+1}), \lambda^k - \hat{\lambda}\rangle + \nu_f \|x^{k+1} - x^*\|^2 + \nu_g \|y^{k+1} - y^*\|^2. \tag{2.9}$$

From the identity $\|a - c\|_G^2 - \|b - c\|_G^2 = 2(a - c)^T G(a - b) - \|a - b\|_G^2$ and (2.2), it follows that

$$\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 = 2(u^k - u^*)^T G_1 (u^k - \hat{u}) - \|G_0(u^k - \hat{u})\|_G^2. \tag{2.10}$$

By (2.9), we have

$$\begin{aligned}\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 \geq\ & 2\|u^k - \hat{u}\|_{G_1}^2 - \|u^k - \hat{u}\|_{G_1 G_0}^2 + 2\langle A(x^k - x^{k+1}), \lambda^k - \hat{\lambda}\rangle \\ & + 2\nu_f \|x^{k+1} - x^*\|^2 + 2\nu_g \|y^{k+1} - y^*\|^2,\end{aligned} \tag{2.11}$$

and thus (2.3) follows. □

In the next theorem, we bound $h(u^k - \hat{u})$ from zero by applying the Cauchy-Schwarz inequality to its cross term $2(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1})$. If $P = \mathbf{0}$, a more refined bound is obtained to give γ a wider range of convergence.

Theorem 2.2. *Assume Assumptions 1 and 2. (1) When $P \neq \mathbf{0}$, if γ obeys*

$$(2 - \gamma)P \succ (\gamma - 1)\beta A^T A \tag{2.12}$$

then there exists $\eta > 0$ such that

$$\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 \geq \eta \|u^k - u^{k+1}\|_G^2 + 2\nu_f \|x^{k+1} - x^*\|^2 + 2\nu_g \|y^{k+1} - y^*\|^2. \tag{2.13}$$

(2) When $P = \mathbf{0}$, if

$$\gamma \in (0, \frac{1}{2}(1 + \sqrt{5})), \tag{2.14}$$

then there exist $\eta > 0$ *such that*

$$\begin{aligned}&\left(\|u^k - u^*\|_G^2 + \frac{\beta}{\rho}\|r^k\|^2\right) - \left(\|u^{k+1} - u^*\|_G^2 + \frac{\beta}{\rho}\|r^{k+1}\|^2\right)\\ \geq\ &\eta\|u^k - u^{k+1}\|_G^2 + 2\nu_f\|x^k - x^{k+1}\|^2 + 2\nu_f\|x^{k+1} - x^*\|^2 + 2\nu_g\|y^{k+1} - y^*\|^2,\end{aligned} \tag{2.15}$$

where

$$r^k := Ax^k + By^k - b$$

is the residual at iteration k.

If we set $\gamma = 1$*, then we have*

$$\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 \geq \eta\|u^k - u^{k+1}\|_G^2 + 2\nu_f\|x^k - x^{k+1}\|^2 + 2\nu_f\|x^{k+1} - x^*\|^2 + 2\nu_g\|y^{k+1} - y^*\|^2, \tag{2.16}$$

where $\eta = 1$.

Proof.

(1) By the Cauchy-Schwarz inequality, we have

$$2(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1}) \geq -\frac{1}{\rho}\|A(x^k - x^{k+1})\|^2 - \rho\|\lambda^k - \hat{\lambda}\|^2,\ \forall\rho > 0. \tag{2.17}$$

Substituting (2.17) into (2.3) and using $\frac{1}{\gamma}(\lambda^k - \lambda^{k+1}) = \lambda^k - \hat{\lambda}$, we have

$$\begin{aligned}&\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2\\ \geq\ &\|x^k - x^{k+1}\|^2_{\hat{P} - \frac{1}{\rho}A^TA} + \|y^k - y^{k+1}\|_Q^2 + \left(\frac{2-\gamma}{\beta} - \rho\right)\frac{1}{\gamma^2}\|\lambda^k - \lambda^{k+1}\|^2\\ &+ 2\nu_f\|x^{k+1} - x^*\|^2 + 2\nu_g\|y^{k+1} - y^*\|^2,\ \forall\rho > 0.\end{aligned} \tag{2.18}$$

To show that (2.13) holds for a certain $\eta > 0$, we only need $\hat{P} - \frac{1}{\rho}A^TA \succ 0$ and $\frac{2-\gamma}{\beta} - \rho > 0$ for a certain $\rho > 0$, which is true if and only if we have $\hat{P} \succ \frac{\beta}{2-\gamma}A^TA$ or, equivalently, (2.12).

(2) For $P = \mathbf{0}$, we first derive a lower bound for the cross term $(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1})$. Applying (1.13) at two consecutive iterations with $P = \mathbf{0}$ and in light of the definition of $\hat{\lambda}$, we have

$$\begin{cases} A^T[\lambda^{k-1} - \beta(Ax^k + By^k - b)] \in \partial f(x^k),\\ A^T\hat{\lambda} \in \partial f(x^{k+1}).\end{cases} \tag{2.19}$$

The difference of the two terms on the left in (2.19) is

$$A^T[\lambda^{k-1} - \beta(Ax^k + By^k - b) - \hat{\lambda}] = A^T(\lambda^k - \hat{\lambda}) - (1-\gamma)\beta A^T(Ax^k + By^k - b). \tag{2.20}$$

By (2.19), (2.20) and (1.16), we get

$$\langle A^T(\lambda^k - \hat{\lambda}), x^k - x^{k+1}\rangle - \langle(1-\gamma)\beta A^T(Ax^k + By^k - b), x^k - x^{k+1}\rangle \geq \nu_f\|x^k - x^{k+1}\|^2, \tag{2.21}$$

to which applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}&(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1})\\ \geq&\langle\sqrt{\beta}(Ax^k + By^k - b), (1-\gamma)\sqrt{\beta}A(x^k - x^{k+1})\rangle + \nu_f\|x^k - x^{k+1}\|^2\\ \geq& -\frac{\beta}{2\rho}\|Ax^k + By^k - b\|^2 - \frac{(1-\gamma)^2\beta\rho}{2}\|A(x^k - x^{k+1})\|^2 + \nu_f\|x^k - x^{k+1}\|^2,\quad \forall\rho > 0.\end{aligned} \tag{2.22}$$

Substituting (2.22) into (2.3) and using $\hat{P} = P + \beta A^T A = \beta A^T A$ and the definition of $\hat{\lambda}$, we have

$$\begin{aligned}
&\|u^k - u^*\|_G^2 + \frac{\beta}{\rho}\|Ax^k + By^k - b\|^2 \\
\geq\ &\|u^{k+1} - u^*\|_G^2 + \frac{\beta}{\rho}\|Ax^{k+1} + By^{k+1} - b\|^2 \\
&+ \beta\left(2 - \gamma - \frac{1}{\rho}\right)\|Ax^{k+1} + By^{k+1} - b\|^2 + \beta\left(1 - (1-\gamma)^2\rho\right)\|A(x^k - x^{k+1})\|^2 \\
&+ \|y^k - y^{k+1}\|_Q^2 + 2\nu_f\|x^k - x^{k+1}\|^2 + 2\nu_f\|x^{k+1} - x^*\|^2 + 2\nu_g\|y^{k+1} - y^*\|^2.
\end{aligned} \tag{2.23}$$

To prove such $\eta > 0$ exists for (2.15), we only need the existence of $\rho > 0$ such that $2 - \gamma - \frac{1}{\rho} > 0$ and $1 - (1-\gamma)^2\rho > 0$, which holds if and only if $2 - \gamma > (1-\gamma)^2$ or, equivalently, $\gamma \in (0, \frac{1+\sqrt{5}}{2})$.

In this case of $P = \mathbf{0}$, if we set $\gamma = 1$, (2.22) reduces to $(\lambda^k - \hat{\lambda})^T A(x^k - x^{k+1}) \geq \nu_f \|x^k - x^{k+1}\|^2$, which substituting into (2.3) gives (2.16) with $\eta = 1$.

□

Now the bounds in Theorem 2.2 are used to give the global convergence of Algorithm 2.

Theorem 2.3 (Global Convergence). *Consider the sequence $\{u^k\}$ generated by Algorithm 2. Under Assumptions 1 and 2 and the additional assumption that $\{u^k\}$ is bounded, for any γ satisfying its conditions given in Theorem 2.2, $\{u^k\}$ converges to a KKT point u^* of (1.1) in the G-norm, namely, $\|u^k - u^*\|_G \to 0$.*

Proof. Being bounded, $\{u^k\}$ has a converging subsequence $\{u^{k_j}\}$. Let $\bar{u} = \lim_{j\to\infty} u^{k_j}$. Next, we will show $\bar{u}$ is a KKT point. Let u^* denote an arbitrary KKT point.

Consider $P \neq \mathbf{0}$ first. From (2.13) we conclude that $\|u^k - u^*\|_G^2$ is monotonically nonincreasing and thus converging, and due to $\eta > 0$, $\|u^k - u^{k+1}\|_G^2 \to 0$. In light of (2.1) where $\hat{P} \succeq 0$ and $Q \succeq 0$, we obtain $\lambda^k - \lambda^{k+1} \to 0$ or equivalently,

$$r^k = (Ax^{k+1} + By^{k+1} - b) \to 0, \quad \text{as } k \to \infty. \tag{2.24}$$

Now consider $P = \mathbf{0}$. From (2.15) we conclude that $\|u^k - u^*\|_G^2 + \frac{\beta}{\rho}\|r^k\|^2$ is monotonically nonincreasing and thus converging. Due to $\eta > 0$, $\|u^k - u^{k+1}\|_G^2 \to 0$, so $\lambda^k - \lambda^{k+1} \to 0$ and (2.24) holds as well. Consequently, $\|u^k - u^*\|_G^2$ also converges.

Therefore, by passing limit on (2.24) over the subsequence, we have for $P = 0$ or not:

$$A\bar{x} + B\bar{y} - b = 0. \tag{2.25}$$

Recall the optimality conditions (1.12) and (1.13):

$$\begin{aligned}
B^T\hat{\lambda} - \beta B^T A(x^k - x^{k+1}) + Q(y^k - y^{k+1}) &\in \partial g(y^{k+1}), \\
A^T\hat{\lambda} + P(x^k - x^{k+1}) &\in \partial f(x^{k+1}).
\end{aligned}$$

Since $\|u^k - u^{k+1}\|_G^2 \to 0$, in light of the definition of G (2.1), we have the following:

- when $P = \mathbf{0}$, $A(x^k - x^{k+1}) \to \mathbf{0}$;
- when $P \neq \mathbf{0}$, the condition (2.12) guarantees $\hat{P} \succ \mathbf{0}$ and thus $x^k - x^{k+1} \to 0$;
- since $Q \succeq \mathbf{0}$, we obtain $Q(y^k - y^{k+1}) \to 0$.

In summary, $\beta B^T A(x^k - x^{k+1})$, $Q(y^k - y^{k+1})$, and $P(x^k - x^{k+1})$ are either 0 or converging to 0 in k, no matter $P = \mathbf{0}$ or not.

Now on both sides of (1.12) and (1.13) taking limit over the *subsequence* and applying Theorem 24.4 of [26], we obtain:

$$B^T\bar{\lambda} \in \partial g(\bar{y}), \tag{2.26}$$

$$A^T\bar{\lambda} \in \partial f(\bar{x}). \tag{2.27}$$

Therefore, together with (2.25), $\bar{u}$ satisfies the KKT condition of (1.1).

Since $\bar{u}$ is a KKT point, we can now let $u^* = \bar{u}$. From $u^{k_j} \to \bar{u}$ in j and the convergence of $\|u^k - u^*\|_G^2$ it follows $\|u^k - u^*\|_G^2 \to 0$ in k. □

Remark 1. *By the definition of G, the convergence $\|u^k - u^*\|_G^2 \to 0$ implies the following:*
(a) $\lambda^k \to \lambda^*$, regardless of the choice of P and Q;
(b) when $P \neq \mathbf{0}$, condition (2.12) guarantees $\hat{P} \succ \mathbf{0}$ and thus $x^k \to x^*$; when $P = \mathbf{0}$, $Ax^k \to Ax^*$;
(c) when $Q \succ \mathbf{0}$, $y^k \to y^*$; otherwise, $By^k \to By^*$ following from (2.24) and (2.25).

Remark 2. *Let us discuss the conditions on γ. If $P \succ 0$, the condition (2.12) is always be satisfied for $0 < \gamma \leq 1$. However, in this case, γ can go greater than 1, which often leads to faster convergence in practice. If $P \not\succ 0$, the condition (2.12) requires γ to lie in $(0, \bar{\gamma})$ where $0 < \bar{\gamma} < 1$ depends on β, P, and $A^T A$. A larger β would allow a larger $\bar{\gamma}$.*

In particular, when the x-subproblem is solved using prox-linear ($P = \frac{\beta}{\tau} I - \beta A^T A$), condition (2.12) is guaranteed by

$$\tau \|A\|^2 + \gamma < 2. \tag{2.28}$$

When the x-subproblem is solved by one-step gradient descent ($P = \frac{1}{\alpha} I - H_f - \beta A^T A$, where $H_f = \nabla^2 f$ and f is quadratic), a sufficient condition for (2.12) is

$$\frac{\beta \|A\|^2}{\frac{1}{\alpha} - \|H_f\|} + \gamma < 2. \tag{2.29}$$

Remark 3. *The assumption on the boundedness of the sequence $\{u^k\}$ can be guaranteed by various conditions. Since (2.13) and (2.15) imply that $\|u^k - u^*\|_G^2$ is bounded, $\{u^k\}$ must be bounded if $\hat{P} \succ 0$ and $Q \succ 0$. Furthermore, if $P = \mathbf{0}$ and $Q = \mathbf{0}$, we have the boundedness of $\{(Ax^k, \lambda^k)\}$ (since $\|u^k - u^*\|_G^2$ is bounded) and that of $\{By^k\}$ by (2.6), so in this case, $\{u^k\}$ is bounded if*
(i) matrix A has full column rank whenever $P = \mathbf{0}$; and
(ii) matrix B has full column rank whenever $Q = \mathbf{0}$.
In addition, the boundedness of $\{u^k\}$ is guaranteed if the objective functions are coercive.

3. Global linear convergence.

In this section, we establish the global linear convergence results for Algorithm 2 that are described in Tables 1.1 and 1.2. We take three steps. First, using (2.13) for $P \neq 0$ and (2.15) for $P = 0$, as well as the assumptions in Table 1.1, we show that there exists $\delta > 0$ such that

$$\|u^k - u^*\|_G^2 \geq (1 + \delta) \|u^{k+1} - u^*\|_G^2, \tag{3.1}$$

where $u^* = \lim_{k \to \infty} u^k$ is given by Theorem 2.3. We call (3.1) the Q-linear convergence of $\{u^k\}$ in G-(semi)norm. Next, using (3.1) and the definition of G, we obtain the Q-linear convergent quantities in Table 1.2. Finally, the R-linear convergence in Table 2 is established.

3.1. Linear convergence in G-(semi)norm.

We first assume $\gamma = 1$, which allows us to simplify the proof presentation. At the end of this subsection, we explain why the results for $\gamma = 1$ can be extended to $\gamma \neq 1$ that satisfies the conditions of Theorem 2.2. Note that for $\gamma = 1$, we have (2.16) instead of (2.15). Hence, no matter $P = 0$ or $P \neq 0$, both inequalities (2.13) and (2.16) have the form

$$\|u^k - u^*\|_G^2 - \|u^{k+1} - u^*\|_G^2 \geq C,$$

where C stands for their right-hand sides. To show (3.1), it is sufficient to establish

$$C \geq \delta \|u^{k+1} - u^*\|_G^2. \tag{3.2}$$

The challenge is that $\|u^{k+1}-u^*\|_G^2$ is the sum of $\|x^{k+1}-x^*\|_{\hat{P}}^2$, $\|y^{k+1}-y^*\|_Q^2$, and $\frac{1}{\beta\gamma}\|\lambda^{k+1}-\lambda^*\|^2$, but C does not contain terms like $\|y^{k+1}-y^*\|^2$ and $\|\lambda^{k+1}-\lambda^*\|^2$. Therefore, we shall bound $\|\lambda^{k+1}-\lambda^*\|^2$ and $\|y^{k+1}-y^*\|_Q^2$ from the existing terms in C or using the strong convexity assumptions. This is done in a series of lemmas below.

Lemma 3.1 (For scenario 1, cases 3 and 4, and scenario 3). *Suppose that B has full column rank. For any $\mu_1 > 0$, we have*

$$\|y^{k+1}-y^*\|^2 \le c_1\|x^{k+1}-x^*\|^2 + c_2\|\lambda^k-\lambda^{k+1}\|^2, \tag{3.3}$$

where $c_1 := (1+\frac{1}{\mu_1})\|A\|^2 \cdot \lambda_{\min}^{-1}(B^TB) > 0$ and $c_2 := (1+\mu_1)(\beta\gamma)^{-2}\cdot\lambda_{\min}^{-1}(B^TB) > 0$.

Proof. By (2.6), we have $\|B(y^{k+1}-y^*)\|^2 = \|A(x^{k+1}-x^*) - \frac{1}{\beta\gamma}(\lambda^k-\lambda^{k+1})\|^2$. Then apply the following inequality

$$\|u+v\|^2 \le \left(1+\frac{1}{\mu_1}\right)\|u\|^2 + (1+\mu_1)\|v\|^2,\ \forall \mu_2 > 0, \tag{3.4}$$

to its right hand side. ∎

Lemma 3.2 (For scenarios 1 and 2). *Suppose that ∇f is Lipschitz continuous with constant L_f and A has full row rank. For any $\mu_2 > 1$, we have*

$$\|\hat{\lambda}-\lambda^*\|^2 \le c_3\|x^{k+1}-x^*\|^2 + c_4\|x^k-x^{k+1}\|^2, \tag{3.5}$$

where $c_3 := L_f^2(1-\frac{1}{\mu_2})^{-1}\lambda_{\min}^{-1}(AA^T) > 0$ and $c_4 := \mu_2\|P\|^2\lambda_{\min}^{-1}(AA^T) > 0$.

Proof. By the optimality conditions (1.10) and (1.13) together with the Lipschitz continuity of ∇f, we have

$$\|A^T(\hat{\lambda}-\lambda^*) + P(x^k-x^{k+1})\|^2 = \|\nabla f(x^{k+1}) - \nabla f(x^*)\|^2 \le L_f^2\|x^{k+1}-x^*\|^2. \tag{3.6}$$

Then apply the following basic inequality:

$$\|u+v\|^2 \ge \left(1-\frac{1}{\mu_2}\right)\|u\|^2 + (1-\mu_2)\|v\|^2,\ \forall \mu_2 > 0, \tag{3.7}$$

to the left hand side of (3.6). We require $\mu_2 > 1$ so that $(1-\frac{1}{\mu_2}) > 0$. ∎

Lemma 3.3 (For scenarios 3 and 4). *Suppose ∇f and ∇g are Lipschitz continuous and $[A, B]$ has full row rank. Let $\bar{c} := \lambda_{\min}^{-1}([A,B][A,B]^T) > 0$. For any $\mu_3 > 1$ and $\mu_4 > 0$, we have*

$$\|\hat{\lambda}-\lambda^*\|^2 \le c_5\|x^k-x^{k+1}\|^2 + c_6\|y^k-y^{k+1}\|_Q^2 + c_7\|x^{k+1}-x^*\|^2 + c_8\|y^{k+1}-y^*\|^2, \tag{3.8}$$

where $c_5 = \mu_3(1+\frac{1}{\mu_4})\|[P^T, -\beta A^TB]\|^2\bar{c} > 0$, $c_6 = \mu_3(1+\mu_4)\|Q\|^2\bar{c} \ge 0$, $c_7 = (1-\frac{1}{\mu_3})^{-1}L_f^2\bar{c} > 0$, and $c_8 = (1-\frac{1}{\mu_3})^{-1}L_g^2\bar{c} > 0$.

Proof. Combining the optimality conditions (1.9), (1.10), (1.13), and (1.12) together with the Lipschitz continuity of ∇f and ∇g, we have

$$\begin{aligned}&\left\|\begin{bmatrix}A^T\\B^T\end{bmatrix}(\hat{\lambda}-\lambda^*) + \begin{bmatrix}P\\-\beta B^TA\end{bmatrix}(x^k-x^{k+1}) + \begin{bmatrix}\mathbf{0}\\Q\end{bmatrix}(y^k-y^{k+1})\right\|^2\\ =&\|\nabla f(x^{k+1})-\nabla f(x^*)\|^2 + \|\nabla g(y^{k+1})-\nabla g(y^*)\|^2\\ \le& L_f^2\|x^{k+1}-x^*\|^2 + L_g^2\|y^{k+1}-y^*\|^2.\end{aligned} \tag{3.9}$$

Similarly, we apply the basic inequalities (3.4) and (3.7) to its left hand side. ∎

Remark 4. *Lemma 3.3 makes a* nonessential *assumption that $[A, B]$ has full row rank, since otherwise certain rows of $[A, B]$ can be eliminated without changing the solution, assuming that $Ax + By = b$ is* *consistent. Indeed, if $[A, B]$ does not have full row rank and the initial multiplier λ^0 is in the range space of $[A, B]$ (letting $\lambda^0 = 0$ suffices), then λ^k, $k = 1, 2, \ldots$, always stay in the range space of $[A, B]$, so do $\hat{\lambda}$ and λ^*.*

Suppose rank$([A, B]) = r < p$. Without loss of generality, assuming the first r rows of $[A, B]$ (denoted by $[A_r, B_r]$) are linearly independent, we have

$$[A, B] = \begin{bmatrix} I \\ L \end{bmatrix} [A_r, B_r],$$

where $I \in \mathbb{R}^{r\times r}$ is the identity matrix and $L \in \mathbb{R}^{(p-r)\times r}$. It follows that

$$\lambda^k = \begin{bmatrix} I \\ L \end{bmatrix} \lambda_r^k,\ \hat{\lambda} = \begin{bmatrix} I \\ L \end{bmatrix} \hat{\lambda}_r,\ \lambda^* = \begin{bmatrix} I \\ L \end{bmatrix} \lambda_r^*.$$

and thus

$$\begin{bmatrix} A^T \\ B^T \end{bmatrix} (\hat{\lambda} - \lambda^*) = \begin{bmatrix} A_r^T \\ B_r^T \end{bmatrix} (I + L^T L)(\hat{\lambda}_r - \lambda_r^*).$$

Then we have

$$\|\hat{\lambda} - \lambda^*\|^2 \leq \bar{c}' \cdot \left\| \begin{bmatrix} A^T \\ B^T \end{bmatrix} (\hat{\lambda} - \lambda^*) \right\|^2,$$

where $\bar{c}' = \lambda_{\min}^{-1}(EE^T)\|I + L^T L\| > 0$ since $E := (I + L^T L)[A_r, B_r]$ has full row rank. Following the same proof procedure, we obtain Lemma 3.3 immediately.

With the above lemmas, we now prove the following main theorem of this subsection.

Theorem 3.4 (Q-linear convergence of u^k in G-(semi)norm). *Under the same assumptions of Theorem 2.3 and $\gamma = 1$, for all scenarios in Table 1.1, there exists $\delta > 0$ such that (3.1) holds.*

Proof. Consider the case of $P = 0$ and the corresponding inequality (2.16). In this case $\hat{P} = \beta A^T A \succeq 0$. Let C denote the right-hand side of (2.16).

Scenarios 1 and 2 (recall in both scenarios, f is strongly convex, ∇f is Lipschitz continuous, and A has full row rank). Note that C contains the terms on the right side of (3.5) with strictly positive coefficients. Hence, applying Lemma 3.2 to C, we can obtain

$$C \geq (c_9\|x^{k+1} - x^*\|^2 + c_{10}\|y^{k+1} - y^*\|^2 + c_{11}\|\lambda^{k+1} - \lambda^*\|^2) + (c_{12}\|y^k - y^{k+1}\|_Q^2 + c_{13}\|\lambda^k - \lambda^{k+1}\|^2) \quad (3.10)$$

with $c_9, c_{11} > 0$, $c_{10} = 2\nu_g \geq 0$, $c_{12} = \eta > 0$, and $c_{13} = \eta/(\beta\gamma) > 0$. We have $c_9 > 0$ because only a fraction of $2\nu_f\|x^{k+1} - x^*\|^2$ is used with Lemma 3.2; $c_9\|x^{k+1} - x^*\|^2$ is unused so it stays. The same principle is applied below to get strictly positive coefficients, and we do not re-state it. For proof brevity, we do not necessarily specify the values of c_i.

For scenario 1 with $Q = 0$, $\|u^{k+1} - u^*\|_G^2 = \|x^{k+1} - x^*\|_{\hat{P}}^2 + \frac{1}{\beta\gamma}\|\lambda^{k+1} - \lambda^*\|^2$. Since $\|x^{k+1} - x^*\|^2 \geq \lambda_{\max}(\hat{P})^{-1}\|x^{k+1} - x^*\|_{\hat{P}}^2$, (3.2) follows from (3.10) with $\delta = \min\{c_9\lambda_{\max}^{-1}(\hat{P}), c_{11}\beta\gamma\} > 0$.

For scenario 1 with $Q \succ 0$, $\|u^{k+1} - u^*\|_G^2 = \|x^{k+1} - x^*\|_{\hat{P}}^2 + \|y^{k+1} - x^*\|_Q^2 + \frac{1}{\beta\gamma}\|\lambda^{k+1} - \lambda^*\|^2$. Since c_{10} is not necessarily strictly positive, we shall apply Lemma 3.1 to (3.10) and obtain

$$C \geq (c_{14}\|x^{k+1} - x^*\|^2 + c_{15}\|y^{k+1} - y^*\|^2 + c_{11}\|\lambda^{k+1} - \lambda^*\|^2) + c_{12}\|y^k - y^{k+1}\|_Q^2 \quad (3.11)$$

where $c_{14}, c_{15}, c_{11}, c_{12} > 0$. So, it leads to (3.1) with $\delta = \min\{c_{14}\lambda_{\max}^{-1}(\hat{P}), c_{15}\lambda_{\max}^{-1}(Q), c_{11}\beta\gamma\} > 0$.

Scenario 2 (recall it is scenario 1 plus that g is strongly convex). We have $c_{10} = 2\nu_g > 0$ in (3.10), which gives (3.1) with $\delta = \min\{c_9 \lambda_{\max}^{-1}(\hat{P}), c_{10}\lambda_{\max}^{-1}(Q), c_{11}\beta\gamma\} > 0$. Note that we have used the convention that if $Q = 0$, then $\lambda_{\max}^{-1}(Q) = \infty$.

Scenario 3 (recall f is strongly convex, both ∇f and ∇g are Lipschitz continuous). We apply Lemma 3.1 to get $\|y^{k+1} - y^*\|^2$ with which we then apply Lemma 3.3 (or Remark 4) to obtain

$$C \geq c_{16}\|x^{k+1} - x^*\|^2 + c_{17}\|y^{k+1} - y^*\|^2 + c_{18}\|\lambda^{k+1} - \lambda^*\|^2, \tag{3.12}$$

where $c_{16}, c_{17}, c_{18} > 0$ and the terms $\|x^k - x^{k+1}\|^2$, $\|y^k - y^{k+1}\|^2$, and $\|\lambda^k - \lambda^{k+1}\|^2$ with nonnegative coefficients have been dropped from the right-hand side of (3.12). From (3.12), we obtain (3.1) with $\delta = \min\{c_{16}\lambda_{\max}^{-1}(\hat{P}), c_{17}\lambda_{\max}^{-1}(Q), c_{18}\beta\gamma\} > 0$.

Scenario 4 (recall it is scenario 3 plus that g is strongly convex). Since $c_{11} = 2\nu_g > 0$ in (3.10), we can directly apply Lemma 3.3 to get (3.1) with $\delta > 0$ in a way similar to scenario 3.

Now consider the case of $P \neq 0$ and the corresponding inequality (2.11). Inequalities (2.11) and (2.16) are similar except (2.16) has the extra term $\|x^k - x^{k+1}\|^2$ with a strictly positive coefficient in its right-hand side. This term is needed when Lemma 3.2 is applied. However, the assumptions of the theorem ensure $\hat{P} \succ 0$ whenever $P \neq 0$. Therefore, in (2.11), the term $\|u^k - u^{k+1}\|_G^2$, which contains $\|x^k - x^{k+1}\|_{\hat{P}}^2$, can spare out a term $c_{19}\|x^k - x^{k+1}\|^2$ with $c_{19} > 0$. Therefore, following the same arguments for the case of $P = 0$, we get (3.1) with certain $\delta > 0$. □

Now we extend the result in Theorem 3.4 (for $\gamma = 1$) to $\gamma \neq 1$ in the following theorem.

THEOREM 3.5. *Under the same assumptions of Theorem 2.3 and $\gamma \neq 1$, for all scenarios in Table 1.1,*

1. *if $P \neq 0$, there exists $\delta > 0$ such that (3.1) holds;*
2. *if $P = 0$, there exists $\delta > 0$ such that*

$$\|u^k - u^*\|_G^2 + \frac{\beta}{\rho}\|r^k\|^2 \geq (1+\delta)\left(\|u^{k+1} - u^*\|_G^2 + \frac{\beta}{\rho}\|r^{k+1}\|^2\right). \tag{3.13}$$

Proof. When $\gamma \neq 1$, which causes $\lambda^{k+1} \neq \hat{\lambda}$. We shall bound $\|\lambda^{k+1} - \lambda^*\|^2$ but Lemmas 3.2 and 3.3 only give bounds on $\|\hat{\lambda} - \lambda^*\|^2$. Noticing that $(\hat{\lambda} - \lambda^*) - (\lambda^{k+1} - \lambda^*) = \hat{\lambda} - \lambda^{k+1} = (\gamma - 1)r^{k+1}$ and C contains a strictly positive term in $\|\lambda^k - \lambda^{k+1}\|^2 = \gamma^2\|r^{k+1}\|^2$, we can bound $\|\lambda^{k+1} - \lambda^*\|^2$ by a positively weighted sum of $\|\hat{\lambda} - \lambda^*\|^2$ and $\|\lambda^k - \lambda^{k+1}\|^2$.

If $P \neq 0$, the rest of the proof follows from that of Theorem 3.4.

If $P = 0$, $\gamma \neq 1$ leads to (2.15), which extends $\|u^i - u^*\|_G^2$ in (2.16) to $\|u^i - u^*\|_G^2 + \frac{\beta}{\rho}\|r^i\|^2$, for $i = k, k+1$. Since C contains $\|\lambda^k - \lambda^{k+1}\|^2 = \gamma^2\|r^{k+1}\|^2$ with a strictly positive coefficient, one obtains (3.13) by using this term and following the proof of Theorem 3.4. □

3.2. Explicit formulas of the linear rate.

To keep the proof of Theorem 3.4 easy to follow, we have avoided giving the explicit formulas of c_i's and thus also those of δ. To give the reader an idea what quantities affect δ, we discuss the value of δ for case 1 of different scenarios with $\gamma = 1$.

Scenario 1 and 2:

$$\delta = 2\nu_f \bigg/ \left(\beta\|A\|^2 + \frac{L_f^2}{\beta\lambda_{\min}(AA^T)}\right). \tag{3.14}$$

Since it is better to have larger δ, we can choose $\beta = \frac{L_f}{\|A\|\sqrt{\lambda_{\min}(AA^T)}}$ and obtain

$$\delta_{\max} = \frac{1}{\kappa_A \kappa_f}, \tag{3.15}$$

where $\kappa_A := \sqrt{\lambda_{\max}(AA^T)/\lambda_{\min}(AA^T)}$ is the condition number of matrix A, and $\kappa_f = L_f/\nu_f$ is the condition number of function f. Not surprisingly, the convergence rate is negatively affected by the condition numbers of A and f.

Scenario 3:

$$\delta = \min\left\{\frac{2\beta\nu_f}{\beta^2\|A\|^2 + c_7 + c_1c_8}, \frac{\beta^2\lambda_{\min}(A^TA) + 2\beta\nu_f}{c_5}, \frac{1}{c_2c_8}\right\}. \tag{3.16}$$

The formulas of c_i's are given in the previous subsection, which involve some arbitrary constants $\mu_1 > 0$ and $\mu_3 > 1$ ($\mu_4 = \infty$ here since $Q = 0$). Therefore, we can maximize δ over $\mu_1 > 0$ and $\mu_3 > 1$ as well as the parameter $\beta > 0$ to get better rate.

Scenario 4:

$$\delta = \min\left\{\frac{2\beta\nu_f}{\beta^2\|A\|^2 + c_7}, \frac{\beta^2\lambda_{\min}(A^TA) + 2\beta\nu_f}{c_5}, \frac{2\beta\nu_g}{c_8}\right\}. \tag{3.17}$$

Similarly, the value of δ can be optimized over $\mu_1 > 0$ and $\beta > 0$.

The formulas of δ for scenarios 3 and 4 appear to be more complicated than the nice formula (3.15) for scenarios 1 and 2. However, a close look at these formulas reveals that the convergence rate is negatively affected by the condition numbers of the constraint matrices A, B and $[A, B]$, as well as the condition numbers of the objective functions f and g.

Due to page limit, we leave other cases and further analysis to future research.

3.3. Q-linear convergent quantities. From the definition of G, which depends on P and Q, it is easy to see that the Q-linear convergence of $u^k = (x^k; y^k; \lambda^k)$ translates to the Q-linear convergence results in Table 1.2. For example, in case 1 ($P = 0$ and $Q = 0$), $\|u^{k+1} - u^*\|_G^2 = \|x^{k+1} - x^*\|_{\hat{P}}^2 + \frac{1}{\beta\gamma}\|\lambda^{k+1} - \lambda^*\|^2$, where $\hat{P} = P + \beta AA^T = \beta A^TA$. Hence, (Ax^k, λ^k) converges Q-linearly. Examining $\|u^{k+1} - u^*\|_G^2$ gives the results for cases 2, 3, 4.

3.4. R-linear convergent quantities. By the definition of R-linear convergence, any part of a Q-linear convergent quantity converges R-linearly. For example, in case 1 ($P = 0$ and $Q = 0$), the Q-linear convergence of (Ax^k, λ^k) in Table 1.2 gives the R-linear convergence of Ax^k and λ^k. Therefore, to establish Table 1.2, it remains to show the R-linear convergence of x^k in cases 1 and 3 and that of y^k in cases 1 and 2. Our approach is to bound their errors by existing R-linear convergent quantities.

Theorem 3.6 (R-linear convergence). *The following statements hold.*

1. *In cases 1 and 3, if λ^k converges R-linearly, then x^k converges R-linearly.*
2. *In cases 1 and 2, scenario 1, if λ^k and x^k both converge R-linearly, then By^k converges R-linearly. In addition, if B has full column rank, then y^k converges R-linearly.*
3. *In cases 1 and 2, scenarios 2–4, if λ^k and x^k both converge R-linearly, then y^k converges R-linearly.*

Proof. We only show the result for $\gamma = 1$ (thus $\hat{\lambda} = \lambda^{k+1}$); for $\gamma \neq 1$ (thus $\hat{\lambda} \neq \lambda^{k+1}$), the results follow from those for $\gamma = 1$ and the R-linear convergence of $\|\hat{\lambda} - \lambda^{k+1}\|^2$, which itself follows from (1.15) and the R-linear convergence of λ^k (thus that of $\lambda^k - \lambda^{k+1}$).

1. By (2.5) and $\hat{P} = \beta A^TA$, we have $\nu_f\|x^{k+1} - x^*\|^2 \leq \|A\|\|x^{k+1} - x^*\|\|\lambda^{k+1} - \lambda^*\|$, which implies

$$\|x^{k+1} - x^*\|^2 \leq \frac{\|A\|^2}{\nu_f^2}\|\lambda^{k+1} - \lambda^*\|^2. \tag{3.18}$$

2. The result follows from (2.6).

3. Scenario 3 assumes the full column rank of B, so the result follows from (2.6). In scenarios 2 and 4, g is strongly convex. Recall (2.4) with $\hat{\lambda} = \lambda^{k+1}$:

$$\langle y^{k+1} - y^*,\ B^T\left(\lambda^{k+1} - \lambda^* - \beta A(x^k - x^{k+1})\right) + Q(y^k - y^{k+1})\rangle \geq \nu_g\|y^{k+1} - y^*\|^2. \tag{3.19}$$

By the Cauchy-Schwarz inequality and $Q = \mathbf{0}$, we have

$$\nu_g \|y^{k+1} - y^*\| \leq \|B\| \|\lambda^{k+1} - \lambda^* - \beta A(x^k - x^{k+1})\|. \tag{3.20}$$

Therefore, the result follows from the R-linear convergence of x^k and λ^k. □

4. Applications.

This section describes several well-known optimization models on which Algorithm 2 not only enjoys global linear convergence but also often has easy-to-solve subproblems.

4.1. Convex regularization.

The following convex regularization model has been widely used in various applications:

$$\min_y f(By - b) + g(y) \tag{4.1}$$

where f is often a *strongly convex* function with Lipschitz continuous gradient, and g is a convex function which is very versatile across different applications. In particular g can be nonsmooth (e.g., projection to a convex set, ℓ_1-norm). Here, f and g are often referred to as the loss (or data fidelity) function and the regularization function, respectively. Model (4.1) can be reformulated to

$$\min_{x,y} f(x) + g(y), \quad \text{s.t. } x + By = b \tag{4.2}$$

and be solved by Algorithm 2. With many popular choices of f and g and also with proper P and Q, the x- and y-subproblems are easy to solve. If B has full column rank or g is strongly convex, then Algorithm 2 converges at a global linear rate.

4.2. Sparse optimization.

In recent years, the problem of recovering sparse vectors and low-rank matrices has received tremendous attention from researchers and engineers, particularly those in the areas of compressive sensing, machine learning, and statistics.

Elastic net (augmented ℓ_1) model. To recover a sparse vector $y^0 \in \mathbb{R}^n$ from linear measurements $b = By^0 \in \mathbb{R}^m$, the elastic net model solves

$$\min_y \ \|y\|_1 + \alpha \|y\|^2 + \frac{1}{2\mu}\|Ay - b\|^2, \tag{4.3}$$

where $A \in \mathbb{R}^{m \times n}$, $\alpha > 0$ and $\mu > 0$ are parameters, and the ℓ_1 norm $\|y\|_1 := \sum_{i=1}^n |y_i|$ is known to promote sparsity in the solution. It has been shown that the elastic model can effectively recover sparse vectors and outperform Lasso ($\alpha = 0$) on reported real-world regression problems [27]. With the constraint $x = y$, (4.3) can be reformulated as:

$$\begin{aligned} &\min_{x,y} \ \|y\|_1 + \alpha \|x\|^2 + \frac{1}{2\mu}\|Ax - b\|^2 \\ &\text{s.t. } x - y = 0. \end{aligned} \tag{4.4}$$

Augmented nuclear-norm model. Similarly, the elastic net model can be extended for recovering low-rank matrices. To recover a low-rank matrix $Y^0 \in \mathbb{R}^{n_1 \times n_2}$ from linear measurements $b = \mathcal{B}(Y^0) \in \mathbb{R}^m$, the augmented nuclear-norm model solves

$$\min_Y \ \|Y\|_* + \alpha \|Y\|_F^2 + \frac{1}{2\mu}\|\mathcal{A}(Y) - b\|^2, \tag{4.5}$$

where $\alpha > 0$ and $\mu > 0$ are parameters, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear operator, $\| \cdot \|_F$ denotes the Frobenius norm, and the nuclear norm $\|Y\|_*$ denotes the sum of singular values of Y which is known to promote low-rankness in the solution. By variable splitting $X = Y$, (4.5) can be reformulated as:

$$\begin{aligned} &\min_{X,Y} \|Y\|_* + \alpha\|X\|_F^2 + \frac{1}{2\mu}\|\mathcal{A}(X) - b\|^2 \\ &\text{s.t. } X - Y = 0. \end{aligned} \tag{4.6}$$

In (4.4) and (4.6), the functions $f(x) = \alpha\|x\|^2 + \frac{1}{2\mu}\|Ax - b\|^2$ and $f(X) = \alpha\|X\|_F^2 + \frac{1}{2\mu}\|\mathcal{A}(X) - b\|^2$ are strongly convex and have Lipschitz continuous gradient; the functions $g(y) := \|y\|_1$ and $g(Y) := \|Y\|_*$ are convex and nonsmooth. In fact, $\|\cdot\|^2$ and $\|\cdot\|_F^2$ can also be replaced by many other choices of functions that are strongly convex and have Lipschitz continuous gradient, or become so when restricted to a bounded set. Note that if $\alpha = 0$ then the functions f may not be strongly convex if the matrix A and the linear operator $\mathcal{A}$ do not have full column rank. In many applications, this is indeed the case since the number of observations of y and Y is usually smaller than their dimensions (i.e., $m < n$ and $m < n_1 \cdot n_2$). However, the parameter $\alpha > 0$ guarantees the strong convexity of f, and hence the global linear convergence of Algorithm 2 when applied to (4.4) and (4.6). In addition, it has been shown in [28] that for most compressive sensing problems, with a moderately small α, problems (4.4) and (4.6) return solutions as if $\alpha = 0$.

4.3. Consensus and sharing optimization. Consider in a network of N nodes, the problem of minimizing the sum of N functions, one from each node, over a common variable x. This problem can be written as

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^N f_i(x). \tag{4.7}$$

Let each node i keep vector $x_i \in \mathbb{R}^n$ as its copy of x. To reach a consensus among x_i, $i = 1, \dots, N$, a common approach is to introduce a global common variable y and get

$$\min_{\{x_i\},y} \sum_{i=1}^N f_i(x_i), \quad \text{s.t. } x_i - y = 0,\ i = 1, \dots, N. \tag{4.8}$$

This is the well-known global consensus problem; see [7] for a review. With an objective function g on the global variable y, we have the global variable consensus problem with regularization:

$$\min_{\{x_i\},y} \sum_{i=1}^N f_i(x_i) + g(y), \quad \text{s.t. } x_i - y = 0,\ i = 1, \dots, N, \tag{4.9}$$

where $g(y)$ is a convex function,

The following sharing problem is also nicely reviewed in [7]:

$$\min_{\{x_i\},y} \sum_{i=1}^N f_i(x_i) + g\left(\sum_{i=1}^N y_i\right), \quad \text{s.t. } x_i - y_i = 0,\ i = 1, \dots, N, \tag{4.10}$$

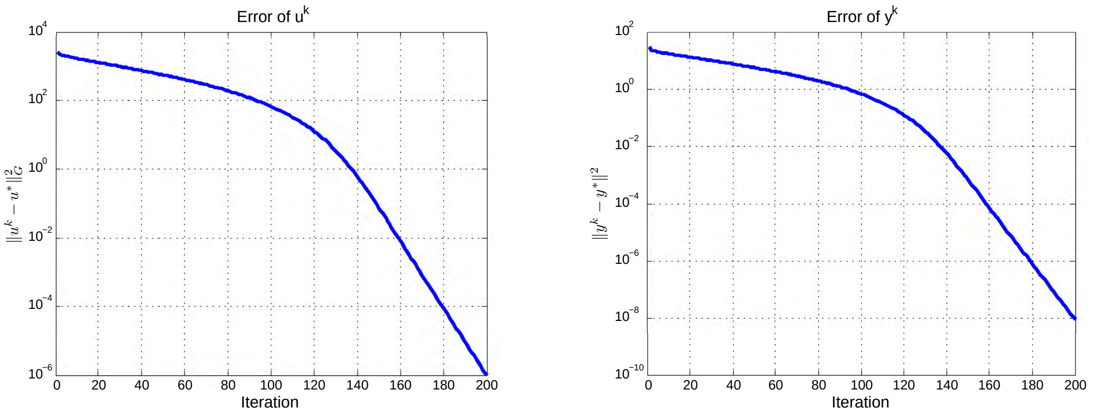
where f_i's are local cost functions and g is the shared cost function by all the nodes i.

Algorithm 2 applied to the problems (4.8), (4.9) and (4.10) converges linearly if each function f_i is strongly convex and has Lipschitz continuous gradient. The resulting ADM is particularly suitable for distributed implementation, since the x-subproblem can be decomposed into N independent x_i-subproblems, and the update to the multiplier λ can also be done at each node i.

5. Numerical demonstration. We present the results of some simple numerical tests to demonstrate the linear convergence of Algorithm 2. The numerical performance is not the focus of this paper and will be investigated more thoroughly in future research.

5.1. Elastic net. We apply Algorithm 2 with $P = 0$ and $Q = 0$ to a small elastic net problem (4.4), where the feature matrix A has $m = 250$ examples and $n = 1000$ features. We first generated the matrix A from the standard Gaussian distribution $\mathcal{N}(0, 1)$ and then orthonormalized its rows. A sparse vector $x^0 \in \mathbb{R}^n$ was generated with 25 nonzero entries, each sampled from the standard Gaussian distribution. The observation vector $b \in \mathbb{R}^m$ was then computed by $b = Ax^0 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 10^{-3}I)$. We chose the model parameters $\alpha = 0.1$ and $\mu = 10^{-2}$, which we found to yield reasonable accuracy for recovering the sparse solution. We initialized all the variables at zero and set the algorithm parameters $\beta = 100$ and $\gamma = 1$. We ran the algorithm for 200 iterations and recorded the errors at each iteration with respect to a precomputed reference solution u^*.

Figure 5.1(a) shows the decreasing behavior of $\|u^k - u^*\|_G^2 (:= \beta\|x^k - x^*\|^2 + \|\lambda^k - \lambda^*\|^2/\beta)$ as the algorithm progresses. Since variable y is not contained in the G-norm, we also plot the convergence curve of $\|y^k - y^*\|^2$ in Figure 5.1(b). We observe that both u^k and y^k converge at similar linear rates. In addition, the convergence appears to have different stages. The later stage exhibits faster convergence rate than the earlier stage. This can be clearly seen in Figure 5.2 which depicts the Q-linear rate $\|u^{k+1} - u^*\|_G^2 / \|u^k - u^*\|_G^2$.



(a) $\|u^k - u^*\|_G^2$ vs. iteration

(b) $\|y^k - y^*\|^2$ vs. iteration

FIG. 5.1. *Convergence curves of ADM for the elastic net problem.*

Here, the strong convexity constant of f is $\nu_f = 2\alpha + \lambda_{\min}(A^T A)/\mu = 2\alpha$ and the Lipschitz constant of ∇f is $L_f = 2\alpha + \lambda_{\max}(A^T A)/\mu = 2\alpha + 1/\mu$. By (3.14), our bound for the global linear rate is $(1+\delta)^{-1} = 0.998$, which roughly matches the early-stage rate shown in the figure. However, our theoretical bound is rather conservative, since it is a global worst-case bound and it does not take into account the properties of the ℓ_1 norm and the solution. In fact, the optimal solution x^* is very sparse and x^k will also become sparse after a number of iterations. Let $\mathcal{S}$ be an index set of the nonzero support of $(x^k - x^*)$, and $A_{\mathcal{S}}$ be a submatrix composed of those columns of A indexed by $\mathcal{S}$. Then, the constants ν_f and L_f in our bound can be effectively replaced by $\bar{\nu}_f = 2\alpha + \lambda_{\min}(A_{\mathcal{S}}^T A_{\mathcal{S}})/\mu$ and $\bar{L}_f = 2\alpha + \lambda_{\max}(A_{\mathcal{S}}^T A_{\mathcal{S}})/\mu$, thereby accounting for the faster convergence rate in the later stage. For example, letting $\mathcal{S}$ be the nonzero support of the optimal solution x^*, we obtain an estimate of the (asymptotic) linear rate $(1+\delta)^{-1} = 0.817$, which well matches the later-stage rate.

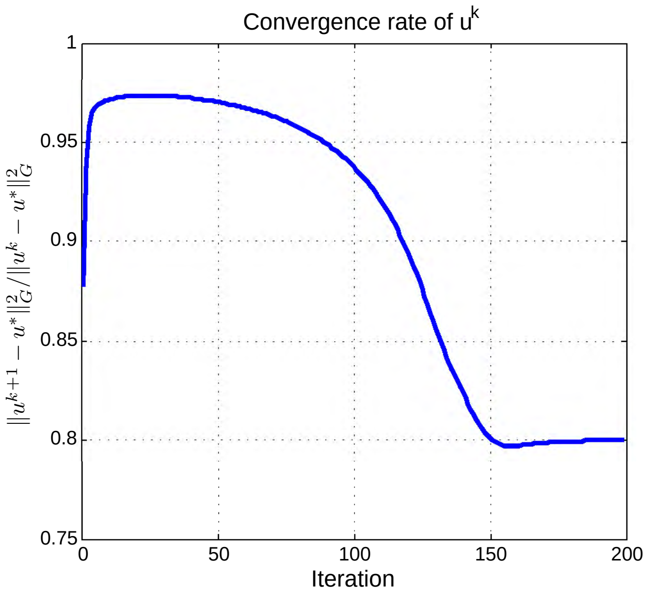


FIG. 5.2. *Q-linear convergence rate of ADM for the elastic net problem.*

5.2. Distributed Lasso. We consider solving the Lasso problem in a distributed way [29]:

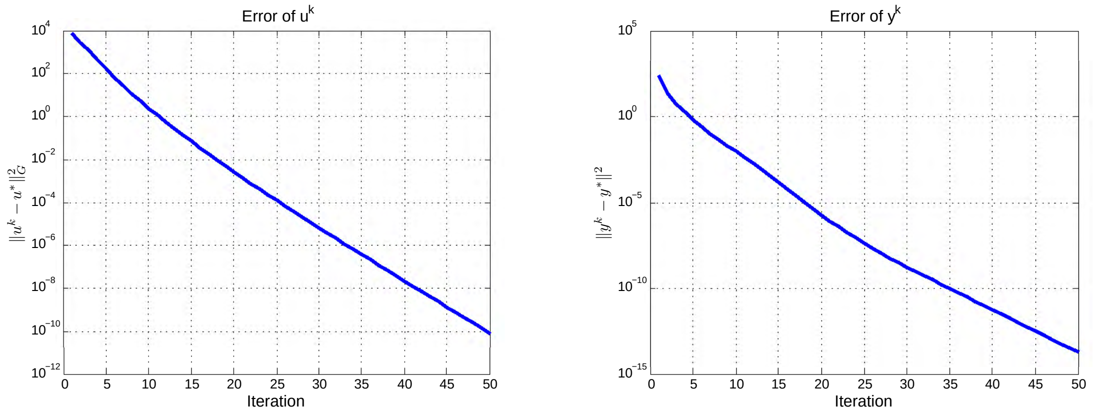
$$\begin{aligned}\min_{\{x_i\},y} \quad & \sum_{i=1}^{N} \frac{1}{2\mu}\|A_i x_i - b_i\|^2 + \|y\|_1 \\ \text{s.t.} \quad & x_i - y = 0,\ i = 1,\dots,N,\end{aligned} \tag{5.1}$$

which is an instance of the global consensus problem with regularization (4.9).

We apply Algorithm 2 with $P = 0$ and $Q = 0$ to a small distributed Lasso problem (5.1) with $N = 5$, where each A_i has $m = 600$ examples and $n = 500$ features. Each A_i is a tall matrix and has full column rank, yielding a strongly convex objective function in x_i. Therefore, Algorithm 2 is guaranteed to converge linearly.

We generated the data similarly as in the elastic net test. We randomly generated each A_i from the standard Gaussian distribution $\mathcal{N}(0,1)$, and then simply scaled its columns to have a unit length. We generated a sparse vector $x^0 \in \mathbb{R}^n$ with 250 nonzero entries, each sampled from the $\mathcal{N}(0,1)$ distribution. Each $b_i \in \mathbb{R}^m$ was then computed by $b_i = A_i x^0 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 10^{-3} I)$. We chose the model parameter $\mu = 0.1$, which we found to yield reasonably good recovery quality. From the initial point at zero, we ran the algorithm with parameters $\beta = 10$ and $\gamma = 1$ for 50 iterations and computed the iterative errors.

Figure 5.3 demonstrates the clear linear convergence behavior of $\|u^k - u^*\|_G^2$ and $\|y^k - y^*\|^2$. In Figure 5.4, the Q-linear convergence rate of $\|u^k - u^*\|_G^2$ is depicted. For this problem, the strong convexity constant is $\nu_f = \min_i\{\lambda_{\min}(A_i^T A_i)/\mu\}$ and the Lipschitz constant is $L_f = \max_i\{\lambda_{\max}(A_i^T A_i)/\mu\}$. However, the condition number ν_f/L_f in this test is relatively big, and hence the theoretical linear rate specified by (3.15) is not a very tight bound for the observed fast rate. Note that all x_i's tend to be equal and become sparse after a number of iterations. Similar to our previous discussion in Section 5.1, we can estimate the asymptotic linear rate by letting $\bar{\nu}_f = \lambda_{\min}(A_{\mathcal{S}}^T A_{\mathcal{S}})/(\mu N)$ and $\bar{L}_f = \lambda_{\max}(A_{\mathcal{S}}^T A_{\mathcal{S}})/(\mu N)$, where $A \in \mathbb{R}^{Nm\times n}$ is formed by stacking all the matrices A_i $(i = 1,\dots,N)$, and $\mathcal{S}$ is an index set of the nonzero support of x^*. We obtained the asymptotic linear rate to be $(1+\delta)^{-1} = 0.779$, which appears to be a much tighter bound.



(a) $\|u^k - u^*\|_G^2$ vs. iteration

(b) $\|y^k - y^*\|^2$ vs. iteration

FIG. 5.3. *Convergence curves of ADM for the distributed Lasso problem.*

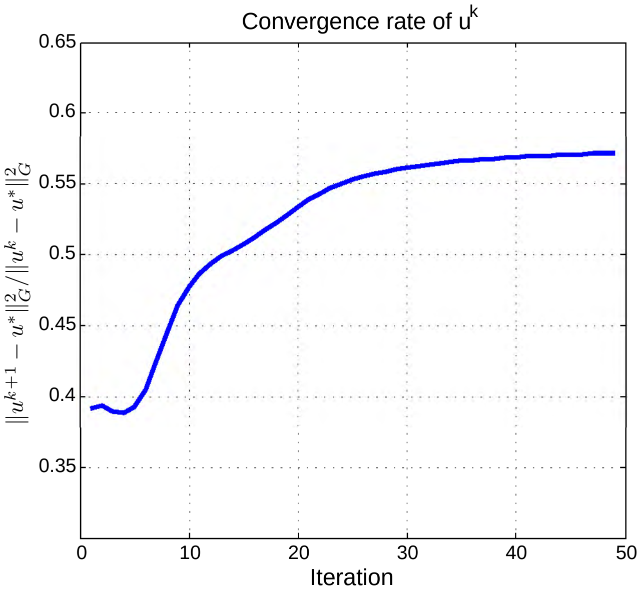


FIG. 5.4. *Q-linear convergence rate of ADM for the distributed Lasso problem.*

6. Conclusions. In this paper, we provide sufficient conditions for the global linear convergence of a general class of ADMs which solve subproblems either exactly or approximately in a certain manner. Among the conditions is a function that is strongly convex and has Lipschitz continuous gradient. These sufficient conditions cover a wide range of applications. We also extend the existing convergence theory to allow more generality on the step size γ for updating the multipliers.

In practice, how to choose the penalty parameter β is always an important issue. Our convergence rate analysis provides more insights on how the penalty parameter β affects the convergence speed, thereby providing some theoretical guidance for choosing β.

Acknowledgements. The authors would like to thank Professors Stephen Boyd, Qing Ling, Shiqian Ma, Zaiwen Wen, Xiaoming Yuan, and Yin Zhang for their valuable discussions and comments that have helped improve this manuscript. The authors' work is supported by ARL, NSF, and ONR.

REFERENCES

[1] J.M. Mendel and C.S. Burrus. *Maximum-likelihood deconvolution: a journey into model-based signal processing*. springer-Verlag New York, 1990.

[2] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[3] E. Esser. Applications of lagrangian-based alternating direction methods and connections to split bregman. *CAM report 09-31, UCLA*, 2009.

[4] T. Goldstein and S. Osher. The split bregman method for l1 regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.

[5] T. Goldstein, X. Bresson, and S. Osher. Geometric applications of the split bregman method: Segmentation and surface reconstruction. *Journal of Scientific Computing*, 45(1):272–293, 2010.

[6] J.F. Cai, S. Osher, and Z. Shen. Split bregman methods and frame based image restoration. *Multiscale modeling & simulation*, 8(2):337, 2009.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[8] J. Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1-2):250–278, 2011.

[9] W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. *TR11-06, Department of Computational and Applied Mathematics, Rice University*, 2011.

[10] H. Jiang, W. Deng, and Z. Shen. Surveillance video processing using compressive sensing. *Inverse Problems and Imaging*, 6(2), 2012.

[11] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[12] R. Glowinski and A. Marrocco. *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*. Laboria, 1975.

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

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

[15] B. He and X. Yuan. On non-ergodic convergence rate of douglas-rachford alternating direction method of multipliers. 2012.

[16] T. Goldstein, B. O'Donoghue, and S. Setzer. Fast alternating direction optimization methods. *CAM report 12-35, UCLA*, May 2012.

[17] Z.Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Arxiv preprint arXiv:1208.3922*, 2012.

[18] J. Eckstein, D.P. Bertsekas, Massachusetts Institute of Technology. Laboratory for Information, and Decision Systems. *An alternating direction method for linear programming*. Division of Research, Harvard Business School, 1990.

[19] D. Boley. Linear convergence of admm on a model problem. *TR 12-009, Department of Computer Science and Engineering, University of Minnesota*, 2012.

[20] R. Glowinski. Numerical methods for nonlinear variational problems. 1984.

[21] B. He, L.Z. Liao, D. Han, and H. Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

[22] X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

[23] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1):81–101, 1994.

[24] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[25] D. Goldfarb and W. Yin. Parametric maximum flow algorithms for fast total variation minimization. *SIAM Journal on Scientific Computing*, 31(5):3712–3743, 2009.

[26] R.T. Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1997.

[27] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[28] M.J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *Arxiv preprint arXiv:1201.4615*, 2012.

[29] G. Mateos, J.A. Bazerque, and G.B. Giannakis. Distributed sparse linear regression. *Signal Processing, IEEE Transactions on*, 58(10):5262–5276, 2010.