Computational Optimization and Applications (2020) 77:729–754
https://doi.org/10.1007/s10589-020-00221-y

Convergence rates for an inexact ADMM applied to separable convex optimization

William W. Hager[1] · **Hongchao Zhang**[2]

Received: 13 January 2020 / Published online: 25 August 2020

Abstract
Convergence rates are established for an inexact accelerated alternating direction method of multipliers (I-ADMM) for general separable convex optimization with a linear constraint. Both ergodic and non-ergodic iterates are analyzed. Relative to the iteration number k, the convergence rate is $\mathcal{O}(1/k)$ in a convex setting and $\mathcal{O}(1/k^2)$ in a strongly convex setting. When an error bound condition holds, the algorithm is 2-step linearly convergent. The I-ADMM is designed so that the accuracy of the inexact iteration preserves the global convergence rates of the *exact* iteration, leading to better numerical performance in the test problems.

Keywords Separable convex optimization · Alternating direction method of multipliers · ADMM · Accelerated gradient method · Inexact methods · Global convergence · Convergence rates

Mathematics Subject Classification 90C06 · 90C25 · 65Y20

1 Introduction

We consider a convex, separable linearly constrained optimization problem

The authors gratefully acknowledge support by the National Science Foundation under Grants 1819002 and 1819161, and by the Office of Naval Research under Grants N00014-15-1-2048 and N00014-18-1-2100.

✉ William W. Hager
hager@ufl.edu
http://people.clas.ufl.edu/hager/

Hongchao Zhang
hozhang@math.lsu.edu
http://math.lsu.edu/~hozhang/

[1] Department of Mathematics, University of Florida, PO Box 118105, Gainesville, FL 32611-8105, USA

[2] Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA

$$\min \ \Phi(\mathbf{x}) \ \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\mathbf{A}$ is N by n. By a separable convex problem, we mean that the objective function is a sum of m independent parts, and the matrix is partitioned compatibly as in

$$\Phi(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i) \quad \text{and} \quad \mathbf{A}\mathbf{x} = \sum_{i=1}^{m} \mathbf{A}_i\mathbf{x}_i. \tag{1.2}$$

Here f_i is convex and Lipschitz continuously differentiable, h_i is a proper closed convex function (possibly nonsmooth), and $\mathbf{A}_i$ is N by n_i with $\sum_{i=1}^{m} n_i = n$. There is no column independence assumption for the $\mathbf{A}_i$. Constraints of the form $\mathbf{x}_i \in \mathcal{X}_i$, where $\mathcal{X}_i$ is a closed convex set, can be incorporated in the optimization problem by letting h_i be the indicator function of $\mathcal{X}_i$. That is, $h_i(\mathbf{x}_i) = \infty$ when $\mathbf{x}_i \notin \mathcal{X}_i$. The problem (1.1)–(1.2) has attracted extensive research due to its importance in areas such as image processing, statistical learning, and compressed sensing. See the recent survey [2] and its references.

It is assumed that there exists a solution $\mathbf{x}^*$ to (1.1)–(1.2) and an associated Lagrange multiplier $\lambda^* \in \mathbb{R}^N$ such that the following first-order optimality conditions hold: $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ and for $i = 1, 2, \ldots, m$ and for all $\mathbf{u} \in \mathbb{R}^{n_i}$, we have

$$\langle \nabla f_i(\mathbf{x}_i^*) + \mathbf{A}_i^{\mathsf{T}} \lambda^*, \mathbf{u} - \mathbf{x}_i^* \rangle + h_i(\mathbf{u}) \geq h_i(\mathbf{x}_i^*), \tag{1.3}$$

where ∇ denotes the gradient.

A popular strategy for solving (1.1)–(1.2) is the alternating direction method of multipliers (ADMM) [16, 17]: For $i = 1, \ldots, m$,

$$\begin{cases} \mathbf{x}_i^{k+1} \in \arg\min\limits_{\mathbf{x}_i \in \mathbb{R}^{n_i}} \mathcal{L}_\rho(\mathbf{x}_1^{k+1}, \ldots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^k, \ldots, \mathbf{x}_m^k, \lambda^k), \\ \lambda^{k+1} = \lambda^k + \rho(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}), \end{cases} \tag{1.4}$$

where ρ is a penalty parameter and $\mathcal{L}_\rho$ is the augmented Lagrangian defined by

$$\mathcal{L}_\rho(\mathbf{x}, \lambda) = \Phi(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2. \tag{1.5}$$

Early ADMMs only consider problem (1.1)–(1.2) with $m = 2$ corresponding to a 2-block structure. In this case, the global convergence and complexity can be found in [12, 28]. When $m \geq 3$, the ADMM strategy (1.4) is not necessarily convergent [4], although its practical efficiency has been observed in many recent applications [40, 41]. Many recent papers, including [3, 5, 6, 11, 18, 24, 26, 27, 32, 33], develop modifications to ADMM to ensure convergence when $m \geq 3$. The approach we have taken employs a back substitution step to complement the ADMM forward substitution step. This modification was first introduced in [26, 27].

Much of the CPU time in an ADMM iteration is associated with the solution of the minimization subproblems. If $m = 1$, then ADMM reduces to the augmented Lagrangian method, for which the first relative error criteria based on the residual in an iteration emanates from [37], while more recent work includes [13, 39]. For $m = 2$ or larger, inexact approaches to the ADMM subproblems have been based on an absolute summable error criterion as in [9, 12, 19], a combined adaptive/absolute summable error criterion [31], a relative error criteria [14, 15], proximal regularizations [7, 25], and linearized subproblems and reduced multiplier update steps [30].

The approach taken in our I-ADMM emanates from our earlier work [10, 20, 21] on a Bregman Operator Splitting algorithm with a variable stepsize (BOSVS) with application to image processing. In the current paper, the penalty term in the accelerated gradient algorithm of [21] is linearized so as to make the solution of the I-ADMM subproblem trivial; there is essentially no reduction in the size of the multiplier update step. The I-ADMM is designed so that the accuracy of the inexact solution of the ADMM subproblems is high enough to preserve the global convergence rates of the *exact* iteration. The global convergence results for I-ADMM are similar to those presented in [21]. However, there is no convergence rate analysis in [21]. In this paper, we focus on the convergence rate of I-ADMM. In particular, relative to the iteration number k, the convergence rate for I-ADMM is $\mathcal{O}(1/k)$ for ergodic iterates in the convex setting and $\mathcal{O}(1/k^2)$ for both ergodic and nonergodic iterates in a strongly convex setting. When an error bound condition holds, I-ADMM is 2-step linearly convergent. These convergence rates are consistent with those obtained for ADMM schemes that solve subproblems exactly including the $\mathcal{O}(1/k)$ rates in [28, 35, 38] for ergodic iterates, and the linear rates obtained in [23, 42] for a 2-block ADMM, and in [30] for the multi-block case and a sufficiently small stepsize in the multiplier update. For a more extensive review of linear convergence results for ADMMs, see [43]. But again, almost all the sublinear or linear convergence rate analysis is based on either a single linearization step to solve the subproblem or the exact solution of the (proximal) subproblem. An advantage of our inexact scheme, compared to the exact iteration, is that the computing time to achieve a given error tolerance is reduced, while maintaining global convergence and its rate.

The paper is organized as follows. Section 2 gives an overview of the inexact ADMM (I-ADMM) that will be analyzed. Section 3 reviews the global convergence results found in a companion paper [22]. These global convergence results are similar to those established for the inexact ADMM of [21]. Section 4 establishes a $\mathcal{O}(1/k)$ convergence rate of for ergodic iterates, and under a strong convexity assumption, an $\mathcal{O}(1/k^2)$ rate for both ergodic and nonergodic iterates. Section 5 gives 2-step linear convergence results when an error bound condition holds. Finally, Section 6 shows the observed convergence in some image recovery problems.

1.1 Notation

Throughout the paper, c denotes a generic positive constant which is independent of parameters such as the iteration number k or the index $i \in [1, m]$. Let $\mathcal{W}^*$ denote the set of solution/multiplier pairs $(\mathbf{x}^*, \lambda^*)$ of (1.1)–(1.2) satisfying (1.3), while $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ is a generic solution/multiplier pair. $\mathcal{L}$ (without the ρ subscript) stands for $\mathcal{L}_0$. For $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T}\mathbf{y}$ is the standard inner product, where the superscript $^\mathsf{T}$ denotes transpose. The Euclidean vector norm, denoted $\| \cdot \|$, is defined by $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ and $\|\mathbf{x}\|_\mathbf{G} = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{G}\mathbf{x}}$ for a positive definite matrix $\mathbf{G}$. For any matrix $\mathbf{A}$, the matrix norm induced by the Euclidean vector norm is the largest singular value of $\mathbf{A}$. For a symmetric matrix, the Euclidean norm is the largest absolute eigenvalue. In addition, $\mathbf{A} \succ \mathbf{0}$ and $\mathbf{A} \succeq \mathbf{0}$ mean that the matrix $\mathbf{A}$ is positive definite and positive semidefinite, respectively. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla f(\mathbf{x})$ is the gradient of f at $\mathbf{x}$, a column vector. More generally, $\partial f(\mathbf{x})$ denotes the subdifferential at $\mathbf{x}$. A function $h : \mathbb{R}^n \mapsto \mathbb{R}$ is convex with modulus $\mu \geq 0$ if

$$h((1-\theta)\mathbf{x} + \theta\mathbf{y}) \leq (1-\theta)h(\mathbf{x}) + \theta h(\mathbf{y}) - \theta(1-\theta)(\mu/2)\|\mathbf{x} - \mathbf{y}\|^2$$

for all $\mathbf{u}$ and $\mathbf{v} \in \mathbb{R}^n$ and $\theta \in [0, 1]$. If $\mu > 0$, then h is strongly convex. The prox operator associated with h is defined by

$$\text{prox}_{\,h}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left(h(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 \right).$$

2 Algorithm structure

The structure of our I-ADMM algorithm is given in Algorithm 2.1. The algorithm generates sequences $\mathbf{x}^k$, $\mathbf{y}^k$, $\mathbf{z}^k$, and R^k. Both $\mathbf{x}^k$ and $\mathbf{z}^k$ are updated in Step 1, R^k is updated in Step 2, and $\mathbf{y}^k$ is updated in Step 3. The error is estimated in Step 2. The matrix $\mathbf{Q}$ in Step 3 is an m by m block diagonal matrix whose i-th diagonal block, denoted $\mathbf{Q}_i$, is chosen to satisfy the conditions:

$$\mathbf{Q}_i \succ \mathbf{0} \quad \text{and} \quad \overline{\mathbf{Q}}_i := \mathbf{Q}_i - \mathbf{A}_i^\mathsf{T}\mathbf{A}_i \succeq \mathbf{0}. \tag{2.1}$$

For example, we could take $\mathbf{Q}_i = \gamma_i \mathbf{I}$ where $\gamma_i \geq \|\mathbf{A}_i^\mathsf{T}\mathbf{A}_i\|$. Condition (2.1) is required for showing global convergence of our I-ADMM. Recent studies show that for the 2-block case ($m = 2$) and an exact ADMM, the requirement that $\overline{\mathbf{Q}}_i$ is positive semidefinite can be relaxed [8, 29]. The matrix $\mathbf{M}$ in Step 3 is the m by m block lower triangular matrix defined by

$$\mathbf{M}_{ij} = \begin{cases} \mathbf{A}_i^\mathsf{T}\mathbf{A}_j & \text{if } j < i, \\ \mathbf{Q}_i & \text{if } j = i, \\ \mathbf{0} & \text{if } j > i. \end{cases} \tag{2.2}$$

By (2.1), M is nonsingular. The solution $\mathbf{y}^{k+1}$ of the block upper triangular system $\mathbf{M}^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^k) = \alpha\mathbf{Q}(\mathbf{z}^k - \mathbf{y}^k)$ can be obtained by back substitution.

Parameters: ρ, $\delta_{\min}$, $\theta_i > 0$, $\alpha \in (0,1)$, $\sigma \in (0,1)$

Starting guess: $\mathbf{x}^1$ and $\boldsymbol{\lambda}^1$.

Initialize: $\mathbf{y}^1 = \mathbf{x}^1$, $k = 1$ and $\Gamma_i^0 = 0$, $1 \le i \le m$, $\epsilon^0 = \infty$

Step 1: For $i = 1, \ldots, m$

Generate $\mathbf{x}_i^{k+1}$, $\mathbf{z}_i^k$, and r_i^k by Algorithm 2.2.

End

Step 2: If $\epsilon^k := \theta_1\|\mathbf{z}^k - \mathbf{y}^k\| + \theta_2\|\mathbf{A}\mathbf{z}^k - \mathbf{b}\| + \theta_3\sqrt{R^k}$ is sufficiently small, then terminate, where $R^k = \sum_{i=1}^m r_i^k$.

Step 3: Find $\mathbf{y}^{k+1}$ by solving $\mathbf{Q}^{-1}\mathbf{M}^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^k) = \alpha(\mathbf{z}^k - \mathbf{y}^k)$

$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha\rho(\mathbf{A}\mathbf{z}^k - \mathbf{b})$, where $\mathbf{Q}$ and $\mathbf{M}$ are defined in (2.1) and (2.2), respectively.

Step 4: $k := k + 1$, and go to Step 1.

Alg. 2.1. I-ADMM algorithm.

In Step 1 of Algorithm 2.1, we approximate the minimizer in the $\mathbf{x}_i$ subproblem of the ADMM algorithm (1.4) using the accelerated gradient method of Algorithm 2.2, which is a modification of Algorithm 5.1 in [21]. Compared with Algorithm 5.1 in [21], Algorithm 2.2 has a slightly different stopping condition in Step 1b, and a proximal term to generate $\mathbf{u}_i^l$ in Step 1a, where

$$\mathbf{b}_i^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j\mathbf{z}_j^k - \sum_{j>i} \mathbf{A}_j\mathbf{y}_j^k. \tag{2.3}$$

Inner loop of Step 1, an accelerated gradient method:

Initialize: $\mathbf{a}_i^0 = \mathbf{u}_i^0 = \mathbf{x}_i^k$ and $\alpha^1 = 1$.

For $l = 1, 2, \ldots$

1a. Choose $\delta^l \geq \delta_{\min}$ and when $l > 1$, choose $\alpha^l \in (0, 1)$ such that

$$f_i(\overline{\mathbf{a}}_i^l) + \langle \nabla f_i(\overline{\mathbf{a}}_i^l), \mathbf{a}_i^l - \overline{\mathbf{a}}_i^l \rangle + \tfrac{(1-\sigma)\delta^l}{2\alpha^l} \|\mathbf{a}_i^l - \overline{\mathbf{a}}_i^l\|^2 \geq f_i(\mathbf{a}_i^l),$$

where $\mathbf{a}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^l$, $\overline{\mathbf{a}}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^{l-1}$, and $\mathbf{u}_i^l = \arg\min\{P(\mathbf{u}) + \frac{\varrho}{2}\|\mathbf{u} - \mathbf{y}_i^k\|_{\mathbf{Q}_i}^2 + h_i(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{n_i}\}$ with

$$P(\mathbf{u}) = \langle \nabla f_i(\overline{\mathbf{a}}_i^l), \mathbf{u} \rangle + \tfrac{\delta^l}{2}\|\mathbf{u} - \mathbf{u}_i^{l-1}\|^2 + \tfrac{\rho}{2}\|\mathbf{A}_i \mathbf{u} - \mathbf{b}_i^k + \boldsymbol{\lambda}^k/\rho\|^2,$$

and $\mathbf{b}_i^k$ defined in (2.3).

1b. If $\gamma^l = (1/\delta^1)\prod_{j=2}^{l}(1-\alpha^j)^{-1} \geq \Gamma_i^{k-1}$, where $\gamma^1 = 1/\delta^1$, and $\|\mathbf{a}_i^l - \mathbf{x}_i^k\|/\sqrt{\gamma^l} \leq \psi(\epsilon^{k-1})$, then break.

Next

1c. Set $\mathbf{x}_i^{k+1} = \mathbf{u}_i^l$, $\mathbf{z}_i^k = \mathbf{a}_i^l$, $\Gamma_i^k = \gamma^l$, and $r_i^k = (1/\Gamma_i^k)\sum_{j=1}^{l}\|\mathbf{u}_i^j - \mathbf{u}_i^{j-1}\|^2$.

ALG. 2.2. *Inner loop in Step* 1 *of Algorithm* 2.1.

The termination condition for Algorithm 2.2 appears in Step 1b. In this step, ψ is a nonnegative function for which $\psi(0) = 0$ and $\psi(s) > 0$ for $s > 0$ with ψ continuous at $s = 0$. For example, $\psi(t) = t$. Two different ways are developed in [21] for choosing the parameters δ^l and α^l in Step 1a. If a Lipschitz constant ζ_i of f_i is known, then we could take

$$\delta^l = \frac{1}{(1-\sigma)}\frac{2\zeta_i}{l} \quad \text{and} \quad \alpha^l = \frac{2}{l+1} \in (0, 1], \tag{2.4}$$

in which case, we have

$$\frac{(1-\sigma)\delta^l}{\alpha^l} = \frac{(l+1)\zeta_i}{l} > \zeta_i.$$

This relation along with a Taylor series expansion of f_i around $\overline{\mathbf{a}}_i^l$ implies that the line search condition in Step 1a of Algorithm 2.2 is satisfied for each l.

A different, adaptive way to choose to choose δ^l and α^l, that does not require knowledge of the Lipschitz constant for f_i, is the following: Choose $\delta_0^l \in [\delta_{\min}, \delta_{\max}]$, where $0 < \delta_{\min} < \delta_{\max} < \infty$ are fixed constants, independent of k and l, and set

$$\delta^l = \frac{2}{\theta^l + \sqrt{(\theta^l)^2 + 4\theta^l \Lambda^{l-1}}} \quad \text{and} \quad \alpha^l = \frac{1}{1 + \delta^l \Lambda^{l-1}}, \quad \text{where}$$
$$\Lambda^l = \sum_{i=1}^{l} 1/\delta^i, \quad \Lambda^0 = 0, \quad \text{and} \quad \theta^l = 1/(\delta_0^l \eta^j) \text{ with } \eta > 1. \tag{2.5}$$

Here the integer $j \geq 0$ is chosen as small a possible while satisfying the inequality in Step 1a. It can be shown that

$$\frac{\delta^l}{\alpha^l} = \frac{1}{\theta^l} = \delta_0^l \eta^j. \tag{2.6}$$

Since $\eta > 1$, the ratio δ^l/α^l appearing in Step 1a tends to infinity as j tends to infinity; consequently, the inequality in Step 1a is satisfied for j sufficiently large.

The stopping condition in Step 1b is elucidated using the following function:

$$\overline{L}_i^k(\mathbf{u}) = L_i^k(\mathbf{u}) + \frac{\rho}{2}(\mathbf{u} - \mathbf{y}_i^k)^\mathsf{T}\overline{\mathbf{Q}}_i(\mathbf{u} - \mathbf{y}_i^k), \quad \text{where}$$
$$L_i^k(\mathbf{u}) = f_i(\mathbf{u}) + h_i(\mathbf{u}) + \frac{\rho}{2}\|\mathbf{A}_i\mathbf{u} - \mathbf{b}_i^k + \lambda^k/\rho\|^2,$$

and $\mathbf{b}_i^k$ is defined in (2.3). As pointed out in Lemma 3.1 of the next section, for either of the parameter choices (2.4) or (2.5), the iterates $\mathbf{a}_i^l$ of Algorithm 2.2 converge to the minimizer of the function $\overline{L}_i^k$ at rate $\mathcal{O}(1/l)$, while the objective values converge at rate $\mathcal{O}(1/l^2)$, which is optimal for first-order methods applied to general convex, possibly nonsmooth optimization problems. We let l_i^k denote the terminating value of l in Step 1b.

Remark 2.1 For the two parameter choices (2.4) and (2.5), it has been shown [21, pp. 227–228] that in Step 1b, $\gamma^l \geq l^2\Theta$ for some constant $\Theta > 0$, independent of k and l. Consequently, the conditions in Step 1b are satisfied for l sufficiently large.

3 Global convergence

The global convergence analysis of the accelerated ADMM in this paper with a linearized penalty term is similar to the global convergence analysis of the accelerated scheme in [21]. Hence, this section simply states the main results, while a supplementary arXiv document [22] provides the detailed analysis. The first result concerns the convergence of the iterates in Step 1 of I-ADMM under the assumption that the sequence

$$\xi^l := \delta^l \alpha^l \gamma^l$$

is nondecreasing. For either of the parameter choices (2.4) or (2.5), it is shown in [21, pp. 227–228] that $\xi^l = 1$.

Lemma 3.1 *If the sequence ξ^l is nonincreasing, then for each $i \in [1,m]$ and $L \geq 1$, we have*

$$\rho\nu_i \|\mathbf{a}_i^L - \overline{\mathbf{x}}_i^k\|^2 + \frac{\mu_{h,i}}{2}\sum_{l=1}^{L}\|\overline{\mathbf{x}}_i^k - \mathbf{a}_i^L\|^2 + \frac{\sigma}{\gamma^L}\sum_{l=1}^{L}\xi^l\|\mathbf{u}_i^l - \mathbf{u}_i^{l-1}\|^2 \leq \frac{\|\mathbf{x}_i^k - \overline{\mathbf{x}}_i^k\|^2}{\gamma^L}, \tag{3.1}$$

where $\mu_{h,i}$ is the modulus of convexity of h_i, $\nu_i > 0$ is the smallest eigenvalue of $\mathbf{Q}_i$, and

$$\overline{\mathbf{x}}_i^k = \arg\min\{\overline{L}_i^k(\mathbf{u}) \,:\, \mathbf{u} \in \mathbb{R}^{n_i}\}. \tag{3.2}$$

Since $\overline{L}_i^k$ is strongly convex, it has a unique minimizer. The following decay property plays an important role in the global convergence analysis.

Lemma 3.2 *Let $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ be any solution/multiplier pair for (1.1)–(1.2), let $\mathbf{x}^k$, $\mathbf{y}^k$, $\mathbf{z}^k$, $\mathbf{u}_k^l$, and λ^k be the iterates generated by Algorithm 2.1, and define*

$$\begin{aligned} E_k &= \rho\|\mathbf{y}^k - \mathbf{x}^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^k - \lambda^*\|^2 + \alpha\sum_{i=1}^{m}\frac{\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2}{\Gamma_i^k} \quad \text{and} \\ E_k^- &= \rho\|\mathbf{y}^k - \mathbf{x}^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^k - \lambda^*\|^2 + \alpha\sum_{i=1}^{m}\frac{\|\mathbf{x}_i^k - \mathbf{x}_i^*\|^2}{\Gamma_i^{k-1}}, \end{aligned} \tag{3.3}$$

where $\mathbf{P} = \mathbf{M}\mathbf{Q}^{-1}\mathbf{M}^\mathsf{T}$. If $\xi^l := \delta^l\alpha^l\gamma^l = 1$ for each l, then

$$\begin{aligned} &E_k - E_{k+1} \geq E_k - E_{k+1}^- \\ &\geq \alpha\left(2\Delta^k + \sigma R^k + \rho(1-\alpha)(\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{Q}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2) + \sum_{i=1}^{m}\mu_{h,i}\|\mathbf{z}_i^k - \mathbf{x}_i^*\|^2\right), \end{aligned} \tag{3.4}$$

where R^k is the residual defined in Step 2, $\mu_{h,i}$ is the modulus of convexity of h_i, and

$$\Delta^k = \mathcal{L}(\mathbf{z}^k, \lambda^*) - \Phi(\mathbf{x}^*) \geq 0. \tag{3.5}$$

Recall that $\mathcal{L} = \mathcal{L}_0$ is the ordinary Lagrangian associated with (1.1). This decay property is used to obtain the following global convergence result for I-ADMM.

Theorem 3.3 *Suppose the parameters δ^l and α^l in Algorithm 2.2 are chosen according to either (2.4) or (2.5). If I-ADMM performs an infinite number of iterations generating $\mathbf{y}^k$, $\mathbf{z}^k$, and λ^k, then the sequences $\mathbf{y}^k$ and $\mathbf{z}^k$ both approach a common limit $\mathbf{x}^*$, λ^k approaches a limit λ^*, and $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$.*

Theorem 3.3 considers the case of an infinite number of iterations. The following lemma considers the case where $\epsilon^k = 0$ within a finite number of iterations.

Lemma 3.4 *If* $\epsilon^k = 0$ *in Algorithm* 2.1, *then* $\mathbf{x}^{k+1} = \mathbf{x}^k = \mathbf{y}^k = \mathbf{z}^k$ *solves* (1.1)–(1.2) *and* $(\mathbf{x}^k, \lambda^k) \in \mathcal{W}^*$.

Proof If $\epsilon^k = 0$, then $r_i^k = 0$ for each i. It follows that

$$\mathbf{x}_i^k = \mathbf{u}_i^0 = \mathbf{u}_i^1 = \ldots = \mathbf{u}_i^l. \tag{3.6}$$

By Step 1c, $\mathbf{u}_i^l = \mathbf{x}_i^{k+1}$. By the definitions $\mathbf{a}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^l$ and $\overline{\mathbf{a}}_i^l = (1-\alpha^l)\mathbf{a}_i^{l-1} + \alpha^l \mathbf{u}_i^{l-1}$ where $\mathbf{a}_i^0 = \mathbf{u}_i^0 = \mathbf{x}_i^k$, we have $\mathbf{a}_i^l = \overline{\mathbf{a}}_i^l = \mathbf{x}_i^k$ for each l due to (3.6). Again, by Step 1c, $\mathbf{z}_i^k = \mathbf{x}_i^k$. Consequently, we have $\mathbf{x}^{k+1} = \mathbf{x}^k = \mathbf{z}^k$.

Let $\mathbf{x}^*$ denote $\mathbf{x}^k$. Then $\mathbf{x}^* = \mathbf{x}^{k+1} = \mathbf{x}^k = \mathbf{z}^k$. Since $\epsilon^k = 0$, Step 2 of Algorithm 2.1 implies that $\mathbf{y}^k = \mathbf{z}^k = \mathbf{x}^*$ and $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Consequently, we have

$$\mathbf{b}_i^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j \mathbf{z}_j^k - \sum_{j>i} \mathbf{A}_j \mathbf{y}_j^k = \mathbf{b} - \sum_{j<i} \mathbf{A}_j \mathbf{x}_j^* - \sum_{j>i} \mathbf{A}_j \mathbf{x}_j^* = \mathbf{A}_i \mathbf{x}_i^*.$$

With this substitution in $P(\mathbf{u})$ in Step 1a, it follows that $\mathbf{u}_i^l = \mathbf{x}_i^*$ minimizes over $\mathbf{u}$ the function

$$\langle \nabla f_i(\mathbf{x}_i^*), \mathbf{u} \rangle + \frac{\delta^l}{2}\|\mathbf{u} - \mathbf{x}_i^*\|^2 + \frac{\rho}{2}\|\mathbf{A}_i(\mathbf{u} - \mathbf{x}_i^*) + \lambda^k/\rho\|^2 + \frac{\rho}{2}\|\mathbf{u} - \mathbf{x}_i^*\|_{\mathbf{Q}_i}^2 + h_i(\mathbf{u}).$$

The first-order optimality condition for this minimizer $\mathbf{x}_i^*$ is the same as the first-order optimality condition (1.3), but with λ^* replaced by λ^k. Hence, $(\mathbf{x}^*, \lambda^k) \in \mathcal{W}^*$. □

Remark 3.1 In this paper, we have focused on algorithms based on an inexact minimization of $\overline{L}_i^k$ in Step 1 of Algorithm 2.1. In cases where f_i and h_i are simple enough that the exact minimizer $\overline{\mathbf{x}}_i^k$ of $\overline{L}_i^k$ can be quickly evaluated, we could simply set $\mathbf{x}_i^{k+1} = \mathbf{z}_i^k = \overline{\mathbf{x}}_i^k$, and $r_i^k = 0$ in Step 1 of I-ADMM, and proceed to Step 2. The global convergence results still hold.

4 Sublinear convergence rates

In this section, sublinear convergences rates are established for I-ADMM. We first establish an $\mathcal{O}(1/t)$ convergence rate for the ergodic iterates

$$\overline{\mathbf{z}}^t = \frac{1}{t}\sum_{k=1}^{t} \mathbf{z}^k \tag{4.1}$$

generated by I-ADMM.

Theorem 4.1 *Let* $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ *be any primal/dual solution pair for* (1.1)–(1.2) *and let* $\mathbf{z}^k$ *be generated by I-ADMM with* $\delta^l \alpha^l \gamma^l = 1$ *for each* l *and* k. *Then, we have*

$$\mathcal{L}(\overline{\mathbf{z}}^t, \lambda^*) - \Phi(\mathbf{x}^*) \le \frac{E_1}{2\alpha t},$$

where $\overline{\mathbf{z}}^t$ *is defined in* (4.1) *and* E_k *is defined in* (3.3).

Proof Discarding several nonnegative terms from (3.4), we have

$$2\alpha\Delta^k + E_{k+1} \le E_k.$$

Adding this inequality over k between 1 and t yields

$$2\alpha \sum_{k=1}^{t} \Delta^k + E_{t+1} \le E_1.$$

Hence, by the definition of Δ^k in (3.5), we have

$$2\alpha \sum_{k=1}^{t} \left[\mathcal{L}(\mathbf{z}^k, \lambda^*) - \Phi(\mathbf{x}^*)\right] \le E_1.$$

By the convexity of Φ and the definition (4.1), it follows that

$$2\alpha t\left[\mathcal{L}(\overline{\mathbf{z}}^t, \lambda^*) - \Phi(\mathbf{x}^*)\right] \le E_1.$$

This completes the proof. □

Note that the minimum of $\mathcal{L}(\mathbf{x}, \lambda^*)$ over $\mathbf{x} \in \mathbb{R}^n$ is attained at $\mathbf{x} = \mathbf{x}^*$, and $\mathcal{L}(\mathbf{x}^*, \lambda^*) = \Phi(\mathbf{x}^*)$. Hence, Theorem 4.1 bounds the difference between $\mathcal{L}(\overline{\mathbf{z}}^t, \lambda^*)$ and the minimum of $\mathcal{L}(\cdot, \lambda^*)$. We will strengthen the convergence rate to $\mathcal{O}(1/t^2)$ when a strong convexity assumption holds, and also obtain a convergence rate for nonergodic iterates.

Assumption 4.1 If $\mu_{f,i} \ge 0$ and $\mu_{h,i} \ge 0$ are the convexity moduli of f_i and h_i respectively, then

$$\mu = \min\,\{\mu_{f,i} + 3\mu_{h,i} \,:\, i = 1, \ldots, m\} > 0. \tag{4.2}$$

In the following theorem, we suppose that at the k-th iteration, the penalty parameter ρ is chosen in the following way:

$$\rho_k = (k_0 + k)\theta, \tag{4.3}$$

where

$$\theta = \frac{\alpha\mu}{8\|\mathbf{P}\|} \quad \text{and} \quad k_0 = \frac{4\|\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}\|}{\alpha(1-\alpha)}, \tag{4.4}$$

with μ defined in Assumption 4.1, $\alpha \in (0, 1)$ is the parameter in Algorithm 2.1, and $\mathbf{P} = \mathbf{M}\mathbf{Q}^{-1}\mathbf{M}^{\mathsf{T}}$. We have the following theorem:

Theorem 4.2 *Let* $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ *be any solution/multiplier pair for* (1.1)–(1.2), *let* $\mathbf{x}^k, \mathbf{y}^k, \mathbf{z}^k$ *and* λ^k *be generated by I-ADMM, and assume that Assumption* 4.1 *holds and* $\delta^l \alpha^l \gamma^l = 1$ *for each* l *and* k. *Suppose that for every* k, ρ_k *is given by* (4.3) *and* Γ_i^k *satisfies*

$$\frac{k}{\Gamma_i^k} \geq \frac{k+1}{\Gamma_i^{k+1}}, \quad 1 \leq i \leq m. \tag{4.5}$$

Then, for all $t > 0$, *we have*

$$\mathcal{L}(\tilde{\mathbf{z}}^t, \lambda^*) - \Phi(\mathbf{x}^*) \leq \frac{2\overline{c}}{\alpha[t(t+1) + 2k_0 t]} \tag{4.6}$$

and

$$\|\mathbf{y}^{t+1} - \mathbf{x}^*\|^2 \leq \frac{\overline{c}}{(t+k_0)^2 \theta}, \tag{4.7}$$

where

$$\tilde{\mathbf{z}}^t = \frac{2}{t(t+1) + 2k_0 t} \sum_{k=1}^{t} ((k_0 + k)\mathbf{z}^k), \tag{4.8}$$

and

$$\overline{c} = \frac{1}{\theta} \|\lambda^1 - \lambda^*\|^2 + \alpha(k_0 + 1) \sum_{i=1}^{m} \frac{\|\mathbf{x}_i^1 - \mathbf{x}_i^*\|^2}{\Gamma_i^1} + k_0^2 \theta \|\mathbf{y}^1 - \mathbf{x}^*\|_{\mathbf{P}}^2. \tag{4.9}$$

Proof By Assumption 4.1 and the definition (3.5) of Δ^k, we have

$$\Delta^k = \mathcal{L}(\mathbf{z}^k, \lambda^*) - \mathcal{L}(\mathbf{x}^*, \lambda^*) \geq \sum_{i=1}^{m} \frac{\mu_{f,i} + \mu_{h,i}}{2} \|\mathbf{z}_i^k - \mathbf{x}_i^*\|^2 = \sum_{i=1}^{m} \frac{\mu_{f,i} + \mu_{h,i}}{2} \|\mathbf{z}_{e,i}^k\|^2,$$

where $\mathbf{z}_e^k = \mathbf{z}^k - \mathbf{x}^*$. The inequality (3.4) of Lemma 3.2 relates the error in two consecutive iterations, where the ρ in (3.4) is the penalty at iteration k. Combining this with the definition of μ in Assumption 4.1, we have

$$\begin{aligned} &\alpha\left(\Delta^k + \frac{\mu}{2}\|\mathbf{z}_e^k\|^2 + \rho_k(1-\alpha)\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{Q}}^2\right) \\ &\quad \leq \rho_k(\|\mathbf{y}_e^k\|_{\mathbf{P}}^2 - \|\mathbf{y}_e^{k+1}\|_{\mathbf{P}}^2) + \frac{1}{\rho_k}(\|\lambda_e^k\|^2 - \|\lambda_e^{k+1}\|^2) + \alpha \sum_{i=1}^{m} \frac{\|\mathbf{x}_{e,i}^k\|^2 - \|\mathbf{x}_{e,i}^{k+1}\|^2}{\Gamma_i^k}, \end{aligned} \tag{4.10}$$

where $\mathbf{x}_e^k = \mathbf{x}^k - \mathbf{x}^*$, $\mathbf{y}_e^k = \mathbf{y}^k - \mathbf{x}^*$, and $\lambda_e^k = \lambda^k - \lambda^*$.

For any matrix $\mathbf{P}$, it follows from an eigendecomposition that

$$\mathbf{x}^{\mathsf{T}}\mathbf{x} \geq \frac{\mathbf{x}^{\mathsf{T}}\mathbf{P}\mathbf{x}}{\|\mathbf{P}\|} \quad \text{and} \quad \mathbf{x}^{\mathsf{T}}\mathbf{Q}\mathbf{x} \geq \frac{\mathbf{x}^{\mathsf{T}}\mathbf{P}\mathbf{x}}{\|\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}\|}.$$

The second inequality is deduced from the first when $\mathbf{x}$ is replaced by $\mathbf{Q}^{1/2}\mathbf{x}$ and $\mathbf{P}$ is replaced by $\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}$. This yields the following lower bound for terms on the left side of (4.10):

$$\begin{aligned}\frac{\mu}{2}\|\mathbf{z}_e^k\|^2 + \rho_k(1-\alpha)\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{Q}}^2 &\geq \frac{\mu}{2\|\mathbf{P}\|}\|\mathbf{z}_e^k\|_{\mathbf{P}}^2 + \frac{\rho_k(1-\alpha)}{\|\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}\|}\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{P}}^2 \\ &\geq \frac{\mu}{2\|\mathbf{P}\|}\left(\|\mathbf{z}_e^k\|_{\mathbf{P}}^2 + \|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{P}}^2\right) \\ &\geq \frac{\mu}{2\|\mathbf{P}\|}\left(2\|\mathbf{z}_e^k\|_{\mathbf{P}}^2 + \|\mathbf{y}_e^k\|_{\mathbf{P}} - 2\|\mathbf{z}_e^k\|\|\mathbf{y}_e^k\|\right) \\ &\geq \frac{\mu}{4\|\mathbf{P}\|}\|\mathbf{y}_e^k\|_{\mathbf{P}} = \frac{2\theta}{\alpha}\|\mathbf{y}_e^k\|_{\mathbf{P}}.\end{aligned} \tag{4.11}$$

The second inequality is due to the special form of ρ_k in (4.3) and (4.4), and the last inequality is due to the relation

$$ab \leq \frac{1}{2}\left(2a^2 + \frac{1}{2}b^2\right).$$

The inequality (4.11) is incorporated in the left side of (4.10). We multiply the resulting inequality by $K := k_0 + k$, substitute $\rho_k = K\theta$, exploit the assumption (4.5) and the inequality $K(K-2) \leq (K-1)^2$ to obtain

$$\begin{aligned}\alpha K \Delta^k \leq& \theta\left((K-1)^2\|\mathbf{y}_e^k\|_{\mathbf{P}}^2 - K^2\|\mathbf{y}_e^{k+1}\|_{\mathbf{P}}^2\right) + \frac{1}{\theta}(\|\lambda_e^k\|^2 - \|\lambda_e^{k+1}\|^2) \\ &+ \alpha\sum_{i=1}^m\left(\frac{K\|\mathbf{x}_{e,i}^k\|^2}{\Gamma_i^k} - \frac{(K+1)\|\mathbf{x}_{e,i}^{k+1}\|^2}{\Gamma_i^{k+1}}\right).\end{aligned}$$

Summing this inequality for k between 1 and t, with $K = k_0 + k$, yields

$$\alpha\sum_{k=1}^t(k_0+k)\Delta^k + (k_0+t)^2\theta\|\mathbf{y}^{t+1} - \mathbf{x}^*\|_{\mathbf{P}}^2 \leq \bar{c}, \tag{4.12}$$

where $\bar{c}$ is defined in (4.9). Substituting for Δ^k using (3.5) and discarding the $\mathbf{y}^{t+1}$ term, we have

$$\alpha\sum_{k=1}^t(k_0+k)\left[\mathcal{L}(\mathbf{z}^k, \lambda^*) - \Phi(\mathbf{x}^*)\right] \leq \bar{c}. \tag{4.13}$$

The convexity of Φ and the definition of $\tilde{\mathbf{z}}^k$ in (4.8) yield

$$\mathcal{L}(\tilde{\mathbf{z}}^k, \lambda^*)) \leq \frac{2}{t(t+1)+2k_0t}\sum_{k=1}^t(k_0+k)\mathcal{L}(\mathbf{z}^k, \lambda^*),$$

which together with (4.13) gives (4.6). In addition, since $\Delta^k \geq 0$, (4.12) also implies (4.7). □

As noted at the end of Sect. 2, for either of the parameter choices (2.4) or (2.5), $\gamma^l \geq l^2\Theta$ for some constant $\Theta > 0$, independent of k and l. Hence, for l sufficiently large, the requirement (4.5) at iteration $k+1$ is satisfied.

5 Linear convergence

For the analysis of linear convergence rate of I-ADMM, we assume that ψ has the additional property that $\psi(t) \leq c_\psi t$ for all $t \geq 0$, where $c_\psi > 0$ is a constant. Let us define

$$e_i(\mathbf{y}, \lambda) = \|\mathbf{y}_i - \operatorname{prox}_{h_i}(\mathbf{y}_i - \nabla f_i(\mathbf{y}_i) - \mathbf{A}_i^\top \lambda)\|. \tag{5.1}$$

We begin with the following lemma.

Lemma 5.1 *If the parameters δ^l and α^l in Algorithm 2.2 are chosen according to either (2.4) or (2.5) and $\psi(t) \leq c_\psi t$, then for any $k \geq 2$, we have*

$$\sum_{i=1}^m e_i(\mathbf{y}^{k+1}, \lambda^{k+1}) \leq c(d_k + d_{k-1}), \tag{5.2}$$

where $c > 0$ is a generic constant which only depends on the problem data and algorithm parameters such as ρ and *c_ψ and*

$$d_k = \|\mathbf{y}^k - \mathbf{z}^k\| + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\| + \sqrt{R^k}. \tag{5.3}$$

Proof For any $\mathbf{p}_i$ and $\mathbf{q}_i \in \mathbb{R}^{n_i}$, $i = 1, 2$, it follows from the triangle inequality and the nonexpansive property of the prox operator that

$$\begin{aligned} &\|\mathbf{p}_1 - \operatorname{prox}_{h_i}(\mathbf{q}_1)\| \\ &\quad = \|[\mathbf{p}_2 - \operatorname{prox}_{h_i}(\mathbf{q}_2)] + [\mathbf{p}_1 - \mathbf{p}_2] + [\operatorname{prox}_{h_i}(\mathbf{q}_2) - \operatorname{prox}_{h_i}(\mathbf{q}_1)]\| \\ &\quad \leq \|\mathbf{p}_2 - \operatorname{prox}_{h_i}(\mathbf{q}_2)\| + \|\mathbf{p}_1 - \mathbf{p}_2\| + \|\mathbf{q}_1 - \mathbf{q}_2\|. \end{aligned} \tag{5.4}$$

We identify $\|\mathbf{p}_1 - \operatorname{prox}_{h_i}(\mathbf{q}_1)\|$ with $e_i(\mathbf{y}^{k+1}, \lambda^{k+1})$ and $\|\mathbf{p}_2 - \operatorname{prox}_{h_i}(\mathbf{q}_2)\|$ with $e_i(\mathbf{z}^k, \lambda^k)$, and use (5.4) to obtain the following bound for $e_i(\mathbf{y}^{k+1}, \lambda^{k+1})$ in terms of $e_i(\mathbf{z}^k, \lambda^k)$:

$$e_i(\mathbf{y}^{k+1}, \lambda^{k+1}) \leq e_i(\mathbf{z}^k, \lambda^k) + (2 + \zeta_i)\|\mathbf{y}_i^{k+1} - \mathbf{z}_i^k\| + \|\mathbf{A}_i^\top(\lambda^{k+1} - \lambda^k)\|,$$

where ζ_i is the Lipschitz constant for ∇f_i. The update formula for λ^{k+1} implies that $\lambda^{k+1} - \lambda^k = \alpha\rho(\mathbf{A}\mathbf{z}^k - \mathbf{b}) = \alpha\rho\mathbf{r}_k$, where $\mathbf{r}_k = \mathbf{A}\mathbf{z}^k - \mathbf{b}$. With this substitution, the bound for $e_i(\mathbf{y}^{k+1}, \lambda_i^{k+1})$ becomes

$$e_i(\mathbf{y}^{k+1}, \lambda^{k+1}) \leq e_i(\mathbf{z}^k, \lambda^k) + (2 + \zeta_i)\|\mathbf{y}_i^{k+1} - \mathbf{z}_i^k\| + \alpha\rho\|\mathbf{A}_i^\top \mathbf{r}^k\|. \tag{5.5}$$

Let $\nu_i > 0$ denote the smallest eigenvalue of $\mathbf{Q}_i$. The analysis is partitioned into two cases:

Case 1 $\Gamma_i^k > 4/(\rho\nu_i)$. Again, by property (5.4), we have

$$e_i(\mathbf{z}^k, \lambda^k) \le e_i(\overline{\mathbf{x}}^k, \lambda^k) + (2 + \zeta_i)\|\mathbf{z}_i^k - \overline{\mathbf{x}}_i^k\|, \tag{5.6}$$

where $\overline{\mathbf{x}}^k$ is given in (3.2). The first-order optimality conditions for $\overline{\mathbf{x}}_i^k$ can be written

$$\overline{\mathbf{x}}_i^k = \text{ prox }_{h_i}\Big(\overline{\mathbf{x}}_i^k - \nabla f_i(\overline{\mathbf{x}}_i^k) - \rho\mathbf{A}_i^\mathsf{T}(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k + \lambda^k/\rho) - \rho\mathbf{Q}_i(\overline{\mathbf{x}}_i^k - \mathbf{y}_i^k)\Big).$$

Using this formula for the first $\overline{\mathbf{x}}_i^k$ on the right side of the identity

$$e_i(\overline{\mathbf{x}}^k, \lambda) = \|\overline{\mathbf{x}}_i^k - \text{ prox }_{h_i}(\overline{\mathbf{x}}_i^k - \nabla f_i(\overline{\mathbf{x}}_i^k) - \mathbf{A}_i^\mathsf{T}\lambda)\|,$$

along with the nonexpansive property of prox operator, we have

$$e_i(\overline{\mathbf{x}}^k, \lambda^k) \le \rho\Big(\|\mathbf{A}_i^\mathsf{T}(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k)\| + \|\mathbf{Q}_i(\overline{\mathbf{x}}_i^k - \mathbf{y}_i^k)\|\Big).$$

The definition of $\mathbf{b}_i^k$ yields

$$\begin{aligned}\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k &= \sum_{j<i}\mathbf{A}_j\mathbf{z}_j^k + \sum_{j\ge i}\mathbf{A}_j\mathbf{y}_j^k - \mathbf{b}\\ &= \mathbf{A}\mathbf{z}^k - \mathbf{b} + \sum_{j\ge i}\mathbf{A}_j(\mathbf{y}_j^k - \mathbf{z}_j^k)\\ &= \mathbf{r}_k + \sum_{j\ge i}\mathbf{A}_j(\mathbf{y}_j^k - \mathbf{z}_j^k).\end{aligned}$$

It follows that

$$\|\mathbf{A}_i^\mathsf{T}(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k)\| \le c(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\|), \tag{5.7}$$

and

$$e_i(\overline{\mathbf{x}}^k, \lambda^k) \le c(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \|\overline{\mathbf{x}}_i^k - \mathbf{z}_i^k\|). \tag{5.8}$$

Combining this with (5.6) gives

$$e_i(\mathbf{z}^k, \lambda^k) \le c(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \|\overline{\mathbf{x}}_i^k - \mathbf{z}_i^k\|).$$

Now, by Lemma 3.1, we have

$$\begin{aligned}\sqrt{\rho\nu_i}\|\mathbf{z}_i^k - \overline{\mathbf{x}}_i^k\| &\le \frac{\|\mathbf{x}_i^k - \overline{\mathbf{x}}_i^k\|}{\sqrt{\Gamma_i^k}}\\ &\le \frac{\|\mathbf{x}_i^k - \mathbf{z}_i^k\| + \|\mathbf{z}_i^k - \overline{\mathbf{x}}_i^k\|}{\sqrt{\Gamma_i^k}}.\end{aligned} \tag{5.9}$$

The stopping condition in Step 1b gives

$$\frac{\|\mathbf{x}_i^k - \mathbf{z}_i^k\|}{\sqrt{\Gamma_i^k}} \leq \psi(\epsilon^{k-1}) \leq c\epsilon^{k-1}. \tag{5.10}$$

Hence, by (5.9) we have

$$\left(\frac{-1 + \sqrt{\Gamma_i^k \rho \nu_i}}{\sqrt{\Gamma_i^k}}\right)\|\mathbf{z}_i^k - \overline{\mathbf{x}}_i^k\| \leq \frac{\|\mathbf{x}_i^k - \mathbf{z}_i^k\|}{\sqrt{\Gamma_i^k}} \leq c\epsilon^{k-1}.$$

Therefore, the Case 1 condition $\Gamma_i^k > 4/(\rho\nu_i)$ implies that

$$\|\mathbf{z}_i^k - \overline{\mathbf{x}}_i^k\| \leq c\epsilon^{k-1},$$

and by (5.8), we have

$$e_i(\mathbf{z}^k, \boldsymbol{\lambda}^k) \leq c(\epsilon^{k-1} + \|\mathbf{y}^k - \mathbf{z}^k\| + \|\mathbf{r}_k\|). \tag{5.11}$$

Case 2 $\Gamma_i^k \leq 4/(\rho\nu_i)$. It is shown in [21, pp. 227–228] that when the parameters δ^l and α^l are chosen according to either (2.4) or (2.5), there exists a constant $\Theta > 0$, independent of k and l, such that $\gamma^l \geq l^2\Theta$. Since the γ^l are increasing functions of l and Γ_i^k is the final value of γ^l in Step 1, it follows from the uniform bound on Γ_i^k in Case 2, and the quadratic growth in γ^l, that the final l value in Step 1, which we denote l_i^k, is uniformly bounded as a function of i and k. Also, it follows from the quadratic growth of γ^l and equations (5.18) and (5.20) in [21] that δ^l is uniformly (in k, l, and i) bounded.

By the definition of γ^l in Algorithm 2.2, we have $(1-\alpha^l)\gamma^l = \gamma^{l-1}$, or equivalently, $\alpha^l\gamma^l = \gamma^l - \gamma^{l-1}$ (with the convention that $\gamma^0 = 0$). Summing this identity over l yields

$$\gamma^l = \sum_{j=1}^{l} \alpha^j\gamma^j. \tag{5.12}$$

Next, we multiply the definition $\mathbf{a}_{ik}^j = (1-\alpha^j)\mathbf{a}_{ik}^{j-1} + \alpha^j\mathbf{u}_{ik}^j$ by γ^j and sum over j between 1 and l. Again, exploiting the identity $(1-\alpha^j)\gamma^j = \gamma^{j-1}$ yields

$$\mathbf{a}_{ik}^l = \frac{1}{\gamma^l}\sum_{j=1}^{l}(\gamma^j\alpha^j)\mathbf{u}_{ik}^j. \tag{5.13}$$

It follows from (5.12), that $\mathbf{a}_{ik}^l$ is a convex combination of $\mathbf{u}_{ik}^j$, $1 \leq j \leq l$. If $p_{ik}^j \in [0, 1]$ denotes the coefficients in the convex combination, we have

$$\mathbf{a}_{ik}^l = \sum_{j=1}^{l} p_{ik}^j\mathbf{u}_{ik}^j, \tag{5.14}$$

Since $\mathbf{z}_i^k = \mathbf{a}_{ik}^L$ for $L = l_i^k$, Jensen's inequality gives

$$
\begin{aligned}
e_i(\mathbf{z}^k, \lambda^k) &\le \sum_{l=1}^{l_i^k} p_{ik}^l \|\mathbf{u}_{ik}^l - \operatorname{prox}_{h_i}(\mathbf{z}_i^k - \nabla f_i(\mathbf{z}_i^k) - \mathbf{A}_i^\top \lambda^k)\| \\
&\le \sum_{l=1}^{l_i^k} \|\mathbf{u}_{ik}^l - \operatorname{prox}_{h_i}(\mathbf{z}_i^k - \nabla f_i(\mathbf{z}_i^k) - \mathbf{A}_i^\top \lambda^k)\|.
\end{aligned}
\tag{5.15}
$$

Now, by the formula for $\mathbf{u}_{ik}^l$ in Alg. 2.2, we have $\mathbf{u}_{ik}^l = \operatorname{prox}_{h_i}(\mathbf{q}_2)$, where

$$
\mathbf{q}_2 = \mathbf{u}_{ik}^l - \nabla f_i(\bar{\mathbf{a}}_{ik}^l) - \delta_{ik}^l(\mathbf{u}_{ik}^l - \mathbf{u}_{ik}^{l-1}) - \rho\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k + \lambda^k/\rho) - \rho\mathbf{Q}_i(\mathbf{u}_{ik}^l - \mathbf{y}_i^k).
$$

We utilize (5.4) with $\mathbf{q}_1 = \mathbf{z}_i^k - \nabla f_i(\mathbf{z}_i^k) - \mathbf{A}_i^\top \lambda^k$, with $\mathbf{q}_2$ as given above, and with $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{u}_{ik}^l$. Hence, $\mathbf{p}_2 - \operatorname{prox}_{h_i}(\mathbf{q}_2) = \mathbf{0}$ and by (5.4), it follows that

$$
\begin{aligned}
&\|\mathbf{u}_{ik}^l - \operatorname{prox}_{h_i}(\mathbf{z}_i^k - \nabla f_i(\mathbf{z}_i^k) - \mathbf{A}_i^\top \lambda^k)\| \\
&\quad \le c\Big(\|\mathbf{u}_{ik}^l - \mathbf{z}_i^k\| + \|\bar{\mathbf{a}}_{ik}^l - \mathbf{z}_i^k\| + \|\mathbf{u}_{ik}^l - \mathbf{u}_{ik}^{l-1}\| + \|\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k)\| + \|\mathbf{u}_{ik}^l - \mathbf{y}_i^k\|\Big) \\
&\quad \le c\Big(\|\mathbf{u}_{ik}^l - \mathbf{z}_i^k\| + \|\bar{\mathbf{a}}_{ik}^l - \mathbf{z}_i^k\| + \|\mathbf{u}_{ik}^l - \mathbf{u}_{ik}^{l-1}\| + \|\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k)\| + \|\mathbf{y}_i^k - \mathbf{z}_i^k\|\Big)
\end{aligned}
\tag{5.16}
$$

Each of the terms on the right side of (5.16) is now analyzed.

Based on (5.7), the trailing two terms in (5.16) have the bound

$$
\|\mathbf{A}_i^\top(\mathbf{A}_i\mathbf{y}_i^k - \mathbf{b}_i^k)\| + \|\mathbf{y}_i^k - \mathbf{z}_i^k\| \le c(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\|).
$$

The remaining terms in (5.16) are bounded by $c\sqrt{r_i^k}$ as will now be shown. The bound $\|\mathbf{u}_{ik}^l - \mathbf{u}_{ik}^{l-1}\| \le c\sqrt{r_i^k}$ is a trivial consequence of the definition of r_i^k and the uniform bound on Γ_i^k in Case 2. By the definition $\bar{\mathbf{a}}_{ik}^l = (1-\alpha^l)(\mathbf{a}_{ik}^{l-1} - \mathbf{u}_{ik}^{l-1}) + \mathbf{u}_{ik}^{l-1}$, it follows that

$$
\|\bar{\mathbf{a}}_{ik}^l - \mathbf{z}_i^k\| \le \|\mathbf{a}_{ik}^{l-1} - \mathbf{u}_{ik}^{l-1}\| + \|\mathbf{u}_{ik}^{l-1} - \mathbf{z}_i^k\|.
$$

This inequality and the fact that $\mathbf{z}_i^k = \mathbf{a}_{ik}^l$ for $l = l_i^k$ implies that all the remaining terms in (5.16) have the form $\|\mathbf{a}_{ik}^l - \mathbf{u}_{ik}^t\|$ for some $l \in [1, l_i^k]$ and some $t \in [1, l]$. Combine (5.14), Jensen's inequality, the fact that $l \le l_i^k$ where l_i^k is uniformly bounded in Case 2, and the Schwarz inequality to obtain

$$
\|\mathbf{a}_{ik}^l - \mathbf{u}_{ik}^t\| \le \sum_{j=1}^{l} \left\|\mathbf{u}_{ik}^j - \mathbf{u}_{ik}^t\right\| \le l \sum_{j=1}^{l} \left\|\mathbf{u}_{ik}^j - \mathbf{u}_{ik}^{j-1}\right\| \le c\sqrt{r_i^k},
$$

These bounds for the terms in (5.16) combine to yield

$$
\|\mathbf{u}_{ik}^l - \operatorname{prox}_{h_i}(\mathbf{z}_i^k - \nabla f_i(\mathbf{z}_i^k) - \mathbf{A}_i^\top \lambda^k)\| \le c\left(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k}\right).
$$

Moreover, by (5.15) and the Case 2 uniform bound on l_i^k, we have

$$e_i(\mathbf{z}^k, \lambda^k) \le c\left(\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k} \right).$$

Combine this with the Case 1 lower bound (5.11) gives

$$e_i(\mathbf{z}^k, \lambda^k) \le c\left(\epsilon^{k-1} + \|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k} \right). \tag{5.17}$$

Inserting this in (5.5) yields

$$e_i(\mathbf{y}^{k+1}, \lambda^{k+1}) \le c\left(\epsilon^{k-1} + \|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k} + \|\mathbf{y}^{k+1} - \mathbf{y}^k\| \right).$$

Based on the back substitution formula $\mathbf{y}^{k+1} - \mathbf{y}^k = \alpha \mathbf{M}^{-\mathsf{T}}\mathbf{Q}(\mathbf{z}^k - \mathbf{y}^k)$, this reduces to

$$e_i(\mathbf{y}^{k+1}, \lambda^{k+1}) \le c\left(\epsilon^{k-1} + \|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k} \right).$$

Since $\epsilon^{k-1} \le cd_{k-1}$ and $\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{r_i^k} \le d_k$, the proof is complete. □

The expression E_k defined in (3.3) measures the energy between the current iterate $(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)$ and a given $(\mathbf{x}^*, \mathbf{x}^*, \lambda^*)$. Let E_k^* denote the minimum energy between the iterate and all possible $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$. We will show that when an error bound condition holds, there exists a constant $\kappa < 1$ such that $E_{k+2}^* \le \kappa E_k^*$.

The error bound condition relates the KKT error to the Euclidean distance to $\mathcal{W}^*$. The KKT error K is given by

$$K(\mathbf{x}, \lambda) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\| + \sum_{i=1}^{m} e_i(\mathbf{x}, \lambda). \tag{5.18}$$

When $K(\mathbf{x}, \lambda) = 0$, the first-order optimality conditions hold. The Euclidean distance from $(\mathbf{x}, \lambda)$ to $\mathcal{W}^*$ will be measured by

$$\mathcal{E}(\mathbf{x}, \lambda) = \min\left\{ \rho\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda - \lambda^*\|^2 : (\mathbf{x}^*, \lambda^*) \in \mathcal{W}^* \right\}^{1/2}. \tag{5.19}$$

Note that $\mathbf{P} = \mathbf{M}\mathbf{Q}^{-1}\mathbf{M}^{\mathsf{T}}$ is positive definite since $\mathbf{M}$ is invertible. Also, by [1, Prop. 6.1.2], every solution of (1.1) has exactly the same set of Lagrange multipliers. If $\mathbf{X}^*$ and $\boldsymbol{\Lambda}^*$ denote the set of solutions and multipliers for (1.1), then $\mathcal{W}^* = \mathbf{X}^* \times \boldsymbol{\Lambda}^*$ is a closed, convex set, and there exists a unique $(\tilde{\mathbf{x}}, \tilde{\lambda}) \in \mathcal{W}^*$ that achieves the minimum in (5.19). The local error bound assumption is as follows:

Assumption 5.1 There exist constants $\beta > 0$ and $\eta > 0$ such that $\mathcal{E}(\mathbf{x}, \lambda) \le \eta K(\mathbf{x}, \lambda)$ whenever $\mathcal{E}(\mathbf{x}, \lambda) \le \beta$.

The local error bound condition is equivalent to saying that in a neighborhood of $\mathcal{W}^*$, the Euclidean distance to $\mathcal{W}^*$ is bound by the KKT error, which is often used to analyze the linear convergence behavior of an optimization algorithm. More recently, a partial error bound condition based on the ADMM iterates instead of conditions on the optimization problem is proposed in [34]. Under such conditions, linear convergence is also established for a 2-block ADMM.

A multivalued mapping F is piecewise polyhedral if its graph $\mathrm{Gph}\, F := \{(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in F(\mathbf{x})\}$ is a union of finitely many polyhedral sets. The local error bound condition (Assumption 5.1) holds when ∇f_i is affine and ∂h_i is piecewise polyhedral for $i = 1, \ldots, m$ [23, 36, 42]. Note that when $(\mathbf{x}, \lambda)$ is restricted to a bounded set, the requirement that $\mathcal{E}(\mathbf{x}, \lambda) \le \beta$ can be dropped. That is, when $\mathcal{E}(\mathbf{x}, \lambda) > \beta$, $K(\mathbf{x}, \lambda)$ is strictly positive, and by taking the constant η large enough, the bound $\mathcal{E}(\mathbf{x}, \lambda) \le \eta K(\mathbf{x}, \lambda)$ holds over the entire set. In our analysis, the error bound condition is applied to the iterates $(\mathbf{y}^k, \lambda^k)$ which lie in a bounded set by Lemma 3.2, so the requirement that $\mathcal{E}(\mathbf{x}, \lambda) \le \beta$ is unnecessary.

Theorem 5.2 *If the parameters δ^l and α^l in Algorithm* 2.2 *are chosen according to either* (2.4) *or* (2.5), $\psi(t) \le c_\psi t$, *and Assumption* 5.1 *holds, then there exists $\kappa < 1$ such that $E^*_{k+2} \le \kappa E^*_k$ at every iteration of Algorithm* 2.1.

Proof Let $(\tilde{\mathbf{y}}^{k+1}, \tilde{\lambda}^{k+1}) \in \mathcal{W}^*$ be the unique minimizer in (5.19) corresponding to $(\mathbf{x}, \lambda) = (\mathbf{y}^{k+1}, \lambda^{k+1})$. By the stopping condition in Step 1b of Algorithm 2.2, and the definition of Γ_i^k in Step 1c, the sequence Γ_i^k is nondecreasing in k by Remark 2.1. Since Γ_i^k is nondecreasing in k, it follows from the triangle inequality and the back substitution formula $\mathbf{y}^{k+1} - \mathbf{y}^k = \alpha \mathbf{M}^{-\mathsf{T}} \mathbf{Q}(\mathbf{z}^k - \mathbf{y}^k)$ that for any $i \in [1, m]$, we have

$$\begin{aligned}
\frac{\|\mathbf{x}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|}{\sqrt{\Gamma_i^{k+1}}} &\le \frac{\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\| + \|\mathbf{z}_i^k - \mathbf{y}_i^k\| + \|\mathbf{y}_i^k - \mathbf{y}_i^{k+1}\| + \|\mathbf{y}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|}{\sqrt{\Gamma_i^{k+1}}} \\
&\le \frac{\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\|}{\sqrt{\Gamma_i^k}} + \frac{\|\mathbf{z}_i^k - \mathbf{y}_i^k\| + \|\mathbf{y}_i^k - \mathbf{y}_i^{k+1}\| + \|\mathbf{y}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|}{\sqrt{\Gamma_i^1}} \\
&\le \frac{\|\mathbf{x}_i^{k+1} - \mathbf{z}_i^k\|}{\sqrt{\Gamma_i^k}} + c\left(\|\mathbf{z}^k - \mathbf{y}^k\| + \|\mathbf{y}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|\right),
\end{aligned} \tag{5.20}$$

where $c > 0$ denotes a generic constant, independent of k.

As noted earlier, when the parameters δ^l and α^l in Algorithm 2.2 are chosen according to either (2.4) or (2.5), we have $\xi^l = \delta^l \alpha^l \gamma^l = 1$. By equation (3.12) in the supplementary material for this paper with $L = l_i^k$, $\mathbf{u} = \mathbf{a}_i^L = \mathbf{z}_i^k$, $\mathbf{u}_i^L = \mathbf{x}^{k+1}$, and $\mathbf{u}_i^0 = \mathbf{x}_k$, we obtain the relation

$$\frac{\|\mathbf{z}_i^k - \mathbf{x}_i^{k+1}\|}{\sqrt{\Gamma_i^k}} \le \frac{\|\mathbf{z}_i^k - \mathbf{x}_i^k\|}{\sqrt{\Gamma_i^k}} \le \psi(\epsilon^{k-1}),$$

where the last inequality is due to the stopping condition in Step 1b. Combining this with (5.20) yields

$$\frac{\|\mathbf{x}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|}{\sqrt{\Gamma_i^{k+1}}} \le \psi(\epsilon^{k-1}) + c\left(\|\mathbf{z}^k - \mathbf{y}^k\| + \|\mathbf{y}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|\right). \tag{5.21}$$

Exploiting the error bound condition, we have

$$\begin{aligned}\|\mathbf{y}^{k+1} - \tilde{\mathbf{y}}^{k+1}\|^2 \le& \sqrt{\|\mathbf{P}^{-1}\|}\|\mathbf{y}^{k+1} - \tilde{\mathbf{y}}^{k+1}\|_{\mathbf{P}} \\ \le& c\mathcal{E}(\mathbf{y}^{k+1}, \lambda^{k+1}) \le cK(\mathbf{y}^{k+1}, \lambda^{k+1}).\end{aligned} \tag{5.22}$$

The constraint violation term in K is estimated as follows:

$$\|\mathbf{A}\mathbf{y}^{k+1} - \mathbf{b}\| \le \|\mathbf{A}\|(\|\mathbf{y}^{k+1} - \mathbf{y}^k\| + \|\mathbf{y}^k - \mathbf{z}^k\|) + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\| \le cd_k,$$

where the last inequality is due to the back substitution formula and the definition (5.3) of d_k. Hence, Lemma 5.1 yields

$$K(\mathbf{y}^{k+1}, \lambda^{k+1}) \le c(d_k + d_{k-1}). \tag{5.23}$$

Combine (5.21)–(5.23) to obtain

$$\frac{\|\mathbf{x}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|}{\sqrt{\Gamma_i^{k+1}}} \le \psi(\epsilon^{k-1}) + c(d_k + d_{k-1}) \le c(d_k + d_{k-1}) \tag{5.24}$$

since $\psi(t) \le c_\psi t$ and $\epsilon^{k-1} \le cd_{k-1}$. Since the energy E_{k+1}^* corresponds to the minimum of E_{k+1} over all $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ and since $(\tilde{\mathbf{y}}^{k+1}, \tilde{\lambda}^{k+1}) \in \mathcal{W}^*$, it follows that

$$E_{k+1}^* \le \rho\|\mathbf{y}^{k+1} - \tilde{\mathbf{y}}^{k+1}\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^{k+1} - \tilde{\lambda}^{k+1}\|^2 + \alpha \sum_{i=1}^m \frac{\|\mathbf{x}_i^{k+1} - \tilde{\mathbf{y}}_i^{k+1}\|^2}{\Gamma_i^{k+1}}.$$

The first two terms on the right are $\mathcal{E}^2(\mathbf{y}^{k+1}, \lambda^{k+1})$, while the last term in bounded by (5.24). We have

$$E_{k+1}^* \le \mathcal{E}^2(\mathbf{y}^{k+1}, \lambda^{k+1}) + c\left(d_k + d_{k-1}\right)^2.$$

Combine this with the error bound condition and (5.23) gives

$$E_{k+1}^* \le c\left(d_k + d_{k-1}\right)^2. \tag{5.25}$$

Suppose that $(\hat{\mathbf{x}}^k, \hat{\lambda}^k) \in \mathcal{W}^*$ is the unique minimizing $(\mathbf{x}^*, \lambda^*) \in \mathcal{W}^*$ associated with E_k^*. By Lemma 3.2 and the fact that $(\hat{\mathbf{x}}^k, \hat{\lambda}^k) \in \mathcal{W}^*$, we have

$$
\begin{aligned}
E_k^* \geq & \rho\|\mathbf{y}^{k+1} - \hat{\mathbf{x}}^k\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^{k+1} - \hat{\lambda}^k\|^2 + \alpha \sum_{i=1}^m \frac{\|\mathbf{x}_i^{k+1} - \hat{\mathbf{x}}_i^k\|^2}{\Gamma_i^k} \\
& + \rho\alpha(1-\alpha)(\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{Q}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2) + \sigma\alpha \sum_{i=1}^m R^k.
\end{aligned}
$$

The first three terms on the right side are bounded from below by E_{k+1}^*, while the last three terms are bounded from below by cd_k^2 by the definition of d_k in (5.3). Hence,

$$
E_k^* \geq E_{k+1}^* + cd_k^2. \tag{5.26}
$$

We replace k by $k-1$ and then use again (5.26) followed by (5.25) to obtain

$$
E_{k-1}^* \geq E_k^* + cd_{k-1}^2 \geq E_{k+1}^* + c(d_k^2 + d_{k-1}^2) \geq (1+c)E_{k+1}^*,
$$

which completes the proof. □

Another linear convergence result is established when the objective Φ is strongly convex, in which case the solution $\mathbf{x}^*$ of (1.1) is unique. Our assumption is the following:

Assumption 5.2 The objective Φ is strongly convex with modulus $\mu > 0$ and there exist constants $\beta > 0$ and $\eta > 0$ such that

$$
\|\lambda - \tilde{\lambda}\| \leq \eta \sum_{i=1}^m \|e_i(\mathbf{x}^*, \lambda)\| \tag{5.27}
$$

whenever $\|\lambda - \tilde{\lambda}\| \leq \beta$.

The local error bound condition (5.27) holds when ∂h_i is piecewise polyhedral for $i = 1, \ldots, m$ [23, 36, 42]. Similar to the comment before Theorem 5.2, the requirement that $\|\lambda - \tilde{\lambda}\| \leq \beta$ can be dropped since it is applied to the iterates λ^k which lie in a bounded set by Lemma 3.2.

Theorem 5.3 *If the parameters δ^l and α^l in Algorithm* 2.2 *are chosen according to either* (2.4) *or* (2.5), *$\psi(t) \leq c_\psi t$, and Assumption* 5.2 *holds, then there exists $\kappa < 1$ such that $E_{k+2}^* \leq \kappa E_k^*$ at every iteration of Algorithm* 2.1.

Proof By the local error bound condition and by (5.4) with $\mathbf{p}_1 -$ prox $_{h_i}(\mathbf{q}_1)$ identified with $e_i(\mathbf{x}^*, \lambda^{k+1})$ and $\mathbf{p}_2 -$ prox $_{h_i}(\mathbf{q}_2)$ identified with $e_i(\mathbf{z}^k, \lambda^k)$, we have

$$
\begin{aligned}
\|\lambda^{k+1} - \tilde{\lambda}^{k+1}\| \leq & \eta \sum_{i=1}^m e_i(\mathbf{x}^*, \lambda^{k+1}) \\
\leq & c\left(\|\mathbf{z}^k - \mathbf{x}^*\| + \|\lambda^{k+1} - \lambda^k\| + \sum_{i=1}^m e_i(\mathbf{z}^k, \lambda^k) \right),
\end{aligned} \tag{5.28}
$$

where $c > 0$ is a constant. In the later proof, we again use $c > 0$ as a generic constant. By (5.17), it follows that

$$\sum_{i=1}^{m} e_i(\mathbf{z}^k, \lambda^k) \le c\Big(\epsilon^{k-1} + \|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{R^k}\Big).$$

Inserting this in (5.28) and recalling that $\lambda^{k+1} - \lambda^k = \alpha\rho(\mathbf{A}\mathbf{z}^k - \mathbf{b}) = \alpha\rho\mathbf{r}_k$, we have

$$\|\lambda^{k+1} - \tilde{\lambda}^{k+1}\| \le c\Big(\epsilon^{k-1} + \|\mathbf{z}^k - \mathbf{x}^*\| + \|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{R^k}\Big).$$

Since $\epsilon^{k-1} \le cd_{k-1}$ and $\|\mathbf{r}_k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \sqrt{R^k} \le d_k$, it follows that

$$\|\lambda^{k+1} - \tilde{\lambda}^{k+1}\| \le c(d_k + d_{k-1} + \|\mathbf{z}^k - \mathbf{x}^*\|). \tag{5.29}$$

By (5.21) with $\tilde{\mathbf{y}}^{k+1} = \mathbf{x}^*$, we have

$$\frac{\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|}{\sqrt{\Gamma_i^{k+1}}} \le c\big(\epsilon^{k-1} + \|\mathbf{z}^k - \mathbf{y}^k\| + \|\mathbf{y}^{k+1} - \mathbf{x}^*\|\big). \tag{5.30}$$

The triangle inequality and the back substitution formula yield

$$\begin{aligned} \|\mathbf{y}^{k+1} - \mathbf{x}^*\| &\le \|\mathbf{y}^{k+1} - \mathbf{y}^k\| + \|\mathbf{y}^k - \mathbf{z}^k\| + \|\mathbf{z}^k - \mathbf{x}^*\| \\ &\le c\|\mathbf{y}^k - \mathbf{z}^k\| + \|\mathbf{z}^k - \mathbf{x}^*\|. \end{aligned} \tag{5.31}$$

The bounds $\epsilon^{k-1} \le cd_{k-1}$ and $\|\mathbf{y}^k - \mathbf{z}^k\| \le d_k$ in (5.31) and (5.30) give

$$\|\mathbf{y}^{k+1} - \mathbf{x}^*\| \le cd_k + \|\mathbf{z}^k - \mathbf{x}^*\| \text{ and } \frac{\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|}{\sqrt{\Gamma_i^{k+1}}} \le c\big(d_{k-1} + d_k + \|\mathbf{z}^k - \mathbf{x}^*\|\big). \tag{5.32}$$

Combine (5.29) and (5.32) to obtain

$$\begin{aligned} E_{k+1}^* &= \rho\|\mathbf{y}^{k+1} - \mathbf{x}^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^{k+1} - \tilde{\lambda}^{k+1}\|^2 + \alpha\sum_{i=1}^{m}\frac{\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2}{\Gamma_i^{k+1}} \\ &\le c(d_k + d_{k-1} + \|\mathbf{z}^k - \mathbf{x}^*\|)^2. \end{aligned} \tag{5.33}$$

On the other hand, by Lemma 3.2 and the fact that $(\mathbf{x}^*, \tilde{\lambda}^k) \in \mathcal{W}^*$, we have

$$\begin{aligned} E_k^* &\ge \rho\|\mathbf{y}^{k+1} - \mathbf{x}^*\|_{\mathbf{P}}^2 + \frac{1}{\rho}\|\lambda^{k+1} - \tilde{\lambda}^k\|^2 + \alpha\sum_{i=1}^{m}\frac{\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^*\|^2}{\Gamma_i^k} \\ &\quad + \rho\alpha(1-\alpha)(\|\mathbf{y}^k - \mathbf{z}^k\|_{\mathbf{Q}}^2 + \|\mathbf{A}\mathbf{z}^k - \mathbf{b}\|^2) + \sigma\alpha R^k + 2\alpha\Delta^k \\ &\ge E_{k+1}^* + cd_k^2 + \mu\|\mathbf{z}^k - \mathbf{x}^*\|^2, \end{aligned} \tag{5.34}$$

where the last inequality is due to the definition (5.3) of d_k and the strong convexity of Φ:

$$\Delta^k := \Phi(\mathbf{z}^k) - \Phi(\mathbf{x}^*) + (\tilde{\lambda}^k, \mathbf{A}\mathbf{z}^k - \mathbf{b}) \geq \frac{\mu}{2}\|\mathbf{z}^k - \mathbf{x}^*\|^2.$$

Finally, we replace k by $k-1$ in (5.34), and then use again (5.34) followed by (5.33) to obtain

$$E_{k-1}^* \geq E_k^* + cd_{k-1}^2 \geq E_{k+1}^* + c(d_k^2 + d_{k-1}^2) + \mu\|\mathbf{z}^k - \mathbf{x}^*\|^2 \geq (1+c)E_{k+1}^*,$$

which completes the proof. □

6 Numerical experiments

In this section, we compare the performance of I-ADMM to that of two different algorithms: (a) linearized ADMM with one linearization step for each subproblem and (b) exact ADMM where the subproblems are solved either by the conjugate gradient method or by an explicit formula. The conjugate gradient method was well suited for the quadratic subproblems in our test set. We tried using a small number of conjugate gradient iterations to solve a subproblem, such as 5 iterations starting from the solution computed in the previous iteration, but found that the scheme did not converge. Instead we continued the CG iteration until the norm of the gradient was at most 10^{-6}. The one-step ADMM algorithm that we used in (a) for the experiments was the generalized BOSVS algorithm from [21]. This algorithm is globally convergent, and although the penalty term was not linearized, it was possible to quickly solve the subproblems that arise in the imaging test problems using a fast Fourier transform, as explained in [10].

The problems in our experiments were the same image reconstruction problems used in [21]. One image employs a blurred version of the well-known Cameraman image of size 256×256, while the second set of test problems, which arise in partially parallel imaging (PPI), are found in [10]. The observed PPI data, corresponding to 3 different images, are denoted data 1, data 2, and data 3. These image reconstruction problem can be formulated as

$$\min_{\mathbf{u}} \ \frac{1}{2}\|\mathbf{F}\mathbf{u} - \mathbf{f}\|^2 + \alpha\|\mathbf{u}\|_{TV} + \beta\|\boldsymbol{\Psi}^{\mathsf{T}}\mathbf{u}\|_1, \tag{6.1}$$

where $\mathbf{f}$ is the given image data, $\mathbf{F}$ is a matrix describing the imaging device, $\|\cdot\|_{TV}$ is the total variation norm, $\|\cdot\|_1$ is the ℓ_1 norm, $\boldsymbol{\Psi}$ is a wavelet transform, and $\alpha > 0$ and $\beta > 0$ are weights. The first term in the objective is the data fidelity term, while the next two terms are for regularization; they are designed to enhance edges and increase image sparsity. In our experiments, $\boldsymbol{\Psi}$ is a normalized Haar wavelet with four levels and $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathsf{T}} = I$. The problem (6.1) is equivalent to

$$\min_{(\mathbf{u},\mathbf{v},\mathbf{w})} \ \frac{1}{2}\|\mathbf{F}\mathbf{u} - \mathbf{f}\|^2 + \alpha\|\mathbf{w}\|_{1,2} + \beta\|\mathbf{v}\|_1 \ \text{ subject to } \mathbf{B}\mathbf{u} = \mathbf{w}, \ \boldsymbol{\Psi}^{\mathsf{T}}\mathbf{u} = \mathbf{v}, \tag{6.2}$$

where $\mathbf{Bu} = \nabla \mathbf{u}$ and $(\nabla \mathbf{u})_i$ is the vector of finite differences in the image along the coordinate directions at the i-th pixel in the image, $\|\mathbf{w}\|_{1,2} = \sum_{i=1}^{N} \|(\nabla \mathbf{u})_i\|_2$, and N is the total number of pixels in the image.

The problem (6.2) has the structure appearing in (1.1)–(1.2) with $h_1 := 0$, $f_1(\mathbf{u}) = 1/2\|\mathbf{Fu} - \mathbf{f}\|^2$, $h_2(\mathbf{w}) = \|\mathbf{w}\|_{1,2}$, $f_2 := 0$, $h_3(\mathbf{v}) = \|\mathbf{v}\|_1$, $f_3 := 0$,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{B} \\ \boldsymbol{\Psi}^\top \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} -\mathbf{I} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{A}_3 = \begin{pmatrix} \mathbf{0} \\ -\mathbf{I} \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

The algorithm parameters α^l and δ^l were chosen as in (2.5). Since $f_2 = f_3 = 0$, the second and third subproblems are solved in closed form, due to the simple structure of h_2 and h_3. Only the first subproblem is solved inexactly. At iteration k, the solution of this subproblem approximates the solution of

$$\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{Fu} - \mathbf{f}\|^2 + \frac{\rho}{2}\|\mathbf{Bu} - \mathbf{w}^k + \rho^{-1}\boldsymbol{\lambda}^k\|^2 + \frac{\rho}{2}\|\boldsymbol{\Psi}^\top\mathbf{u} - \mathbf{v}^k + \rho^{-1}\boldsymbol{\mu}^k\|^2,$$

where $\boldsymbol{\lambda}^k$ and $\boldsymbol{\mu}^k$ are the Lagrange multipliers at iteration k for the constraints $\mathbf{Bu} = \mathbf{w}$ and $\boldsymbol{\Psi}^\top\mathbf{u} = \mathbf{v}$ respectively. Details of the experimental setup can be found in [21]. The i-th block diagonal element of $\mathbf{Q}$ was taken to be a multiple γ_i of the identity $\mathbf{I}$. According to the assumptions of IADM, γ_1 should be chosen large enough that $\gamma_1\mathbf{I} - \mathbf{A}_1^\top\mathbf{A}_1$ is positive semidefinite, where

$$\mathbf{A}_1^\top\mathbf{A}_1 = \mathbf{B}^\top\mathbf{B} + \boldsymbol{\Psi}\boldsymbol{\Psi}^\top.$$

However, a closer inspection of the global convergence proof reveals that for convergence, it is sufficient to have

$$\gamma_1\|\mathbf{z}^k - \mathbf{y}^k\|^2 \geq \|\mathbf{A}_1(\mathbf{z}^k - \mathbf{y}^k)\|^2 \tag{6.3}$$

in each iteration. Instead of computing the largest eigenvalue of $\mathbf{A}_1^\top\mathbf{A}_1$, we simply start with $\gamma_1 = 4$ and multiply it by a constant factor (3 in the experiments) whenever the inequality (6.3) is violated. Within a finite number of iterations, γ_1 is large enough that (6.3) always holds.

Figure 1 plots the logarithm of the relative objective error versus the CPU time for the four test problems and the three methods. Note that the first few iterations of the exact ADMM for Data 3 have error greater than one, so they missing from the plot. Observe that I-ADMM performed better than the exact ADMM and the exact ADMM was generally better than the single linearization step, except possibly in the initial iterations where the high accuracy of the exact ADMM was not helpful. I-ADMM gave better performance both initially and asymptotically.

7 Conclusion

We propose an inexact alternating direction method of multipliers, I-ADMM, for solving separable convex linearly constrained optimization problems, where the objective is the sum of smooth and relatively simple nonsmooth terms. The

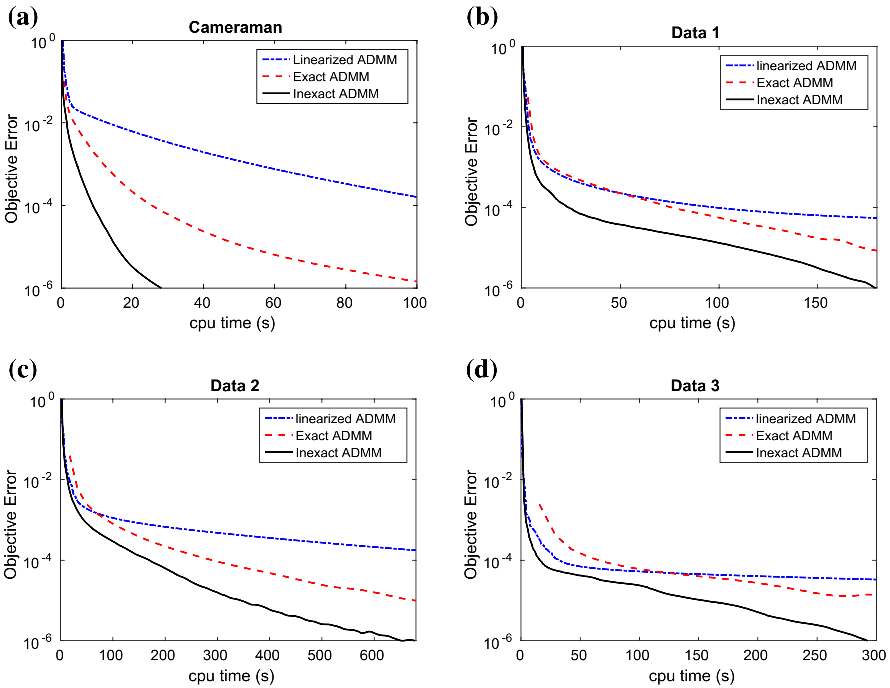

Fig. 1 Base-10 logarithm of the relative objective error versus CPU time for the test problems

nonsmooth terms could be infinite, so the algorithms and analysis include problems with additional convex constraints. This I-ADMM emanates for our earlier work [10, 20, 21] on a Bregman Operator Splitting algorithm with a variable stepsize (BOSVS). The subproblems are solved using an accelerated gradient algorithm that employs a linearization of both the smooth objective and the penalty term. We establish an $\mathcal{O}(1/k)$ ergodic convergence rate for I-ADMM, where k is the iteration number. Under a strong convexity assumption, the convergence rate improves to $\mathcal{O}(1/k^2)$ for both ergodic and nonergodic iterates. When an error bound condition holds, 2-step linear convergence is established for nonergodic iterates. The convergence rates for I-ADMM are consistent with convergence rates obtained for exact ADMM schemes such as those in [23, 28, 30, 35, 38, 42]. As observed in the numerical experiments, an advantage of the inexact scheme is that the computing time to achieve a given error tolerance is reduced, when compared to the the exact iteration, since the accuracy of the subproblem solutions are adaptively increased as the iterates converge so as to achieve the same convergence rates as the exact algorithms.

References

1. Bertsekas, D.P.: Convex Analysis and Optimization. Athena Scientific, Belmont (2003)

2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Mach. Learn. **3**, 1–122 (2010)
3. Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**, 39–73 (2017)
4. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**, 57–79 (2016)
5. Chen, C., Li, M., Liu, X., Ye, Y.: On the convergence of multi-block alternating direction method of multipliers and block coordinate descent method (2015). arXiv:1508.00193
6. Chen, C., Shen, Y., You, Y.: On the convergence analysis of the alternating direction method of multipliers with three blocks. Abstr. Appl. Anal. (2013). https://doi.org/10.1155/2013/183961
7. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**, 81–101 (1994)
8. Chen, J.W., Wang, Y.Y., He, H.J., Lv, Y.B.: Convergence analysis of positive-indefinite proximal ADMM with a Glowinski's relaxation factor. Numer. Algorithms (2019). https://doi.org/10.1007/s11075-019-00731-9
9. Chen, L., Sun, D., Toh, K.: An efficient inexact symmetric Gauss–Seidel based majorized ADMM for high-dimensional convex composite conic programming. Math. Program. **161**, 237–270 (2017)
10. Chen, Y., Hager, W.W., Yashtini, M., Ye, X., Zhang, H.: Bregman operator splitting with variable stepsize for total variation image reconstruction. Comput. Optim. Appl. **54**, 317–342 (2013)
11. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set Valued Var. Anal. **25**, 829–858 (2017)
12. Eckstein, J., Bertsekas, D.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
13. Eckstein, J., Silva, P.J.S.: A practical relative error criterion for augmented Lagrangians. Math. Program. **141**, 319–348 (2013)
14. Eckstein, J., Yao, W.: Approximate ADMM algorithms derived from Lagrangian splitting. Comput. Optim. Appl. **68**, 363–405 (2017)
15. Eckstein, J., Yao, W.: Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. Math. Program. **170**, 417–444 (2018)
16. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
17. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
18. Goldfarb, D., Ma, S.: Fast multiple-splitting algorithms for convex optimization. SIAM J. Optim. **22**, 533–556 (2012)
19. Gol'shtein, E., Tret'yakov, N.: Modified Lagrangians in convex programming and their generalizations. In: Huard, P. (ed.) Point-to-Set Maps and Mathematical Programming. Mathematical Programming Studies, vol. 10, pp. 86–97. Springer, Berlin (1979)
20. Hager, W.W., Yashtini, M., Zhang, H.: An $O(1/k)$ convergence rate for the variable stepsize Bregman operator splitting algorithm. SIAM J. Numer. Anal. **54**, 1535–1556 (2016)
21. Hager, W.W., Zhang, H.: Inexact alternating direction methods of multipliers for separable convex optimization. Comput. Optim. Appl. **73**, 201–235 (2019)
22. Hager, W.W., Zhang, H.: Convergence rates for an inexact ADMM applied to separable convex optimization (2020). arXiv:2001.02503
23. Han, D., Sun, D., Zhang, L.: Linear rate convergence of the alternating direction method of multipliers for convex composite programming. Math. Oper. Res. **43**, 622–637 (2018)
24. Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**, 227–238 (2012)
25. He, B., Liao, L., Han, D., Yan, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**, 103–118 (2002)
26. He, B., Tao, M., Xu, M., Yuan, X.: An alternating direction-based contraction method for linearly constrained separable convex programming problems. Optimization **62**, 573–596 (2013)
27. He, B., Tao, M., Yuan, X.: Alternating direction method with Gaussian back substitution for separable convex programming. SIAM J. Optim. **22**, 313–340 (2012)
28. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
29. He, B.S., Ma, F., Yuan, X.M.: Optimally linearizing the alternating direction method of multipliers for convex programming. Comput. Optim. Appl. **75**, 361–388 (2020)

30. Hong, M., Luo, Z.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**, 165–199 (2017)
31. Li, M., Liao, L., Yuan, X.: Inexact alternating direction methods of multipliers with logarithmic–quadratic proximal regularization. J. Optim. Theory Appl. **159**, 412–436 (2013)
32. Li, M., Sun, D., Toh, K.C.: A convergent 3-block semi-proximal ADMM for for convex minimization problems with one strongly convex block. Asia Pac. J. Oper. Res. **32**, 1–19 (2015)
33. Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multiblock variables. SIAM J. Optim. **25**, 1478–1497 (2015)
34. Liu, Y., Yuan, X., Zeng, S., Zhang, J.: Partial error bound conditions for the linear convergence rate of the alternating direction method of multipliers. SIAM J. Numer. Anal. **56**, 2095–2123 (2018)
35. Monteiro, R.D.C., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**, 475–507 (2013)
36. Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Math. Program. Study **14**, 206–214 (1981)
37. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control **14**, 877–898 (1976)
38. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**, 269–297 (2014)
39. Solodov, M.V., Svaiter, B.F.: An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. Math. Oper. Res. **25**, 214–230 (2000)
40. Tao, M., Yuan, X.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**, 57–81 (2011)
41. Wen, Z., Goldfarb, D., Yin, W.: Alternating direction augmented Lagrangian methods for semidefinite programming. Math. Program. Comput. **2**, 203–230 (2010)
42. Yang, W.H., Han, D.: Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. SIAM J. Numer. Anal. **54**, 625–640 (2016)
43. Yuan, X., Zeng, S.Z., Zhang, J.: Discerning the linear convergence of ADMM for structured convex optimization through the lens of variational analysis. J. Mach. Learn. Res. **21**, 1–75 (2020). https://jmlr.org/papers/v21/18-562.html