Chapter 4
Smoothing Alternating Direction Methods for Fully Nonsmooth Constrained Convex Optimization

Quoc Tran-Dinh and Volkan Cevher

Abstract We propose two new alternating direction methods to solve "fully" nonsmooth constrained convex problems. Our algorithms have the best known worst-case iteration-complexity guarantee under mild assumptions for both the objective residual and feasibility gap. Through theoretical analysis, we show how to update all the algorithmic parameters automatically with clear impact on the convergence performance. We also provide a representative numerical example showing the advantages of our methods over the classical alternating direction methods using a well-known feasibility problem.

Keywords Gap reduction technique · Smooth alternating minimization · Smooth alternating direction method of multipliers · Homotopy smoothing technique · Nonsmooth constrained convex optimization

AMS Subject Classifications 90C25, 90C06, 90-08

Q. Tran-Dinh (✉)
Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC, USA
e-mail: quoctd@email.unc.edu

V. Cevher
Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
e-mail: volkan.cevher@epfl.ch

P. Giselsson, A. Rantzer (eds.), *Large-Scale and Distributed Optimization*, Lecture Notes in Mathematics 2227,
https://doi.org/10.1007/978-3-319-97478-1_4

4.1 Introduction

In this paper, we aim at developing new optimization algorithms to solve nonsmooth constrained convex optimization problems of the form:

$$f^{\star} := \begin{cases} \min\limits_{x:=(u,v)\in\mathbb{R}^p} & \big\{f(x) := g(u) + h(v)\big\}, \\ \text{s.t.} & Au + Bv = c, \end{cases} \tag{4.1}$$

where $g : \mathbb{R}^{p_1} \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^{p_2} \to \mathbb{R} \cup \{+\infty\}$ are two proper, closed, and convex functions, $A \in \mathbb{R}^{n\times p_1}$, $B \in \mathbb{R}^{n\times p_2}$, $c \in \mathbb{R}^n$ are given, and $p := p_1 + p_2$. Although our proposed methods can solve (4.1) with both smooth and nonsmooth objective functions, we are more interested in the case where both f and g are nonsmooth. In this case, we refer to (4.1) as a "fully" nonsmooth problem since, except for convexity, we do not require any structure assumptions on g and h such as Lipschitz continuous gradient or strong convexity. Problem (4.1) covers many prominent applications such as convex feasibility problems [2], support vector machine [5], matrix completion [8], basis pursuit [33], among many others.

Associated with the primal problem (4.1), we also look at the dual problem:

$$d^{\star} := \min_{\lambda\in\mathbb{R}^n} \big\{d(\lambda) := g^*(A^{\top}\lambda) + h^*(B^{\top}\lambda) - \langle c, \lambda\rangle\big\}, \tag{4.2}$$

where g^* and h^* are the Fenchel conjugates [30] of g and h, respectively; d is the dual function; λ is the dual variable; and $d^{\star}$ denotes the dual optimal value. The convex template (4.1) also manifests itself when we apply convex splitting techniques to decompose the composite objective f into two terms g and h that are coupled via linear constraints. It can also include convex constraints on u and v via indicator functions.

This paper develops a new primal-dual algorithmic framework to solve (4.1) which processes g and h in an alternating fashion to obtain approximately numerical solutions. The alternating optimization approach has regained popularity due to its ability to decentralize data, decompose problem components, and distribute computation in large-scale problems. The underlying theory for the classical alternating optimization methods, such as the alternating direction method of multipliers (ADMM) or the alternating minimization algorithm (AMA), is mature as they have their roots from the splitting methods in monotone inclusions and other classical approaches, such as forward-backward splitting, Douglas-Rachford splitting, Dykstra projections, and Hauzageau's methods [1, 2].

Alternating optimization strategies often provide computational advantages as compared to processing both terms jointly. This approach leads to several methods and variants for solving (4.1) as can be found in the literature, see, e.g., [4, 7, 10–13, 15, 17, 19–22, 24, 28, 31, 32, 34, 37–39]. Among those, ADMM and AMA are the most popular ones. Unlike the standard AMA and ADMM methods and their variants mentioned here, we focus on the case that the objective functions g and

h are nonsmooth and the sum f does not have a "tractable" proximal operator. As an example, in convex feasibility problems, we aim at finding a common point in the intersection of many convex sets. This problem can be formulated into a nonsmooth constrained convex problem (4.1) as we are targeting here. The "full" nonsmoothness of (4.1) creates some fundamental drawbacks for numerical algorithms. First, algorithms that require gradients of the objective function are not applicable. Second, evaluating a proximal operator of the full objective function f becomes impractical. Third, methods using penalty or augmented Lagrangian functions are often inefficient due to complicated subproblems and tuning parameters. A more thorough discussion on our approach and existing methods is postponed to Sect. 4.7. In this paper, we overcome these drawbacks by proposing a combination of different techniques in optimization for solving (4.1).

Our Contributions Our main contribution can be summarized as follows:

(a) (*Theory*) We introduce *a split-gap reduction technique* as a new framework for deriving new alternating direction methods. Our framework unifies the model-based gap reduction technique of [35], smoothing techniques, and the powerful forward-backward and Douglas-Rachford splitting techniques. We establish explicit relations between primal weighting strategy, the parameter choices, and the global convergence rate of the algorithms in our framework.

(b) (*Algorithms and convergence guarantees*) We propose two new smoothing alternating direction optimization algorithms: smoothing alternating minimization algorithm (SAMA), and smoothing alternating direction method of multipliers (SADMM). We derive update rules for all algorithmic parameters including penalty parameters in a heuristic-free fashion. We rigorously characterize the convergence rate of our algorithms for both the objective residual $f(\bar{x}^k) - f^\star$ and the feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$. To the best of our knowledge, this is the best known global convergence rate that can be achieved under mildest assumptions in the literature.

(c) (*Special cases*) We also illustrate that our technique can exploit additional assumptions on A or B, g and h, whenever they are available.

Let us emphasize the following important points of our contribution.

1. (*Mild assumptions*) We only assume that g and h are proper, closed, and convex, the solution set of (4.1) is nonempty, and Slater's condition holds. We also require a technical assumption on the boundedness of the domain of g and h. However, this assumption can be removed by using Lemma 1. Therefore, our methods can solve a broad class of convex optimization problems covered by (4.1).
2. (*Computational complexity*) Our smoothing AMA algorithm essentially has the same per-iteration complexity as the standard AMA [37]. Similarly, our smoothing ADMM has essentially the same per-iteration complexity as the standard ADMM [5]. Although we require additional computation for accelerated steps and averaging, this computation only requires vector-vector additions and scalar-vector multiplications, whose cost is negligible.

3. (*Parameter update*) Our algorithms are heuristic-free in the sense that we update all the parameters automatically at each iteration including the so-called penalty parameter in alternating direction methods [3, 23, 27]. This solves the major drawback in augmented Lagrangian-based methods. We argue that this key feature is important in parallel and distributed implementation, when tuning parameters is impossible to carry out. Intriguingly, our algorithms update their penalty parameters in a decreasing fashion in stark contrast to the classical algorithms.
4. (*Convergence guarantees*) The proposed methods achieve the best known global convergence rate on the primal problem (4.1) as well as on the dual one (4.2) under required assumptions. Moreover, we can explicitly show how the choice of algorithmic parameters can trade-off the convergence guarantee of the objective residual $f(\bar{x}^k)-f^\star$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k-c\|$ in the worst case.

Paper Organization Section 4.2 briefly presents a primal-dual formulation of problem (4.1) under basic assumptions, and characterizes its optimality condition. Section 4.3 deals with a smoothing technique for the primal-dual gap function. Section 4.4 presents a smoothing AMA algorithm and analyzes its convergence. The strongly convex case is also studied in this section. Section 4.5 is devoted to developing a smoothing ADMM algorithm and analyzes its convergence. Section 4.6 presents numerical experiments to verify the performance of our algorithms. We conclude with a discussion of our results in the context of existing work. For clarity of exposition, several technical and new proofs are moved to the Appendix.

Notation In the sequel, we refer to (4.1) as the primal problem. We work on the real and finite dimensional spaces $\mathbb{R}^p$ and $\mathbb{R}^n$, endowed with the inner product $\langle x, \lambda\rangle$ and the standard Euclidean norm $\|\cdot\|$. We use the superscript $\top$ for both the transpose and adjoint operators. For a convex function f, we use ∂f for its subdifferential, and f^* for its Fenchel conjugate. For a convex set $\mathcal{X}$, we use $\delta_{\mathcal{X}}$ for its indicator function, and $\mathrm{ri}(\mathcal{X})$ for its relative interior. We also use $\mathbb{R}_{++}$ for the set of positive real numbers.

For any proper, closed, and convex function $\varphi : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$, the proximal operator is defined as follows:

$$\mathrm{prox}_{\varphi}(x) := \operatorname*{argmin}_{z}\left\{\varphi(z) + (1/2)\|z-x\|^2\right\}. \tag{4.3}$$

Generally, computing prox_{φ} is intractable. However, if prox_{φ} can be efficiently computed in a closed form or in polynomial time, then we say that φ has a *tractable* proximity operator. Several examples can be found, e.g., in [2, 29].

4.2 Preliminaries: Lagrangian Primal-Dual Formulation

This section briefly describes the primal-dual formulation of (4.1) and our fundamental assumptions.

4.2.1 The Dual Problem

Let $x := (u, v) \equiv (u^\top, v^\top)^\top \in \mathbb{R}^p$ be the primal variable, $\mathrm{dom}\,(f) := \mathrm{dom}\,(g) \times \mathrm{dom}\,(h)$, and $\mathcal{D} := \{(u, v) \in \mathrm{dom}\,(f) \mid Au + Bv = c\}$ be the feasible set of (4.1). We define the Lagrange function of (4.1) associated with $Au+Bv = c$ as $\mathcal{L}(x, \lambda) := g(u)+h(v)-\langle \lambda, Au+Bv-c\rangle$, where $\lambda \in \mathbb{R}^n$ is the Lagrange multiplier. We recall the dual problem (4.2) of (4.1) here:

$$d^\star := \min_{\lambda\in\mathbb{R}^n} \left\{ d(\lambda) := \max_u \left\{ \langle A^\top\lambda, u\rangle - g(u) \right\} + \max_v \left\{ \langle B^\top\lambda, v\rangle - h(v) \right\} - c^\top\lambda \right\}, \tag{4.4}$$

where d is the dual function, and two terms can be individually computed as

$$\begin{cases} \varphi(\lambda) := \max\limits_{u\in\mathrm{dom}(g)} \left\{ \langle A^\top\lambda, u\rangle - g(u) \right\} & = g^*(A^\top\lambda), \\ \psi(\lambda) := \max\limits_{v\in\mathrm{dom}(h)} \left\{ \langle B^\top\lambda, v\rangle - h(v) \right\} - c^\top\lambda & = h^*(B^\top\lambda) - c^\top\lambda. \end{cases} \tag{4.5}$$

Let us denote by $u^*(\lambda)$ and $v^*(\lambda)$ one solution of these subproblems, respectively, if they exist. In this case, using the optimality condition, we have $A^\top\lambda \in \partial g(u^*(\lambda))$, which is equivalent to $u^*(\lambda) \in \partial g^*(A^\top\lambda)$. Similarly, $B^\top\lambda \in \partial h(v^*(\lambda))$, which is equivalent to $v^*(\lambda) \in \partial h^*(B^\top\lambda)$. These dual components are convex, but generally nonsmooth. Subgradient or bundle-type methods for directly solving (4.4) are generally inefficient [25, 26].

4.2.2 Basic Assumptions

Let us denote by $\mathcal{X}^\star$ the solution set of (4.1). We say that the *Slater condition* holds for (4.1) if we have

$$\mathrm{ri}(\mathrm{dom}\,(f)) \cap \left\{ (u, v) \in \mathbb{R}^p \mid Au + Bv = c \right\} \neq \emptyset, \tag{4.6}$$

where $\mathrm{ri}(\mathcal{X})$ is the relative interior of $\mathcal{X}$ (see [30]).

For the primal-dual pair (4.1) and (4.4), we require the following assumption:

Assumption 1 *The functions g and h are proper, closed, and convex. The solution set $\mathcal{X}^\star$ of* (4.1) *is nonempty. Either* $\mathrm{dom}\,(f)$ *is polyhedral or the Slater condition* (4.6) *holds.*

Compared to existing methods for solving (4.1) in the literature [4, 7, 10–13, 15, 17, 19–22, 24, 28, 31, 32, 34, 37–39], this assumption is perhaps the mildest one so far. We do not require any strong convexity, error bound, regularity, or Lipschitz gradient assumptions on g and h.

4.2.3 Zero Duality Gap

Under Assumption 1, the solution set $\Lambda^\star$ of the dual problem (4.4) is nonempty and bounded. Moreover, *strong duality* holds, i.e., $f^\star + d^\star = 0$. From the classical duality theory, we have $f(x) + d(\lambda) \geq 0$ for any feasible primal-dual point (x, λ). Hence, the duality gap function G is defined by

$$G(w) := f(x) + d(\lambda) \geq 0, \quad \forall x \in \mathcal{D}, \ \forall \lambda \in \mathbb{R}^n, \tag{4.7}$$

where $w := (x, \lambda)$. Clearly, $G(w^\star) = 0$ (zero duality gap) for any primal-dual solution $w^\star := (x^\star, \lambda^\star) \in \mathcal{X}^\star \times \Lambda^\star$. In addition, $w^\star$ is a saddle point of the Lagrange function; that is $\mathcal{L}(x^\star, \lambda) \leq \mathcal{L}(x^\star, \lambda^\star) = f^\star = -d^\star \leq \mathcal{L}(x, \lambda^\star)$ for all $x \in \operatorname{dom}(f)$ and $\lambda \in \mathbb{R}^n$. The optimality condition of (4.1) can be written as

$$Au^\star + Bv^\star = c, \quad A^\top \lambda^\star \in \partial g(u^\star), \quad \text{and} \quad B^\top \lambda^\star \in \partial h(v^\star). \tag{4.8}$$

4.2.4 Technical Assumption

Apart from Assumption 1, the methods we will develop in the following sections require the following boundedness assumption:

Assumption 2 *Both* $\operatorname{dom}(g)$ *and* $\operatorname{dom}(h)$ *are bounded.*

According to [2, Corollary 17.19], the boundedness of $\operatorname{dom}(g)$ and $\operatorname{dom}(h)$ is equivalent to the Lipschitz continuity of the conjugates g^* and h^*, respectively. Assumption 2 also theoretically restricts the class of problems in (4.1) that we can solve. However, if Assumption 2 does not hold, then we can always add an artificial constraint $\|x\| \leq R$ to (4.1) (or $\|u\| \leq R$ and $\|v\| \leq R$) so that Assumption 2 is satisfied for this modified problem, where $R \in (0, +\infty)$. Under a proper choice of R, this problem is equivalent to (4.1) as showed in the following lemma.

Lemma 1 *Consider two constrained convex optimization problems:*

$$(\mathrm{P}_\infty) \ \ f^\star := \min_{x \in \mathcal{D}} f(x) \qquad \textit{and} \qquad (\mathrm{P}_R) \ \ \bar{f}^\star := \min_{x \in \mathcal{D}} \{ f(x) \mid \|x\| \leq R \},$$

where f *is defined in* (4.1), $\mathcal{D} := \{x = (u, v) \mid Au + Bv = c, \ u \in \operatorname{dom}(g),$ $v \in \operatorname{dom}(h)\}$ *is the feasible set of* (4.1)*, and* $R \in (0, +\infty)$.

If $x^\star$ *is a solution of* (P_∞)*, and* $\|x^\star\| \leq R$*, then it is a solution of* (P_R)*. Conversely, if* $\bar{x}^\star$ *is a solution of* (P_R) *and* $\|\bar{x}^\star\| < R$*, then it is a solution of* (P_∞).

Proof It is obvious that if $x^\star$ is a solution of (P_∞), and $\|x^\star\| \leq R$, then it is a solution of (P_R). Conversely, if $\bar{x}^\star$ is a solution of (P_R), then we have $f(\bar{x}^\star) \leq f(x)$ for all $x \in \mathcal{D}$ and $\|x\| \leq R$. Take any $x \in \mathcal{D} \backslash \mathbb{B}_R$, where $\mathbb{B}_R := \{x \in \mathbb{R}^p \mid \|x\| \leq R\}$ is a ball centered at the origin with radius R. Since $\bar{x}^\star \in \operatorname{int}(\mathbb{B}_R)$, the interior of $\mathbb{B}_R$, there exists $\hat{x}$ on the open segment $(\bar{x}^\star, x)$ such that $\hat{x} = (1 - \tau)\bar{x}^\star + \tau x$

and $\hat{x} \in \mathcal{D} \cap \mathbb{B}_R$, where $\tau \in (0, 1)$. In this case, by convexity of f, we have $f(\bar{x}^\star) \leq f(\hat{x}) = f(1-\tau)\bar{x}^\star + \tau x) \leq (1-\tau)f(x^\star) + \tau f(x)$. Since $\tau \in (0, 1)$, this inequality implies $f(\bar{x}^\star) \leq f(x)$. Therefore, $\bar{x}^\star$ is a solution of (P_∞). □

As suggested by Lemma 1, if we add artificial bounds $\|u\| \leq R$ and $\|v\| \leq R$ to (4.1), then the resulting problem is equivalent to

$$\min_{u,v} \left\{ \hat{g}(u) + \hat{h}(v) \mid Au + Bv = c \right\},$$

where $\hat{g} := g + \delta_{\mathbb{B}_R}$, $\hat{h} := h + \delta_{\mathbb{B}_R}$, and $\delta_{\mathbb{B}_R}$ is the indicator function of the closed ball $\mathbb{B}_R := \{z \mid \|z\| \leq R\}$. This problem has the same form as (4.1). Under Assumption 2, the following quantity:

$$D_f := \sup_{u \in \mathrm{dom}(g),\ \hat{v}, v \in \mathrm{dom}(h)} \left\{ \max\left\{ \|Au + Bv - c\|, \|Au + B(2\hat{v} - v) - c\| \right\} \right\} \tag{4.9}$$

is bounded, i.e., $0 \leq D_f < +\infty$.

Note that, in our algorithms below, since we do not require D_f as an input of the algorithms, this quantity can be heuristically estimated after we terminate the algorithms, and estimate the corresponding artificial radius R based on iteration sequences obtained from the algorithms (see Remark 1).

4.3 Smoothing the Primal-Dual Gap Function

The dual function d defined by (4.4) is convex, but it is generally nonsmooth. Our key idea is to replace the component g^* in (4.5) with a new smoothed approximation g^*_γ to derive new algorithms.

Let us consider the domain $\mathcal{U} := \mathrm{dom}\,(g)$ of g. Associated with $\mathcal{U}$, we choose a proximity function ω, i.e., ω is continuous and strongly convex with the convexity parameter $\mu_\omega = 1 > 0$, and $\mathcal{U} \subseteq \mathrm{dom}\,(\omega)$. In addition, we assume that ω is smooth, and its gradient is Lipschitz continuous with the Lipschitz constant $L_\omega \in [0, +\infty)$.

Given ω, we define the associated Bregman distance

$$b_{\mathcal{U}}(u, \hat{u}) := \omega(u) - \omega(\hat{u}) - \langle \nabla\omega(\hat{u}), u - \hat{u} \rangle. \tag{4.10}$$

Let $\bar{u}_c := \mathrm{argmin}_u\, \omega(u)$ be the prox-center of ω, which exists and is unique. We consider the function $b_{\mathcal{U}}(\cdot, \bar{u}^c)$. Clearly, $b_{\mathcal{U}}(\cdot, \bar{u}^c)$ is smooth and strongly convex with the convexity parameter $\mu_b = \mu_\omega = 1$. Its gradient $\nabla_1 b_{\mathcal{U}}(u, \bar{u}^c) = \nabla\omega(u) - \nabla\omega(\bar{u}^c)$ is Lipschitz continuous with the Lipschitz constant $L_b = L_\omega \geq \mu_\omega = 1$. In addition, $b_{\mathcal{U}}(\bar{u}^c, \bar{u}^c) = 0$ and $\nabla_1 b_{\mathcal{U}}(\bar{u}^c, \bar{u}^c) = 0$.

Given $b_{\mathcal{U}}(\cdot, \bar{u}^c)$, and the conjugate g^* of g, we define

$$g^*_\gamma(z) := \max_{u\in\mathbb{R}^{p_1}} \left\{\langle z, u\rangle - g(u) - \gamma b_{\mathcal{U}}(u, \bar{u}^c)\right\}, \tag{4.11}$$

where $\gamma > 0$ is a smoothness parameter. We denote by $u^*_\gamma(z)$ the solution of the maximization problem in (4.11), i.e.:

$$u^*_\gamma(z) := \arg\max_{u\in\mathbb{R}^{p_1}} \left\{\langle z, u\rangle - g(u) - \gamma b_{\mathcal{U}}(u, \bar{u}^c)\right\}, \tag{4.12}$$

which is well-defined and unique. Clearly, $\nabla g^*_\gamma(z) = u^*_\gamma(z)$ is the gradient of g^*_γ, which has $(1/\gamma)$-Lipschitz gradient. Hence, g^*_γ is $(1/\gamma)$-smooth [2].

Let g^*_γ and ψ be defined by (4.11) and (4.5), respectively, and $\beta > 0$. We consider

$$\begin{cases} d_\gamma(\lambda) & := g^*_\gamma(A^\top\lambda) + \left(h^*(B^\top\lambda) - \langle c, \lambda\rangle\right) = \varphi_\gamma(\lambda) + \psi(\lambda), \\ f_\beta(x) & := g(u) + h(v) + \frac{1}{2\beta}\|Au + Bv - c\|^2, \\ G_{\gamma\beta}(w) & := f_\beta(x) + d_\gamma(\lambda). \end{cases} \tag{4.13}$$

If $\gamma \downarrow 0^+$, then we have $d_\gamma(\lambda) \to d(\lambda)$. Hence, d_γ is a smoothed approximation of d, but it is not fully smooth due to possible nonsmoothness of ψ. For any feasible point $x = (u, v) \in \mathcal{D}$, we have $f_\beta(x) = f(x)$. Here, f_β can be considered as an approximation to f near the feasible set $\mathcal{D}$. Hence, the smoothed gap function $G_{\gamma\beta}$ is an approximation of the duality gap function G in (4.7). Moreover, the smoothed gap function $G_{\gamma\beta}$ is convex. The following lemma shows us how to use $G_{\gamma\beta}$ to characterize the primal-dual solutions for (4.1)–(4.2), whose proof is in section "Proof of Lemma 2: The Primal-Dual Bounds" in Appendix.

Lemma 2 *For any $\bar{x}^k := (\bar{u}^k, \bar{v}^k) \in \mathrm{dom}\,(f)$ and $\bar{\lambda}^k \in \mathbb{R}^n$, it holds that*

$$-\|\lambda^\star\|\|A\bar{u}^k + B\bar{v}^k - c\| \le f(\bar{x}^k) - f^\star \le f(\bar{x}^k) + d(\bar{\lambda}^k). \tag{4.14}$$

Let $\{\bar{w}^k\}$ be an arbitrary sequence in $\mathrm{dom}\,(f) \times \mathbb{R}^n$ and $\{(\gamma_k, \beta_k)\}$ be a sequence in $\mathbb{R}^2_{++}$. Then, the following estimates hold:

$$\begin{cases} f(\bar{x}^k) - f^\star & \le S_k(\bar{w}^k), \\ \|A\bar{u}^k + B\bar{v}^k - c\| & \le 2\beta_k\|\lambda^\star\| + \sqrt{2\beta_k S_k(\bar{w}^k)}, \\ d(\bar{\lambda}^k) - d^\star & \le 2\beta_k\|\lambda^\star\|^2 + \|\lambda^\star\|\sqrt{2\beta_k S_k(\bar{w}^k)} + S_k(\bar{w}^k), \end{cases} \tag{4.15}$$

where $S_k(\bar{w}^k) := G_{\gamma_k\beta_k}(\bar{w}^k) + \gamma_k b_{\mathcal{U}}(u^\star, \bar{u}^c)$, which requires the values of $G_{\gamma\beta}$.

Computing exactly a primal-dual solution $(x^\star, \lambda^\star)$ is impractical. Hence, our objective is to find an approximation $(\bar{x}^k, \bar{\lambda}^k)$ to $(x^\star, \lambda^\star)$ in the following sense:

Definition 1 Given an accuracy $\varepsilon > 0$, a primal-dual point $(\bar{x}^k, \bar{\lambda}^k) \in \mathrm{dom}\,(f) \times \mathbb{R}^n$ is said to be an ε-solution of (4.1)–(4.2) if

$$f(\bar{x}^k) - f^\star \leq \varepsilon, \quad \|A\bar{u}^k + B\bar{v}^k - c\| \leq \varepsilon, \quad \text{and} \quad d(\bar{\lambda}^k) - d^\star \leq \varepsilon.$$

We use the same accuracy parameter ε for each of these terms for simplicity.

We note that by combining $\|A\bar{u}^k + B\bar{v}^k - c\| \leq \varepsilon$ and (4.14), we can guarantee a lower abound $f(\bar{x}^k) - f^\star \geq -\|\lambda^\star\|\varepsilon$. In addition, the domain $\mathrm{dom}\,(f)$ is usually simple (e.g., box, ball, cone, or simplex) so that the constraint $\bar{x}^k \in \mathrm{dom}\,(f)$ can be guaranteed via a closed form projection onto $\mathrm{dom}\,(f)$.

The goal is to generate a primal-dual sequence $\{\bar{w}^k\}$ and a parameter sequence $\{(\gamma_k, \beta_k)\}$ in Lemma 2 such that $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ converges to 0 and $\{(\gamma_k, \beta_k)\}$ also converges to zero. Moreover, the convergence rate of $f(\bar{x}^k) - f^\star$ and $\|A\bar{u}^k + B\bar{v}^k - c\|$ depends on the convergence rate of $\{G_{\gamma_k\beta_k}(\bar{w}^k)\}$ and $\{(\gamma_k, \beta_k)\}$.

4.4 Smoothing Alternating Minimization Algorithm (SAMA)

We propose a new alternating direction method via the application of the accelerated forward-backward splitting to the smoothed gap function. We describe SAMA in three subsections: main steps, initialization, and parameter updates.

4.4.1 Main Steps

At the iteration $k \geq 0$, given $\hat{\lambda}^k \in \mathbb{R}^n$ and the parameters $\gamma_{k+1} > 0$ and $\eta_k > 0$, the main steps of our SAMA consists of two primal alternating direction steps and one dual ascend step as follows:

$$\begin{cases} \hat{u}^{k+1} := \operatorname*{argmin}_{u \in \mathrm{dom}(g)} \left\{ g(u) - \langle A^\top \hat{\lambda}^k, u\rangle + \gamma_{k+1} b_{\mathcal{U}}(u, \bar{u}^c) \right\}, \\ \hat{v}^{k+1} := \operatorname*{argmin}_{v \in \mathrm{dom}(h)} \left\{ h(v) - \langle B^\top \hat{\lambda}^k, v\rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + Bv - c\|^2 \right\}, \\ \bar{\lambda}^{k+1} := \hat{\lambda}^k - \eta_k(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c), \end{cases} \tag{SAMA}$$

where γ_{k+1} and η_k are referred to as the smoothness and the penalty parameter, respectively, and $\bar{u}_c$ is the prox-center of ω in (4.10).

The subproblems in SAMA can often be computed in a closed form. Let us describe two cases. First, if $b_{\mathcal{U}}(\cdot, \bar{u}^c) := (1/2)\|\cdot - \bar{u}^c\|^2$, the standard Euclidean distance, then computing $\hat{u}^{k+1}$ reduces to computing the proximal operator of g, i.e.,

$$\hat{u}^{k+1} = \mathrm{prox}_{\gamma_{k+1}^{-1} g}\big(\bar{u}_c + \gamma_{k+1}^{-1} A^\top \hat{\lambda}^k\big).$$

Second, if we have $B = \mathbb{I}$ or B is orthonormal, then computing $\hat{v}^{k+1}$ reduces to computing the proximal operator of h, i.e.,

$$\hat{v}^{k+1} = \mathrm{prox}_{\eta_k^{-1}h}\big(B^\top(c - A\hat{u}^{k+1}) + \eta_k^{-1}B^\top\hat{\lambda}^k\big).$$

By inspection, it is easy to see that SAMA is an analog of the classical AMA (cf., (4.46)). The first subproblem, due to (4.11), corresponds to the forward step while the last two lines correspond to the backward step. Moreover, if we set $\gamma_{k+1} = 0$ and $\hat{\lambda}^{k+1} = \bar{\lambda}^{k+1}$, SAMA becomes AMA. However, in contrast to the AMA, the SAMA also features a dual acceleration and a primal weighted averaging step:

$$\begin{cases} \hat{\lambda}_k & := (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*, & \text{(dual acceleration)} \\ (\bar{u}^{k+1}, \bar{v}^{k+1}) & := (1-\tau_k)(\bar{u}^k, \bar{v}^k) + \tau_k(\hat{u}^{k+1}, \hat{v}^{k+1}), & \text{(weighted averaging)} \end{cases} \tag{4.16}$$

where $\lambda_k^* := \beta_k^{-1}(c - A\bar{u}^k - B\bar{v}^k)$, and $\tau_k \in (0,1)$ is a given step size. As we will prove in Theorem 1 below, these dual acceleration and primal weighted averaging steps allow us to achieve a better convergence rate on both the primal and the dual spaces compared to standard AMA methods [17].

The following lemma provides conditions showing that the sequence $\{(\bar{x}^k, \bar{\lambda}^k)\}$ generated by (SAMA)–(4.16) maintains the non-monotone gap reduction condition introduced in [36]. The proof of this lemma can be found in section "Proof of Lemma 3: Gap Reduction Condition" in Appendix.

Lemma 3 *Let* $\{\bar{w}^k\}$ *with* $\bar{w}^k := (\bar{u}^k, \bar{v}^k, \bar{\lambda}^k)$ *be the sequence generated by (SAMA)–(4.16). If* $\tau_k \in (0,1]$ *and* $\gamma_k, \beta_k, \eta_k \in \mathbb{R}_{++}$ *satisfy the following conditions:*

$$\begin{array}{ll} (1 + L_b^{-1}\tau_k)\gamma_{k+1} \ge \gamma_k, & \beta_{k+1} \ge (1-\tau_k)\beta_k, \\ (1-\tau_k^2)\gamma_{k+1}\beta_k \ge 2\|A\|^2\tau_k^2, \ \ \textit{and} & 2\|A\|^2\eta_k = \gamma_{k+1}, \end{array} \tag{4.17}$$

then the following non-monotone gap reduction condition holds:

$$G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \le (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \frac{\eta_k\tau_k^2}{4}D_f^2, \tag{4.18}$$

where $G_{\gamma_k\beta_k}$ *is defined by* (4.13) *and* D_f *is defined by* (4.9).

4.4.2 Initialization

We note that we can initialize the algorithm at any starting point $\bar{w}^1 := (\bar{u}^1, \bar{v}^1, \bar{\lambda}^1)$. However, the convergence bounds will depend on $G_{\gamma_1\beta_1}(\bar{w}^1)$. In order to provide transparent convergence results, we propose to use the following initialization in

Lemma 4, whose proof is given in section "Proof of Lemma 4: Bound on $G_{\gamma\beta}$ for the First Iteration" in Appendix.

Lemma 4 *Given $\hat{\lambda}^0 \in \mathbb{R}^m$, $\gamma_1 > 0$, and $\eta_0 > 0$, let $(\bar{u}^1, \bar{v}^1, \bar{\lambda}^1)$ be computed by*

$$\begin{cases} \bar{u}^1 := \operatorname*{argmin}_{u \in \mathrm{dom}(g)} \left\{ g(u) - \langle A^\top \hat{\lambda}^0, u\rangle + \gamma_1 b_{\mathcal{U}}(u, \bar{u}^c) \right\}, \\ \bar{v}^1 := \operatorname*{argmin}_{v \in \mathrm{dom}(h)} \left\{ h(v) - \langle B^\top \hat{\lambda}^0, v\rangle + \frac{\eta_0}{2}\|A\bar{u}^1 + Bv - c\|^2 \right\}, \\ \bar{\lambda}^1 := \hat{\lambda}^0 - \eta_0(A\bar{u}^1 + B\bar{v}^1 - c). \end{cases} \tag{4.19}$$

Then, for any $\beta_1 > 0$, $\bar{w}^1 := (\bar{u}^1, \bar{v}^1, \bar{\lambda}^1)$, and $G_{\gamma\beta}$ defined by (4.13) satisfy

$$\begin{aligned} G_{\gamma_1\beta_1}(\bar{w}^1) \leq \tfrac{\eta_0}{4} D_f^2 + \tfrac{1}{2\eta_0^2}\left[\tfrac{1}{\beta_1} - \tfrac{(5\gamma_1 - 2\eta_0\|A\|^2)\eta_0}{2\gamma_1}\right]\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2 \\ + \eta_0^{-1}\langle\hat{\lambda}^0, \bar{\lambda}^1 - \hat{\lambda}^0\rangle. \end{aligned} \tag{4.20}$$

Consequently, if we choose γ_1, β_1, and η_0 such that $5\gamma_1 > 2\eta_0\|A\|^2$ and $\beta_1 \geq \frac{2\gamma_1}{(5\gamma_1 - 2\eta_0\|A\|^2)\eta_0}$, then $G_{\gamma_1\beta_1}(\bar{w}^1) \leq \frac{\eta_0}{4}D_f^2 + \eta_0^{-1}\langle\hat{\lambda}^0, \bar{\lambda}^1 - \hat{\lambda}^0\rangle$.

4.4.3 Updating the Parameters

For simplicity of presentation, we choose ω as $\omega(u) := \frac{1}{2}\|u - \bar{u}^c\|^2$ for a fixed $\bar{u}_c \in \mathrm{dom}(g)$. In this case, $b_{\mathcal{U}}(\cdot, \bar{u}_c)$ defined by (4.10) becomes $b_{\mathcal{U}}(\cdot, \bar{u}_c) = \frac{1}{2}\| \cdot - \bar{u}_c\|^2$. Hence, we can update τ_k, γ_k, β_k and η_k such that the equality in the conditions (4.17) holds. The following lemma provides **one possibility** to update these parameters whose proof is given in section "Proof of Lemma 5: Parameter Updates" in Appendix.

Lemma 5 *Let $b_{\mathcal{U}}$ be chosen such that $b_{\mathcal{U}}(\cdot, \bar{u}_c) := \frac{1}{2}\| \cdot - \bar{u}_c\|^2$ for a fixed $\bar{u}_c \in \mathrm{dom}(g)$, and $\gamma_1 > 0$. Then, for $k \geq 1$, if τ_k, γ_k, β_k, and η_k are updated by*

$$\tau_k := \frac{3}{k+4}, \quad \gamma_k := \frac{5\gamma_1}{k+4}, \quad \beta_k := \frac{18\|A\|^2(k+5)}{5\gamma_1(k+1)(k+7)}, \quad \textit{and} \quad \eta_k := \frac{5\gamma_1}{2\|A\|^2(k+5)}, \tag{4.21}$$

then they satisfy conditions (4.17). Moreover, the convergence rate of $\{\tau_k\}$ is optimal, and $\beta_k \leq \frac{18\|A\|^2}{5\gamma_1(k+1)}$.

Let us comment here on our weighting strategy and its relation to [12], which places emphasis on the later iterates in averaging by using $\omega_i = i + 1$ as described by (4.45) in Sect. 4.7. In our updates, we consider another weighting scheme (4.45) that places even more emphasis. For this purpose, we use $\omega_i = (i+1)(i+2)$ and rewrite (4.45) in a way to mimic the averaging step in (4.16): $\bar{x}^{k+1} = \frac{1}{k+4}\bar{x}^k + \frac{3}{k+4}x^{k+1}$. Hence, our particular primal weighting scheme (SAMA) uses $\tau_k = \frac{3}{k+4}$.

4.4.4 The New Smoothing AMA Algorithm

Since λ_k^* in the first line of (4.16) requires one matrix-vector multiplication (Au, Bv), we can combine the third line of SAMA and the second line of (4.16) to compute λ_k^* recursively as

$$\lambda_{k+1}^* := \beta_{k+1}^{-1}\big[(1-\tau_k)\beta_k\lambda_k^* + \tau_k\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k)\big]. \tag{4.22}$$

Consequently, each iteration of Algorithm 1 below requires one matrix-vector multiplication (Au, Bv) and one corresponding adjoint operation $(A^\top\lambda, B^\top\lambda)$. Hence, the per-iteration complexity of (SAMA) and the standard AMA (4.46) are essentially the same. Finally, we can combine the main steps (SAMA), (4.16), (4.22), and the update rule (4.21) to complete the smoothing alternating minimization algorithm (SAMA) in Algorithm 1.

Algorithm 1 Smoothing alternating minimization algorithm (SAMA)

```
**Initialization:**
  Fix $\bar{u}_c \in \mathrm{dom}\,(g)$. Choose $\hat{\lambda}^0 \in \mathbb{R}^n$ and $\gamma_1 > 0$.
  Set $\eta_0 := \frac{\gamma_1}{2\|A\|^2}$ and $\beta_1 := \frac{27\|A\|^2}{20\gamma_1}$.
  Compute $\bar{u}^1 := \mathrm{prox}_{\gamma_1^{-1}g}\big(\bar{u}_c + \gamma_1^{-1}A^\top\hat{\lambda}^0\big)$.
  Solve $\bar{v}^1 := \arg\min_v \big\{h(v) - \langle\hat{\lambda}^0, Bv\rangle + \frac{\eta_0}{2}\|A\bar{u}^1 + Bv - c\|^2\big\}$.
  Update $\bar{\lambda}^1 := \hat{\lambda}^0 - \eta_0(A\bar{u}^1 + B\bar{v}^1 - c)$ and $\lambda_1^* := \beta_1^{-1}(c - A\bar{u}^1 - B\bar{v}^1)$.
**Iteration: For** $k = 1$ **to** $k_{\max}$**, perform:**
  Compute $\tau_k := \frac{3}{k+4}$, $\gamma_{k+1} := \frac{5\gamma_1}{k+5}$, $\beta_k := \frac{18\|A\|^2(k+5)}{5\gamma_1(k+1)(k+7)}$ and $\eta_k := \frac{5\gamma_1}{2\|A\|^2(k+5)}$.
  Set $\hat{\lambda}^k := (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda_k^*$.
  Compute $\hat{u}^{k+1} := \mathrm{prox}_{\gamma_{k+1}^{-1}g}\big(\bar{u}_c + \gamma_{k+1}^{-1}A^\top\hat{\lambda}^k\big)$.
  Solve $\hat{v}^{k+1} := \arg\min_v \big\{h(v) - \langle\hat{\lambda}^k, Bv\rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + Bv - c\|^2\big\}$.
  Update $\bar{\lambda}^{k+1} := \hat{\lambda}^k - \eta_k(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c)$.
  Compute $\lambda_{k+1}^* := \beta_{k+1}^{-1}\big[(1-\tau_k)\beta_k\lambda_k^* + \tau_k\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k)\big]$.
  Update $\bar{u}^{k+1} := (1-\tau_k)\bar{u}^k + \tau_k\hat{u}^{k+1}$ and $\bar{v}^{k+1} := (1-\tau_k)\bar{v}^k + \tau_k\hat{v}^{k+1}$.
**End for**
```

We can view Algorithm 1 as a primal-dual method, where we apply Nesterov's accelerated method to the smoothed dual problem while using a weighted averaging scheme $\bar{x}^k = \left(\sum_{i=0}^k \omega_i\right)^{-1} \sum_{i=0}^k \omega_i \hat{x}^i$ for the primal variables. However, Algorithm 1 aims at solving the nonsmooth problem (4.1) without any additional assumption on g and h except for the finiteness of D_f in (4.9).

4.4.5 Convergence Analysis

We prove in section "Proof of Theorem 1: Convergence of Algorithm 1" in Appendix the convergence and the worst-case iteration-complexity of Algorithm 1 in Theorem 1.

Theorem 1 *Assume that $b_{\mathcal{U}}$ is chosen as $b_{\mathcal{U}}(\cdot, \bar{u}_c) := \frac{1}{2}\|\cdot - \bar{u}_c\|^2$ for any fixed $\bar{u}_c \in \mathrm{dom}\,(g)$. Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 1. Then, for any $\gamma_1 > 0$, the following estimates hold*

$$\begin{cases} f(\bar{x}^k) - f^\star & \leq \frac{5\gamma_1}{(k+4)}\left(\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{9D_f^2}{8\|A\|^2(k+3)}\right), \\ \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{36\|A\|^2\|\lambda^\star\|}{5\gamma_1(k+1)} + \frac{6\|A\|}{(k+1)}\sqrt{\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{9D_f^2}{8\|A\|^2(k+7)}}, \\ d(\bar{\lambda}^k) - d^\star & \leq \frac{36\|A\|^2\|\lambda^\star\|^2}{5\gamma_1(k+1)} + \frac{6\|A\|\|\lambda^\star\|}{(k+1)}\sqrt{\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{9D_f^2}{8\|A\|^2(k+7)}} \\ & \quad + \frac{5\gamma_1}{(k+4)}\left(\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{9D_f^2}{8\|A\|^2(k+3)}\right), \end{cases} \tag{4.23}$$

where D_f are defined by (4.9). As a consequence, if we choose $\gamma_1 := \|A\|$, then the worst-case iteration-complexity of Algorithm 1 to achieve an ε-primal-dual solution $(\bar{x}^k, \bar{\lambda}^k)$ of (4.1) and (4.2) in the sense of Definition 1 is $\mathcal{O}\left(\varepsilon^{-1}\right)$.

Theorem 1 shows that the convergence rate of Algorithm 1 consists of two parts. While the first part depends on $\|\bar{u}^c - u^\star\|^2$ which is only $\mathcal{O}(1/k)$, the second part depending on D_f is up to $\mathcal{O}(1/k^2)$. We can obtain the convergence rate of the feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ from the dual convergence as done in [17]. However, this rate is only $\mathcal{O}(1/\sqrt{k})$ when the rate on the dual objective residual $d(\bar{\lambda}^k) - d^\star$ is $\mathcal{O}(1/k)$.

Remark 1 If Assumption 2 fails to hold, then artificial constraints $\|u\| \leq R$ and/or $\|v\| \leq R$ must be added to (4.1). Since Algorithm 1 does not require R as an input, we can estimate R after we terminate this algorithm. Theoretically, the sequence $\left\{(\bar{u}^k, \bar{v}^k)\right\}$ generated by Algorithm 1 converges to $x^\star = (u^\star, v^\star)$ a solution of (4.1). Hence, by Lemma 1, R can roughly be estimated as $R > \sup_k \left\{\|\bar{u}^k\|, \|\bar{v}^k\|\right\}$. Note that, in this case, the objective function of the subproblems in u and v from (SAMA) is also changed from g to $g + \delta_{\mathbb{B}_R}$, and from h to $h + \delta_{\mathbb{B}_R}$, respectively. Practically, by assuming that R is sufficiently large so that $\|u\| \leq R$ and $\|v\| \leq R$ are inactive, we can discard the term $\delta_{\mathbb{B}_R}(u)$, and $\delta_{\mathbb{B}_R}(v)$. Therefore, the computation of $\hat{u}^{k+1}$ and $\hat{v}^{k+1}$ at Step 8 and Step 9, respectively, of Algorithm 1 is unchanged.

4.4.6 Special Case: g is Strongly Convex

We now consider a special case of the constrained problem (4.1) when g is strongly convex. If g is strongly convex with the convexity parameter $\mu_g > 0$, then we can modify Algorithm 1 so that $d(\bar{\lambda}^k) - d^\star \le \mathcal{O}(\frac{1}{k^2})$ in terms of the dual objective function as shown in [17]. However, the convergence rate in terms of the primal objective residual $f(\bar{x}^k) - f^\star$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ we can prove is worse than $\mathcal{O}(\frac{1}{k^2})$.

Let us consider again the dual function φ defined by (4.5). Since g is strongly convex with the strong convexity parameter $\mu_g > 0$, $\nabla\varphi$ is Lipschitz continuous with the Lipschitz constant $L_\varphi := \frac{\|A\|^2}{\mu_g}$. We modify Algorithm 1 in order to obtain a new variant that captures the strong convexity of g and removes the smoothness parameter γ_k. By a similar analysis as in Lemma 3, we can show in section "Proof of Corollary 1: Strong Convexity of g" in Appendix that if the following conditions hold

$$\beta_{k+1} \ge (1-\tau_k)\beta_k \ \text{ and } \ \eta_k\left(\frac{3}{2} + \tau_k - \frac{\|A\|^2\eta_k}{\mu_g}\right) \ge \frac{\tau_k^2}{(1-\tau_k)\beta_k}, \tag{4.24}$$

then

$$G_{\beta_{k+1}}(\bar{w}^{k+1}) \le (1-\tau_k)G_{\beta_k}(\bar{w}^k) + \frac{\tau_k^2\eta_k D_f^2}{4}, \tag{4.25}$$

where $G_{\beta_k}(\bar{w}^k) := f_{\beta_k}(\bar{x}^k) + d(\bar{\lambda}^k)$. The first iterate $\bar{u}^1$ in (4.19) can be computed as

$$\bar{u}^1 := \operatorname*{argmin}_{u\in\mathrm{dom}(f)} \left\{g(u) + \langle\hat{\lambda}^0, Au\rangle\right\}. \tag{4.26}$$

Using (4.26) and new update rules for the parameters in Algorithm 1, we obtain a new variant of Algorithm 1. The following corollary shows the convergence of this variant, whose proof is also moved to section "Proof of Corollary 1: Strong Convexity of g" in Appendix.

Corollary 1 *Let $\{\bar{w}^k\}$ be the sequence generated by Algorithm 1 using (4.26) and the update rules*

$$\tau_k := \frac{3}{k+4}, \quad \eta_k := \frac{\mu_g}{2\|A\|^2}, \quad \text{and} \quad \beta_k := \frac{2\|A\|^2\tau_k^2}{\mu_g(1-\tau_k^2)} = \frac{18\|A\|^2}{\mu_g(k+1)(k+7)}. \tag{4.27}$$

Then, the following estimates hold

$$\begin{cases} f(\bar{x}^k) - f^\star & \leq \frac{9\mu_g D_f^2}{16\|A\|^2(k+3)} & = \mathcal{O}\left(\frac{1}{k}\right), \\ \|A\bar{u}^k + B\bar{v}^k - c\| \leq \frac{36\|A\|^2\|\lambda^\star\|}{\mu_g(k+1)(k+7)} + \frac{9D_f}{2\sqrt{(k+1)(k+3)(k+7)}} & & = \mathcal{O}\left(\frac{1}{k^{3/2}}\right). \end{cases} \tag{4.28}$$

Alternatively, if we use the following update rules in Algorithm 1

$$\tau_k := \frac{3}{k+4}, \quad \eta_k := \frac{\mu_g \tau_k}{\|A\|^2}, \quad and \quad \beta_k := \frac{2\|A\|^2\tau_k}{3\mu_g(1-\tau_k)} = \frac{2\|A\|^2}{\mu_g(k+1)}, \tag{4.29}$$

then

$$\begin{cases} f(\bar{x}^k) - f^\star & \leq \frac{27\mu_g D_f^2}{4\|A\|^2(k+3)^2} & = \mathcal{O}\left(\frac{1}{k^2}\right), \\ \|A\bar{u}^k + B\bar{v}^k - c\| \leq \frac{4\|A\|^2\|\lambda^\star\|}{\mu_g(k+1)} + \frac{3\sqrt{3}}{(k+3)}\frac{D_f}{\sqrt{k+1}} & & = \mathcal{O}\left(\frac{1}{k}\right). \end{cases} \tag{4.30}$$

Here, D_f is defined by (4.9). In both cases, the guarantee of the primal-dual gap function $G(\bar{w}^k) := f(\bar{x}^k) + d(\bar{y}^k)$ is

$$G(\bar{w}^k) + \frac{1}{2\beta_k}\|A\bar{u}^k + B\bar{v}^k - c\|^2 \leq \frac{9\mu_g D_f^2}{4\|A\|^2(k+3)}, \tag{4.31}$$

where β_k is given by either (4.27) or (4.29).

We note that, similar to [17], if we modify Step 11 of Algorithm 1 by $\lambda^*_{k+1} := \lambda^*_k + \frac{1}{\tau_k}\big(\bar{\lambda}^{k+1} - \hat{\lambda}^k\big)$, then we can prove the $\mathcal{O}(\frac{1}{k^2})$-convergence rate for the dual objective residual $d(\lambda^k) - d^\star$ in Algorithm 1 under the strong convexity of g.

4.4.7 Composite Convex Minimization with Linear Operators

A common composite convex minimization formulation in image processing and machine learning [2] is the following problem:

$$\min_{u \in \mathbb{R}_1^p} \left\{ f(u) := g(u) + h(Fu - y) \right\}, \tag{4.32}$$

where g and h are two proper, closed and convex functions (possibly nonsmooth), F is a linear operator from $\mathbb{R}^{p_1}$ to $\mathbb{R}^n$, and $y \in \mathbb{R}^n$ is a given observation vector. We are more interested in the case that g and h are nonsmooth but are equipped with a tractable proximal operator. For example, g and h are both the ℓ_1-norm.

Classical AMA and ADMM methods can solve (4.32) but do not have an $\mathcal{O}(1/k)$ - theoretical convergence rate guarantee without additional smoothness-type, properly proximal terms, or strong convexity-type assumption on g and h. In addition, the ADMM still requires to solve the subproblem at the second line of (4.44) iteratively when F is not orthogonal.

If we introduce a new variable $v := Fu - y$, then we can reformulate (4.32) into (4.1) with $A = F$ and $B = -\mathbb{I}$. In this case, we can apply both Algorithms 1 and 2 (in Sect. 4.5) to solve the resulting problem without additional assumption on g and h except for the boundedness of D_f. However, we only focus on Algorithm 1, which only requires the proximal operator of g and h. The main step of this algorithmic variant can be written explicitly as

$$\begin{cases} \hat{u}^{k+1} := \mathrm{prox}_{\gamma_{k+1}^{-1} g}\big(\bar{u}_c + \gamma_{k+1}^{-1} F^\top \hat{\lambda}^k\big), \\ \hat{v}^{k+1} := \mathrm{prox}_{\eta_k^{-1} h}\big(F\hat{u}^{k+1} - y - \eta_k^{-1}\hat{\lambda}^k\big). \end{cases}$$

Substituting this step into Algorithm 1, we obtain a new variant for solving (4.32) using only the proximal operator of g and h, and matrix-vector multiplications.

4.5 The New Smoothing ADMM Method

For completeness, we present a new alternating direction method of multipliers (ADMM) algorithm for solving (4.1) by applying Douglas-Rachford splitting method to the smoothed dual problem. Our new algorithm, dubbed the smoothing ADMM (SADMM), features similar optimal convergence rate guarantees as SAMA. See Sect. 4.7 for further discussion.

4.5.1 The Main Steps of the Smoothing ADMM Method

The main step of our SADMM scheme is as follows. Given $\hat{\lambda}^k \in \mathbb{R}^n$, $\hat{v}^k \in \mathrm{dom}\,(h)$ and the parameters $\gamma_{k+1} > 0$, $\rho_k > 0$ and $\eta_k > 0$, we compute $(\hat{u}^{k+1}, \hat{v}^{k+1}, \bar{\lambda}^{k+1})$ as follows:

$$\begin{cases} \hat{u}^{k+1} := \operatorname*{argmin}_{u \in \mathrm{dom}(g)} \Big\{ g(u; \gamma_{k+1}) - \langle A^\top \hat{\lambda}^k, u\rangle + \frac{\rho_k}{2}\|Au + B\hat{v}^k - c\|^2 \Big\}, \\ \hat{v}^{k+1} := \operatorname*{arg\,min}_{v \in \mathrm{dom}(h)} \Big\{ h(v) - \langle B^\top \hat{\lambda}^k, v\rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + Bv - c\|^2 \Big\}, \\ \bar{\lambda}^{k+1} := \hat{\lambda}^k - \eta_k\big(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c\big), \end{cases} \tag{SADMM}$$

where $g(u; \gamma) := g(u) + \gamma b_{\mathcal{U}}(u, \bar{u}^c)$. This scheme is different from the standard ADMM scheme (4.44) at two points. First, $\hat{u}^{k+1}$ is computed from the regularized subproblem with $g(\cdot;\gamma)$ instead of g. Second, we use different penalty parameters ρ_k and η_k compared to the standard ADMM scheme (4.44) in Sect. 4.7. The complexity of computing $\hat{u}^{k+1}$ in (SADMM) is essentially the same as computing u^{k+1} in the standard ADMM scheme (4.44) below.

As a special case, if $A = \mathbb{I}$, the identity operator, or A is orthonormal, then we can choose $b_{\mathcal{U}}(\cdot,\bar{u}^c) = (1/2)\|\cdot - \bar{u}^c\|^2$ to obtain a closed form solution of $\hat{u}^{k+1}$ as

$$\hat{u}^{k+1} := \mathrm{prox}_{(\rho_k+\gamma_{k+1})^{-1}g}\left((\rho_k+\gamma_{k+1})^{-1}\big(\gamma_{k+1}\bar{u}^c + A^\top(\hat{\lambda}^k - \rho_k(B\hat{v}^k - c))\big)\right).$$

In addition to (SADMM), our algorithm also requires additional steps

$$\begin{cases} \hat{\lambda}_k & := (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda^*_k, & \text{(dual acceleration)} \\ (\bar{u}^{k+1},\bar{v}^{k+1}) & := (1-\tau_k)(\bar{u}^k,\bar{v}^k) + \tau_k(\hat{u}^{k+1},\hat{v}^{k+1}), & \text{(weighted averaging)} \end{cases} \tag{4.33}$$

as in Algorithm 1, where $\lambda^*_k := \beta_k^{-1}(c - A\bar{u}^k - B\bar{v}^k)$, and $\tau_k \in (0,1)$ is a step size.

We prove in section "Proof of Lemma 6: Gap Reduction Condition" in Appendix the following lemma, which provides conditions on the parameters to guarantee the gap reduction condition.

Lemma 6 *Let* $\{\bar{w}^k\}$ *with* $\bar{w}^k := (\bar{u}^k,\bar{v}^k,\bar{\lambda}^k)$ *be the sequence generated by (SADMM)–(4.33). If* $\tau_k \in (0,1)$ *and* $\gamma_k,\beta_k,\rho_k,\eta_k \in \mathbb{R}_{++}$ *satisfy*

$$\begin{cases} (1-\tau_k)(1+2\tau_k)\eta_k\beta_k \geq 2\tau_k^2, & \gamma_{k+1} \geq \left(\frac{3-2\tau_k}{3-(2-L_b^{-1})\tau_k}\right)\gamma_k, \\ \beta_{k+1} \geq (1-\tau_k)\beta_k, \text{ and } & \gamma_{k+1} \geq \|A\|^2\left(\eta_k + \frac{\rho_k}{\tau_k}\right), \end{cases} \tag{4.34}$$

then the following non-monotone gap reduction condition holds

$$G_{\gamma_{k+1}\beta_{k+1}}(\bar{w}^{k+1}) \leq (1-\tau_k)G_{\gamma_k\beta_k}(\bar{w}^k) + \left(\frac{\tau_k^2\eta_k}{4} + \frac{\tau_k\rho_k}{2}\right)D_f^2, \tag{4.35}$$

where $G_{\gamma_k\beta_k}$ *is defined by* (4.13)*, and* D_f *is defined by* (4.9)*.*

4.5.2 Updating Parameters

The second step of our algorithmic design is to derive an update rule for the parameters to satisfy the conditions (4.34). Lemma 7 shows **one possibility** to update these parameters, whose proof is given in section "Proof of Lemma 7: Parameter Updates" in Appendix.

Lemma 7 *Let $b_{\mathcal{U}}$ be chosen such that $b_{\mathcal{U}}(\cdot, \bar{u}_c) := \frac{1}{2}\|\cdot - \bar{u}_c\|^2$ for a fixed $\bar{u}_c \in \mathrm{dom}\,(g)$, and $\gamma_1 > 0$. Then, for $k \geq 1$, τ_k, γ_k, β_k, ρ_k, and η_k updated by*

$$\begin{array}{lll} \tau_k := \frac{3}{k+4}, & \gamma_k := \frac{3\gamma_1}{k+2}, & \beta_k := \frac{6\|A\|^2(k+3)}{\gamma_1(k+1)(k+10)}, \\ \rho_k := \frac{9\gamma_1}{2\|A\|^2(k+3)(k+4)}, & \eta_k := \frac{3\gamma_1}{2\|A\|^2(k+3)}, & \end{array} \tag{4.36}$$

satisfy (4.34)*. Moreover, $\beta_k \leq \frac{9\|A\|^2}{5\gamma_1(k+1)}$, and the convergence rate of $\{\tau_k\}$ is optimal.*

We note that we have freedom to choose γ_1 in order to trade-off the upper-bound of the primal objective residual $f(\bar{x}^k) - f^\star$ and the primal feasibility gap $\|A\bar{u}^k + B\bar{v}^k - c\|$ as in Algorithm 1.

4.5.3 The Smoothing ADMM Algorithm

Similar to Algorithm 1, we can combine the third line of (SADMM) and the second line of (4.33) to update λ^*_k. In this case, the arithmetic cost-per-iteration of Algorithm 2 is essentially the same as in the standard ADMM scheme (4.44). We also use $\bar{w}^1 = (\bar{u}^1, \bar{v}^1, \bar{\lambda}^1)$ computed by (4.19) at the first iteration. By putting (4.19), (4.36), (SADMM), (4.33) and (4.22) together, we obtain a complete SADMM algorithm as presented in Algorithm 2.

4.5.4 Convergence Analysis

The following theorem with its proof being in section "Proof of Theorem 2: Convergence of Algorithm 2" in Appendix shows the worst-case iteration-complexity of Algorithm 2.

Theorem 2 *Assume that $b_{\mathcal{U}}$ is chosen as $b_{\mathcal{U}}(\cdot, \bar{u}_c) := \frac{1}{2}\|\cdot - \bar{u}_c\|^2$ for a fixed $\bar{u}_c \in \mathrm{dom}\,(g)$. Let $\{(\bar{u}^k, \bar{v}^k, \bar{\lambda}^k)\}$ be the sequence generated by Algorithm 2. Then the following estimates hold*

$$\begin{cases} f(\bar{x}^k) - f^\star & \leq \frac{3\gamma_1}{(k+2)}\left[\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{27D_f^2}{8\|A\|^2(k+3)}\right], \\ \|A\bar{u}^k + B\bar{v}^k - c\| & \leq \frac{18\|A\|^2\|\lambda^\star\|}{5\gamma_1(k+1)} + \frac{6\|A\|}{(k+1)}\sqrt{\|\bar{u}^c - u^\star\|^2 + \frac{27D_f^2}{8\|A\|^2(k+10)}}, \end{cases} \tag{4.37}$$

where D_f is given by (4.9)*. If $\gamma_1 := \|A\|$, then the worst-case iteration-complexity of Algorithm 2 to achieve an ε—solution $\bar{x}^k$ of* (4.1) *is $\mathcal{O}\left(\varepsilon^{-1}\right)$.*

As can be seen from Theorem 2, the term $\frac{6\|A\|}{(k+1)}\left(\frac{\|\bar{u}^c - u^\star\|^2}{2} + \frac{27D_f^2}{8\|A\|^2(k+10)}\right)^{1/2}$ in (4.37) does not depend on the choice of γ_1. If we decrease γ_1, then the upper

Algorithm 2 Smoothing alternating direction method of multipliers (SADMM)

```
Initialization:
  Fix $\bar{u}_c \in \mathrm{dom}(g)$. Choose $\hat{\lambda}^0 \in \mathbb{R}^n$ and $\gamma_1 > 0$.
  Set $\eta_0 := \frac{\gamma_1}{2\|A\|^2}$ and $\beta_1 := \frac{12\|A\|^2}{11\gamma_1}$.
  Compute $\bar{u}^1 := \mathrm{prox}_{\gamma_1^{-1}g}\big(\bar{u}_c + \gamma_1^{-1}A^\top\hat{\lambda}^0\big)$.
  Solve $\bar{v}^1 := \arg\min_v \big\{h(v) - \langle\hat{\lambda}^0, Bv\rangle + \frac{\eta_0}{2}\|A\bar{u}^1 + Bv - c\|^2\big\}$. Set $\hat{v}^1 := \bar{v}^1$.
  Update $\bar{\lambda}^1 := \hat{\lambda}^0 - \eta_0(A\bar{u}^1 + B\bar{v}^1 - c)$ and $\lambda^*_1 := \beta_1^{-1}(c - A\bar{u}^1 - B\bar{v}^1)$.
Iteration: For $k = 1$ to $k_{\max}$, perform:
  Compute $\tau_k := \frac{3}{k+4}$, $\gamma_{k+1} := \frac{3\gamma_1}{k+3}$, $\beta_k := \frac{6\|A\|^2(k+3)}{\gamma_1(k+1)(k+10)}$. Then, set $\eta_k := \frac{3\gamma_1}{2\|A\|^2(k+3)}$ and $\rho_k := \frac{9\gamma_1}{2\|A\|^2(k+3)(k+4)}$.
  Set $\hat{\lambda}^k := (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda^*_k$.
  Solve $\hat{u}^{k+1} := \arg\min_u\big\{g(u) - \langle\hat{\lambda}^k, Au\rangle + \frac{\rho_k}{2}\|Au + B\hat{v}^k - c\|^2 + \gamma_{k+1}b_{\mathcal{U}}(u, \bar{u}^c)\big\}$.
  Solve $\hat{v}^{k+1} := \arg\min_v \big\{h(v) - \langle\hat{\lambda}^k, Bv\rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + Bv - c\|^2\big\}$.
  Update $\bar{\lambda}^{k+1} := \hat{\lambda}^k - \eta_k(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c)$.
  Compute $\lambda^*_{k+1} := \beta_{k+1}^{-1}\big[(1-\tau_k)\beta_k\lambda^*_k + \tau_k\eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k)\big]$.
  Update $\bar{u}^{k+1} := (1-\tau_k)\bar{u}^k + \tau_k\hat{u}^{k+1}$ and $\bar{v}^{k+1} := (1-\tau_k)\bar{v}^k + \tau_k\hat{v}^{k+1}$.
End for
```

bound of $f(\bar{x}^k) - f^\star$ decreases, while the upper bound of $\|A\bar{u}^k + B\bar{v}^k - c\|$ increases, and vice versa. Hence, γ_1 trades off these worse-case bounds. The convergence rate guarantee on the dual objective residual can be easily obtained from the last bound of (4.15).

4.5.5 SAMA vs. SADMM

There are at least two cases, where SAMA theoretically gains advantages over SADMM. First, if A is non-orthogonal. The u-subproblem in (SAMA) can be computed by using prox_g, while in SADMM, the nonorthogonal operator A prevents us from using prox_g. Second, if g is block separable, i.e., $g(u) := \sum_{i=1}^s g_i(u_i)$, then we can choose $g(u;\gamma) := \sum_{i=1}^s \big[g_i(u_i) + \frac{\gamma}{2}\|u_i - \bar{u}_i^c\|^2\big]$, which can be evaluated in parallel. This is not preserved in SADMM. Indeed, for SADMM, the subproblem in u still has the quadratic term $\frac{\rho_k}{2}\|Au + B\hat{v}^k - c\|^2$, which makes it nonseparable even if g is separable.

4.6 Numerical Evidence

We illustrate a "geometric invariant" property of Algorithms 1 and 2 for solving the distance minimization problem (4.39). This problem is classical but solving it efficiently remains an interesting research topic. Various algorithms have been proposed including Douglas-Rachford (DR) splitting, Dykstra's projection, and Hauzageau's method [1, 2]. In this section, we compare our algorithms with these methods.

We consider the following convex feasibility problem with two convex sets:

$$\text{Find } \lambda^\star \text{ such that: } \lambda^\star \in \mathcal{C}_1 \cap \mathcal{C}_2, \tag{4.38}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are two nonempty, closed, and convex sets in $\mathbb{R}^p$. Problem (4.38) may not have solution. Hence, instead of solving (4.38), we consider a problem of finding the best substitution for a point in the intersection $\mathcal{C}_1 \cap \mathcal{C}_2$ even if it is empty. Such a problem can be formulated as

$$d^\star := \min_{\lambda \in \mathbb{R}^n} \left\{ d(\lambda) := d_{\mathcal{C}_1}(\lambda) + d_{\mathcal{C}_2}(\lambda) \right\}, \tag{4.39}$$

where $d_{\mathcal{C}}$ is the Euclidean distance to the set $\mathcal{C}$. Unlike (4.38), the optimal value d^* of (4.39) is always finite as long as $\mathcal{C}_1$ and $\mathcal{C}_2$ are nonempty. Moreover, $d^\star = \mathrm{dist}(\mathcal{C}_1, \mathcal{C}_2)$, the distance between $\mathcal{C}_1$ and $\mathcal{C}_2$. Hence, if $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$, then $d^\star = 0$, see, e.g., [6].

According to [6], our primal template (4.1) for (4.39) then takes the following form

$$\min_{u,v} \left\{ s_{\mathcal{C}_1}(u) + s_{\mathcal{C}_2}(v) \quad | \quad u + v = 0,\ u \in \mathbb{B}_1,\ v \in \mathbb{B}_1 \right\}, \tag{4.40}$$

where $s_{\mathcal{C}_i}$ is the support function of $\mathcal{C}_i$ for $i = 1, 2$, and $\mathbb{B}_r := \{w \mid \|w\| \leq r\}$ for $r > 0$.

Clearly, (4.40) is fully nonsmooth, since $s_{\mathcal{C}_i}$ is convex and nonsmooth for $i = 1, 2$. In addition, (4.40) satisfies Assumption 2. Here, we can even increase the constraint radius, currently 1, to a sufficiently large number such that the constraints $u, v \in \mathbb{B}_r$ of each subproblems in (4.44), (SAMA) and (SADMM) are inactive without changing the underlying problem. In this particular setting, we can choose the prox-center points for u and v as zero since they actually obtain the optimal solution.

If we apply ADMM to solve (4.40), then it can be written explicitly as

$$\begin{cases} u^{k+1} := \mathrm{prox}_{\rho^{-1}s_{\mathcal{C}_1}}(\lambda^k - v^k) \quad = \lambda^k - v^k - \rho^{-1}\pi_{\mathcal{C}_1}\left(\rho(\lambda^k - v^k)\right), \\ v^{k+1} := \mathrm{prox}_{\rho^{-1}s_{\mathcal{C}_2}}(\lambda^k - u^{k+1}) = \lambda^k - u^{k+1} - \rho^{-1}\pi_{\mathcal{C}_2}\left(\rho(\lambda^k - u^{k+1})\right), \\ \lambda^{k+1} := \lambda^k - (u^{k+1} + v^{k+1}), \end{cases}$$

where $\pi_{\mathcal{C}_i}$ is the projection onto $\mathcal{C}_i$ for $i = 1, 2$, and $\rho > 0$ is the penalty parameter. Clearly, multiplying this expression by ρ and using the same notation, we obtain

$$\begin{cases} u^{k+1} := \lambda^k - v^k - \pi_{\mathcal{C}_1}\left(\lambda^k - v^k\right), \\ v^{k+1} := \lambda^k - u^{k+1} - \pi_{\mathcal{C}_2}\left(\lambda^k - u^{k+1}\right), \\ \lambda^{k+1} := \lambda^k - (u^{k+1} + v^{k+1}), \end{cases} \tag{4.41}$$

which shows that this scheme is independent of any parameter ρ. With an elementary transformation, we can write (4.41) as a Douglas-Rachford (DR) splitting scheme

$$\begin{cases} z^k \;\; := z^{k-1} + \pi_{\mathcal{C}_1}\left(2\lambda^k - z^{k-1}\right) - \lambda^k, \\ \lambda^{k+1} := \pi_{\mathcal{C}_2}(z^k). \end{cases} \tag{4.42}$$

To recover u^k and v^k from z^k and λ^k, we can use $u^k := \lambda^{k-1} - z^k$ and $v^k := z^{k-1} - \lambda^k$.

Now, if we apply our SAMA to solve (4.40) using $b_{\mathcal{U}}(u, \bar{u}^c) := (1/2)\|u - \bar{u}^c\|^2$, the two main steps of SAMA becomes

$$\begin{cases} \hat{u}^{k+1} := \mathrm{prox}_{\gamma_{k+1}^{-1} s_{\mathcal{C}_1}}(\bar{u}^c + \gamma_{k+1}^{-1}\hat{\lambda}^k) = \gamma_{k+1}^{-1}\hat{\lambda}^k + \bar{u}^c - \gamma_{k+1}^{-1}\pi_{\mathcal{C}_1}\left(\hat{\lambda}^k + \gamma_{k+1}\bar{u}^c\right), \\ \hat{v}^{k+1} := \mathrm{prox}_{\eta_k^{-1} s_{\mathcal{C}_2}}(\eta_k^{-1}\hat{\lambda}^k - \hat{u}^{k+1}) = \eta_k^{-1}\hat{\lambda}^k - \hat{u}^{k+1} - \eta_k^{-1}\pi_{\mathcal{C}_2}\left(\hat{\lambda}^k - \eta_k\hat{u}^{k+1}\right). \end{cases} \tag{4.43}$$

Clearly, the standard AMA is not applicable to solve (4.40) due to the lack of strong convexity. The standard ADMM applying to (4.40) becomes the alternative projection scheme (4.42) for solving (4.38). This scheme can be arbitrarily slow if the geometry between two sets $\mathcal{C}_1$ and $\mathcal{C}_2$ is ill-posed (see below).

To observe an interesting convergence behavior, we test Dykstra's projection, Hauzageau's method, and the ADMM (4.41) (or its DR form (4.42)), and compare them with our algorithms in the following configuration.

We first choose $\mathcal{C}_i := \{u \in \mathbb{R}^n \mid \langle \mathbf{a}_i, u\rangle \leq b_i\}$ for $i = 1, 2$ as two half-planes in $\mathbb{R}^n$, where $b_1 = b_2 = 0$. Here, the normal vectors are $\mathbf{a}_1 := (\epsilon, \cdots, \epsilon, -1, \cdots, -1)^\top$, and $\mathbf{a}_2 := (0, \cdots, 0, 1, \cdots, 1)^\top$, where $\epsilon > 0$ is a positive angle. The tangent angle ϵ is repeated $\lfloor n/2 \rfloor$ times in $\mathbf{a}_1$, and the zero is repeated $\lfloor n/2 \rfloor$ times in $\mathbf{a}_2$, where $n = 1000$. The starting point is chosen as $u^0 := (1, \cdots, 1)^\top$. By varying ϵ, we can observe the convergence behavior of these five methods.

We note that Dykstra's and Hauzageau's algorithms directly solve the dual problem (4.39), while our methods and ADMM solve both the primal and dual problems (4.40) and (4.39). We compare these algorithms on the absolute dual objective residual $d(\lambda) - d^\star$ of (4.39).

Figure 4.1 shows the convergence of five algorithms with different choices of ϵ.

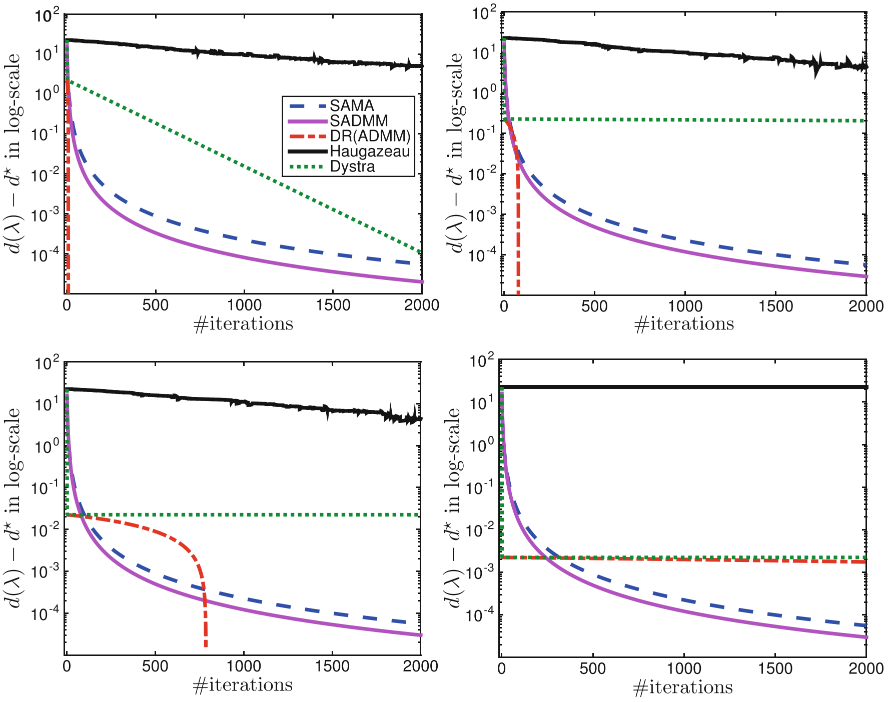

Fig. 4.1 The convergence behavior of five algorithms with different values of ϵ. These plots correspond to $\epsilon = 10^{-1}$ (left-top), 10^{-2} (right-top), 10^{-3} (left-bottom) or 10^{-4} (right-bottom)

We observe Hauzageau's and Dykstra's methods are slow, but Hauzageau's method is extremely slow. The speed of ADMM (or DR splitting) strongly depends on the geometry of the sets, in particular, the tangent angle between two sets. For large values of ϵ, these methods work well, but they become arbitrarily slow when ϵ is decreasing. The objective value of this method drops quickly to a certain level, then is saturated, and makes a very slow progress toward to the optimal value as seen in Fig. 4.1. Since the ADMM scheme (4.41) is independent of its penalty parameter, this is the best performance we can achieve for solving (4.39). Both SAMA and SADMM have almost identical convergence rate for different values of ϵ. These convergence rate reflects the theoretical guarantee, which is $\mathcal{O}(1/k)$ as predicted by our theoretical results.

4.7 Discussion

We have developed a rigorous alternating direction optimization framework for solving constrained convex optimization problems. Our approach is built upon the model-based gap reduction (MGR) technique in [35], and unifies five main ideas: smoothing, gap reduction, alternating direction, acceleration/averaging, and homotopy. By splitting the gap, we have developed two new smooth alternating optimization algorithms: SAMA and SADMM with rigorous convergence guarantees. One important feature of these methods is a heuristic-free parameter update, which has not been proved yet in the literature for AMA and ADMM as we discuss below:

(a) **Alternating direction method of multipliers** (ADMM). The ADMM algorithm can be viewed as the Douglas-Rachford splitting applied to the optimality condition of the dual problem (4.2). As a result, the standard ADMM algorithm generates a primal sequence $\left\{(u^k, v^k)\right\}$ together with a multiplier sequence $\left\{\lambda^k\right\}$ as

$$\begin{cases} u^{k+1} := \underset{u \in \mathrm{dom}(g)}{\mathrm{argmin}} \left\{ g(u) - \langle \lambda^k, Au \rangle + \frac{\eta_k}{2} \|Au + Bv^k - c\|_2^2 \right\} \\ v^{k+1} := \underset{v \in \mathrm{dom}(h)}{\mathrm{argmin}} \left\{ h(v) - \langle \lambda^k, Bv \rangle + \frac{\eta_k}{2} \|Au^{k+1} + Bv - c\|_2^2 \right\} \\ \lambda^{k+1} := \lambda^k - \eta_k\left(Au^{k+1} + Bv^{k+1} - c\right), \end{cases} \tag{4.44}$$

where k denotes the iteration count and $\eta_k > 0$ is a penalty parameter. This basic method is closely related to or equivalent to many other algorithms, such as Spingarn's method of partial inverses, Dykstra's alternating projections, Bregman's iterative algorithms, and can also be motivated from the augmented Lagrangian perspective [5].

The ADMM algorithm serves as a good general-purpose tool for optimization problems arising in the analysis and processing of modern massive datasets. Indeed, its implementations have received a significant amount of engineering effort both in research and in industry. As a result, its global convergence rate characterizations for the template (4.1) is an active research topic, see, e.g., [10–13, 15, 17, 19, 22, 28, 31, 38], and the references quoted therein.

In the constrained setting of (4.1), a global convergence characterization specifically means the following: The algorithm provides us $\bar{x}^k = (\bar{u}^k, \bar{v}^k)$ and we determine the number of iterations k necessary to obtain $f(\bar{x}^k) - f^\star \le \epsilon_f$ and $\|A\bar{u}^k + B\bar{v}^k - c\| \le \epsilon_c$ for some fixed accuracy ϵ_f for the objective and for some—possibly another—fixed accuracy ϵ_c for the linear constraint. Separating constraint feasibility is crucial so that the primal convergence has any significance otherwise we can trivially have $f^\star - f(\bar{x}^k) \le 0$ for some infeasible iterate $\bar{x}^k$.

A key theoretical strategy for obtaining global convergence rates for alternating direction methods is ergodic averaging [10–12, 19, 22, 24, 28, 31, 38]. For instance, as opposed to working with the primal-sequence $x^k := (u^k, v^k)$ from (4.44) directly, we instead choose a sequence of weights $\{\omega_k\} \subset (0, +\infty)$ and then average as follows

$$\bar{x}^k := \Big(\sum_{i=0}^{k} \omega_i\Big)^{-1} \sum_{i=0}^{k} \omega_i x^i. \tag{4.45}$$

The averaged sequence $\bar{x}^k$ then makes it theoretically elementary to obtain the desired type of convergence rate characterizations for (4.1).

Indeed, existing literature critically relies on such weighting strategies in order to obtain global convergence guarantees. For instance, He and Yuan in [19] prove an $\mathcal{O}(1/k)$-convergence rate of their ADMM scheme (4.44) by using the form (4.45) with $\omega_i := 1$ but for both primal and dual variables x as well as λ simultaneously. They provided their guarantee in terms of a gap function for an associated variational inequality for (4.1) and assumed the boundedness on both primal and dual domains. This result is further extended by other authors to different variants of ADMM, including [18, 34, 39]. The same rate is obtained in [12] for a relaxed ADMM variant with similar assumptions along with a weighting strategy that emphasizes the latter iterations by using $\omega_i := k + 1$ in (4.45).

We should note that there are also weighted global convergence characterizations for ADMM, such as $f(\bar{x}^k) - f^\star + \rho\|A\bar{u}^k + B\bar{v}^k - c\|$ for some fixed $\rho > 0$ by Shefi and Teboulle [31]. The authors added proximal terms to the u- and v-subproblems and imposed conditions on three parameters to achieve the $\mathcal{O}(1/k)$-convergence rate jointly between the objective residual and feasibility gap. Intriguingly, this type of convergence rate guarantee does not necessarily imply the $\mathcal{O}(1/k)$-convergence separately on the primal objective residual and feasibility gap as indicated in [31, Theorem 5.2] without additional assumptions.

Interestingly, making additional assumptions on the template is quite common [12, 14, 16, 17]. For instance, the authors in [28] studied a linearized ADMM variant of (4.44) and proved the $\mathcal{O}(1/k)$-rate separately, but required the Lipschitz gradient assumption on either g or h in (4.1). In addition, the authors in [17] require strong convexity on both g and h. In contrast, the authors [14] require the strong convexity of either g or h but need A or B to be full rank as well. In [39] the authors proposed an asynchronous ADMM and showed the $\mathcal{O}(1/k)$ rate on the averaging sequence for a special case of (4.1) where $h = 0$, which trivially has Lipschitz gradient.

Unsurprisingly, these assumptions again limit the applicability of the algorithmic guarantees when, for instance, g and h are non-Lipschitz gradient loss functions or fully non-smooth regularizers, as in Poisson imaging, robust principal component analysis (RPCA), and graphical model learning [9]. Several recent results rely on other type of assumptions such as error bounds, metric regularity, or the well-known Kurdyka-Lojasiewicz condition [7, 20, 21]. Although these conditions cover a wide range of application models, it is unfortunately very hard to verify some quantities related to these assumptions in practice. Other times, the additional assumptions obviate the ADMM choice as they can allow application of a simpler algorithm:

(b) Alternating minimization algorithm (AMA). The AMA algorithm, given below, is guaranteed to converge when g is strongly convex or g^* has Lipschitz gradient [17]:

$$\begin{cases} \hat{u}^{k+1} := \mathop{\mathrm{argmin}}\limits_{u\in\mathrm{dom}(g)} \big\{g(u) - \langle\hat{\lambda}^k, Au\rangle\big\},\\ \hat{v}^{k+1} := \mathop{\mathrm{argmin}}\limits_{v\in\mathrm{dom}(h)} \big\{h(v) - \langle\hat{\lambda}^k, Bv\rangle + \frac{\eta_k}{2}\|A\hat{u}^{k+1} + Bv - c\|_2^2\big\},\\ \hat{\lambda}^{k+1} := \hat{\lambda}^k - \eta_k\big(A\hat{u}^{k+1} + B\hat{v}^{k+1} - c\big), \end{cases} \tag{4.46}$$

where $\eta_k > 0$ is a penalty parameter.

One can view AMA as the forward-backward splitting algorithm applied to the optimality condition of the dual problem (4.2) (cf., [17, 37]). Alternatively, we can motivate the algorithm by using one Lagrange dual step and one augmented Lagrangian dual step between two blocks of variables u and v [4, 32, 37]. Computationally, (4.46) is arguably easier than (4.44). However, it often requires stronger assumptions than ADMM to guarantee convergence [17, 37]. The most obvious assumption is the strong convexity of g.

Acknowledgements QTD's work was supported in part by the NSF-grant No. DMS-1619884, USA. VC's work was supported by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 725594—time-data). The authors would like to acknowledge Dr. C.B., Vu, and Dr. V.Q. Nguyen with their help on verifying the technical proofs and the numerical experiment. The authors also thank Mr. Ahmet Alacaoglu, Mr. Nhan Pham, and Ms. Yuzixuan Zhu for their careful proofreading.

Appendix: Proofs of Technical Results

This appendix provides full proofs of technical results presented in the main text.

Proof of Lemma 2: The Primal-Dual Bounds

First, using the fact that $-d(\lambda) \le -d^\star = f^\star \le \mathcal{L}(x, \lambda^\star) = f(x) + \langle \lambda^\star, Au + Bv - c\rangle \le f(x) + \|\lambda^\star\|\|Au + Bv - c\|$, we get

$$-\|\lambda^\star\|\|Au + Bv - c\| \le f(x) - f^\star \le f(x) + d(\lambda), \tag{4.47}$$

which is exactly the lower bound (4.14).

Next, since $A^\top\lambda^\star \in \partial g(u^\star)$ due to (4.8), by Fenchel-Young's inequality, we have $g(u^\star) + g^*(A^\top\lambda^\star) = \langle A^\top\lambda^\star, u^\star\rangle$, which implies $g^*(A^\top\lambda^\star) = \langle A^\top\lambda^\star, u^\star\rangle - g(u^\star)$. Using this relation and the definition of φ_γ, we have

$$\begin{aligned}\varphi_\gamma(\lambda) &:= \max\left\{\langle A^\top\lambda, u\rangle - g(u) - \gamma b_{\mathcal{U}}(u, \bar{u}^c)\right\} \ge \langle A^\top\lambda, u^\star\rangle - g(u^\star) - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c)\\ &= \langle A^\top\lambda^\star, u^\star\rangle - g(u^\star) + \langle A^\top(\lambda - \lambda^\star), u^\star\rangle - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c)\\ &= g^*(A^\top\lambda^\star) + \langle A^\top(\lambda - \lambda^\star), u^\star\rangle - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c)\\ &= \varphi(\lambda^\star) + \langle \lambda - \lambda^\star, Au^\star\rangle - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c).\end{aligned}$$

Alternatively, we have $\psi(\lambda) \ge \psi(\lambda^\star) + \langle \nabla\psi(\lambda^\star), \lambda - \lambda^\star\rangle$, where $\nabla\psi(\lambda^\star) = B\nabla h^*(B^\top\lambda^\star) - c = Bv^\star - c$ due to the last relation in (4.8), where $\nabla h^*(B^\top\lambda^\star) \in$

$\partial h^*(B^\top\lambda^\star)$ is one subgradient of ∂h^*. Hence, $\psi(\lambda) \geq \psi(\lambda^\star) + \langle \lambda - \lambda^\star, Bv^\star - c\rangle$. Adding this inequality to the last estimation with the fact that $d_\gamma = \varphi_\gamma + \psi$ and $d = \varphi + \psi$, we obtain

$$d_\gamma(\lambda) \geq d(\lambda^\star) + \langle \lambda - \lambda^\star, Au^\star + Bv^\star - c\rangle - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c) \overset{(4.8)}{=} d^\star - \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c) \tag{4.48}$$

Using this inequality with $d^\star = -f^\star$ and the definition (4.13) of f_β we have

$$f(x) - f^\star \overset{(4.13)+(4.48)}{\leq} f_\beta(x) + d_\gamma(\lambda) + \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c) - \frac{1}{2\beta}\|Au + Bv - c\|^2 \tag{4.49}$$

$$= G_{\gamma\beta}(w) + \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c) - \frac{1}{2\beta}\|Au + Bv - c\|^2.$$

Let $S := G_{\gamma\beta}(w) + \gamma b_{\mathcal{U}}(u^\star, \bar{u}^c)$. Then, by dropping the last term $-\frac{1}{2\beta}\|Au + Bv - c\|^2$ in (4.49), we obtain the first inequality of (4.15).

Let $t := \|Au + Bv - c\|$. Using again (4.47) and (4.49), we can see that $\frac{1}{2\beta}t^2 - \|\lambda^\star\|t - S \leq 0$. Solving this quadratic inequation w.r.t. t and noting that $t \geq 0$, we obtain the second bound of (4.15). The last estimate of (4.15) is a direct consequence of (4.49), the first one of (4.15). Finally, from (4.47), we have $f(x) \geq f^\star - \|\lambda^\star\|\|Au + Bv - c\|$. Substituting this into (4.49) we get $d(\lambda) - d^\star - \|\lambda^\star\|\|Au + Bv - c\| \leq S - \frac{1}{2\beta}\|Au + Bv - c\|^2$, which implies

$$d(\lambda) - d^\star \leq S - (1/(2\beta))\|Au + Bv - c\|^2 + \|\lambda^\star\|\|Au + Bv - c\|.$$

By discarding $-(1/(2\beta))\|Au + Bv - c\|^2$ and using the second estimate of (4.15) into the last estimate, we obtain the last inequality of (4.15). □

Convergence Analysis of Algorithm 1

We provide a full proof of Lemmas and Theorems related to the convergence of Algorithm 1. First, we prove the following key lemma, which will be used to prove Lemma 3.

Lemma 8 *Let $\bar{\lambda}^{k+1}$ be generated by (SAMA). Then*

$$d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k\hat{\ell}_{\gamma_{k+1}}(\lambda) + \tfrac{1}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda - \hat{\lambda}^k\rangle$$

$$- \left(\tfrac{1}{\eta_k} - \tfrac{\|A\|^2}{2\gamma_{k+1}}\right)\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 - \tfrac{(1-\tau_k)\gamma_{k+1}}{2}\|u^*_{\gamma_{k+1}}(A^\top\bar{\lambda}^k) - \hat{u}^{k+1}\|^2, \tag{4.50}$$

where

$$\begin{aligned}\hat{\ell}_{\gamma_{k+1}}(\lambda) &:= \varphi_{\gamma_{k+1}}(\hat{\lambda}^k) + \langle \nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k), \lambda - \hat{\lambda}^k\rangle + \psi(\lambda) \\ &\leq d_{\gamma_{k+1}}(\lambda) - \tfrac{\gamma_{k+1}}{2}\|u^*_{\gamma_{k+1}}(A^\top\lambda) - \hat{u}^{k+1}\|^2.\end{aligned} \tag{4.51}$$

In addition, for any z, $\gamma_k, \gamma_{k+1} > 0$, the function g^*_γ defined by (4.11) *satisfies*

$$g^*_{\gamma_{k+1}}(z) \leq g^*_{\gamma_k}(z) + (\gamma_k - \gamma_{k+1})b_{\mathcal{U}}(u^*_{\gamma_{k+1}}(z), \bar{u}^c). \tag{4.52}$$

Proof First, it is well-known that SAMA is equivalent to the proximal-gradient step applying to the smoothed dual problem

$$\min_{\lambda}\left\{\varphi_{\gamma_{k+1}}(\lambda) + \psi(\lambda) : \lambda \in \mathbb{R}^n\right\}.$$

This proximal-gradient step can be presented as

$$\bar{\lambda}^{k+1} := \mathrm{prox}_{\eta_k\psi}\left(\hat{\lambda}^k - \eta_k\nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k)\right).$$

We write down the optimality condition of this corresponding minimization problem of this step as

$$0 \in \partial\psi(\bar{\lambda}^{k+1}) + \nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k) + \eta_k^{-1}(\bar{\lambda}^{k+1} - \hat{\lambda}^k).$$

Using this condition and the convexity of ψ, for any $\nabla\psi(\bar{\lambda}^{k+1}) \in \partial\psi(\bar{\lambda}^{k+1})$, we have

$$\begin{aligned}\psi(\bar{\lambda}^{k+1}) &\leq \psi(\lambda) + \langle\nabla\psi(\bar{\lambda}^{k+1}), \bar{\lambda}^{k+1} - \lambda\rangle \\ &= \psi(\lambda) + \langle\nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k), \lambda - \bar{\lambda}^{k+1}\rangle + \eta_k^{-1}\langle\bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \bar{\lambda}^{k+1}\rangle.\end{aligned} \tag{4.53}$$

Next, by the definition $\varphi_\gamma(\lambda) := g^*_\gamma(A^\top\lambda)$, we can show from (4.11) that $\hat{u}^{k+1} = u^*_{\gamma_{k+1}}(A^\top\hat{\lambda}^k)$. Since g^*_γ is $(1/\gamma)$-Lipschitz gradient continuous, we have

$$\frac{\gamma}{2}\|\nabla g^*_\gamma(z) - \nabla g^*_\gamma(\hat{z})\|^2 \leq g^*_\gamma(z) - g^*_\gamma(\hat{z}) - \langle\nabla g^*_\gamma(\hat{z}), z - \hat{z}\rangle \leq \frac{1}{2\gamma}\|z - \hat{z}\|^2.$$

Using this inequality with $\gamma := \gamma_{k+1}$, $\nabla g^*_{\gamma_{k+1}}(A^\top\lambda) = u^*_{\gamma_{k+1}}(A^\top\lambda)$, $\nabla g^*_{\gamma_{k+1}}(A^\top\hat{\lambda}^k) = u^*_{\gamma_{k+1}}(A^\top\hat{\lambda}^k) = \hat{u}^{k+1}$, and $\nabla\varphi_{\gamma_{k+1}}(\lambda) = A\nabla g^*_{\gamma_{k+1}}(A^\top\lambda)$, we have

$$\begin{aligned}\tfrac{\gamma_{k+1}}{2}\|u^*_{\gamma_{k+1}}(A^\top\lambda) - \hat{u}^{k+1}\|^2 &\leq \varphi_{\gamma_{k+1}}(\lambda) - \varphi_{\gamma_{k+1}}(\hat{\lambda}^k) - \langle\nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k), \lambda - \hat{\lambda}^k\rangle \\ &\leq \tfrac{1}{2\gamma_{k+1}}\|A^\top(\lambda - \hat{\lambda}^k)\|^2 \leq \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\lambda - \hat{\lambda}^k\|^2.\end{aligned} \tag{4.54}$$

Using (4.54) with $\lambda = \bar{\lambda}^{k+1}$, we have

$$\varphi_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq \varphi_{\gamma_{k+1}}(\hat{\lambda}^k) + \langle \nabla\varphi_{\gamma_{k+1}}(\hat{\lambda}^k), \bar{\lambda}^{k+1} - \hat{\lambda}^k\rangle + \frac{\|A\|^2}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2.$$

Summing up this inequality and (4.53), then using the definition of $\hat{\ell}_{\gamma_{k+1}}(\lambda)$ in (4.51), we obtain

$$d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq \hat{\ell}_{\gamma_{k+1}}(\lambda) + \tfrac{1}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \hat{\lambda}^k\rangle - \left(\tfrac{1}{\eta_k} - \tfrac{\|A\|^2}{2\gamma_{k+1}}\right)\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2. \tag{4.55}$$

Here, the second inequality in (4.51) follows from the right-hand side of (4.54).

Now, using (4.55) with $\lambda := \bar{\lambda}^k$, then combining with (4.51), we get

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tfrac{1}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \bar{\lambda}^k - \hat{\lambda}^k\rangle - \left(\tfrac{1}{\eta_k} - \tfrac{\|A\|^2}{2\gamma_{k+1}}\right)\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 \\ -\tfrac{\gamma_{k+1}}{2}\|u^*_{\gamma_{k+1}}(A^\top\bar{\lambda}^k) - \hat{u}^{k+1}\|^2. \end{aligned}$$

Multiplying the last inequality by $1 - \tau_k \in [0, 1]$ and (4.55) by $\tau_k \in [0, 1]$, then summing up the results, we obtain (4.50).

Finally, from (4.11), since $g^*_\gamma(z) := \max_u\{P(u, \gamma; z) := \langle z, u\rangle - g(u) - \gamma b_{\mathcal{U}}(u; \bar{u}^c)\}$, is the maximization of P over u indexing in γ and z, which is concave in u and linear in γ, we have $g^*_\gamma(z)$ is convex w.r.t. $\gamma > 0$. Moreover, $\frac{dg^*_\gamma(z)}{d\gamma} = -b_{\mathcal{U}}(u^*_\gamma(z), \bar{u}^c)$. Hence, using the convexity of g^*_γ w.r.t. $\gamma > 0$, we have $g^*_{\gamma_k}(z) \geq g^*_{\gamma_{k+1}}(z) - (\gamma_k - \gamma_{k+1})b_{\mathcal{U}}(u^*_\gamma(z), \bar{u}^c)$, which is indeed (4.52). □

Proof of Lemma 4: Bound on $G_{\gamma\beta}$ for the First Iteration

Since $\bar{w}^1 := (\bar{u}^1, \bar{v}^1, \bar{\lambda}^1)$ is updated by (4.19), similar to (SAMA), we can use (4.55) with $k = 0$, $\lambda := \hat{\lambda}^0$ and $\hat{\ell}_{\gamma_1}(\hat{\lambda}^0) \leq d_{\gamma_1}(\hat{\lambda}^0)$ to obtain

$$d_{\gamma_1}(\bar{\lambda}^1) \leq d_{\gamma_1}(\hat{\lambda}^0) - \left(\frac{1}{\eta_0} - \frac{\|A\|^2}{2\gamma_1}\right)\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2. \tag{4.56}$$

Since $\bar{v}^1$ solves the second problem in (4.19) and $v^*(\hat{\lambda}^0) \in \mathrm{dom}\,(h)$, we have

$$\begin{aligned} h(v^*(\hat{\lambda}^0)) - \langle \hat{\lambda}^0, Bv^*(\hat{\lambda}^0)\rangle + \tfrac{\eta_0}{2}\|A\bar{u}^1 + Bv^*(\hat{\lambda}^0) - c\|^2 \geq h(\bar{v}^1) \\ -\langle \hat{\lambda}^0, B\bar{v}^1\rangle + \tfrac{\eta_0}{2}\|A\bar{u}^1 + B\bar{v}^1 - c\|^2 + \tfrac{\eta_0}{2}\|B(v^*(\hat{\lambda}^0) - \bar{v}^1)\|^2. \end{aligned}$$

Using D_f in (4.9), this inequality implies

$$h^*(B^\top\hat{\lambda}^0) \le \langle \hat{\lambda}^0, B\bar{v}^1\rangle - h(\bar{v}^1) - \frac{\eta_0}{2}\|A\bar{u}^1 + B\bar{v}^1 - c\|^2 + \frac{\eta_0}{2}\|A\bar{u}^1 + B\bar{v}^1 - c\|D_f. \tag{4.57}$$

Using the definition of d_γ, we further estimate (4.56) using (4.57) as follows:

$$\begin{aligned}
d_{\gamma_1}(\bar{\lambda}^1) &\overset{(4.56)}{\le} \varphi_{\gamma_1}(\hat{\lambda}^0) + \psi(\hat{\lambda}^0) - \left(\tfrac{1}{\eta_0} - \tfrac{\|A\|^2}{2\gamma_1}\right)\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2 \\
&\overset{(4.11)}{=} \langle A\bar{u}^1, \hat{\lambda}^0\rangle - g(\bar{u}^1) - \gamma_1 b_{\mathcal{U}}(\bar{u}^1, \bar{u}^c) + \psi(\hat{\lambda}^0) - \left(\tfrac{1}{\eta_0} - \tfrac{\|A\|^2}{2\gamma_1}\right)\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2 \\
&\overset{(4.57)}{\le} \langle \hat{\lambda}^0, A\bar{u}^1 + B\bar{v}^1 - c\rangle - g(\bar{u}^1) - h(\bar{v}^1) - \gamma_1 b_{\mathcal{U}}(\bar{u}^1, \bar{u}^c) \\
&\quad - \tfrac{\eta_0}{2}\|A\bar{u}^1 + B\bar{v}^1 - c\|^2 - \left(\tfrac{1}{\eta_0} - \tfrac{\|A\|^2}{2\gamma_1}\right)\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2 + \tfrac{\eta_0}{2}\|A\bar{u}^1 + B\bar{v}^1 - c\|D_f \\
&\le -f_{\beta_1}(\bar{x}^1) + \tfrac{1}{2\eta_0^2}\left[\tfrac{1}{\beta_1} - \tfrac{5\eta_0}{2} + \tfrac{\|A\|^2\eta_0^2}{\gamma_1}\right]\|\bar{\lambda}^1 - \hat{\lambda}^0\|^2 + \tfrac{1}{\eta_0}\langle\hat{\lambda}^0, \bar{\lambda}^1 - \hat{\lambda}^0\rangle + \tfrac{\eta_0}{4}D_f^2.
\end{aligned}$$

Since $G_{\gamma_1\beta_1}(\bar{w}^1) = f_{\beta_1}(\bar{x}^1) + d_{\gamma_1}(\bar{\lambda}^1)$, we obtain (4.20) from the last inequality. If $\beta_1 \ge \frac{2\gamma_1}{\eta_0(5\gamma_1 - 2\|A\|^2\eta_0)}$, then (4.20) leads to $G_{\gamma_1\beta_1}(\bar{w}^1) \le \frac{\eta_0}{4}D_f^2 + \frac{1}{\eta_0}\langle\hat{\lambda}^0, \bar{\lambda}^1 - \hat{\lambda}^0\rangle$. □

Proof of Lemma 3: Gap Reduction Condition

For notational simplicity, we first define the following abbreviations

$$\begin{cases}
\bar{z}^k & := A\bar{u}^k + B\bar{v}^k - c \\
\hat{z}^{k+1} & := A\hat{u}^{k+1} + B\hat{v}^{k+1} - c \\
\bar{u}^*_{k+1} & := u^*_{\gamma_{k+1}}(A^\top\bar{\lambda}^k) \ \text{ the solution of (4.11) at } \bar{\lambda}^k, \\
\hat{v}^*_k & := v^*(\hat{\lambda}^k) \in \partial h^*(A^\top\hat{\lambda}^k) \ \text{ a subgradient of } h^* \text{ defined by (4.5) at } A^\top\hat{\lambda}^k, \text{ and} \\
D_k & := \|A\hat{u}^{k+1} + B(2\hat{v}^*_k - \hat{v}^{k+1}) - c\|.
\end{cases}$$

From SAMA, we have $\bar{\lambda}^{k+1} - \hat{\lambda}^k = \eta_k(c - A\hat{u}^{k+1} - B\hat{v}^{k+1}) = -\eta_k\hat{z}^{k+1}$. In addition, by (4.16), we have $\hat{\lambda}^k = (1-\tau_k)\bar{\lambda}^k + \tau_k\lambda^*_k$, which leads to $(1-\tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k - \hat{\lambda}^k = \tau_k(\hat{\lambda}^k - \lambda^*_k)$. Using these expressions into (4.50) with $\lambda := \hat{\lambda}^k$, and then using (4.51) with $\hat{\ell}_{\gamma_{k+1}}(\hat{\lambda}^k) \le d_{\gamma_{k+1}}(\hat{\lambda}^k)$, we obtain

$$\begin{aligned}
d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \le (1-\tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \tau_k\langle\hat{z}^{k+1}, \lambda^*_k - \hat{\lambda}^k\rangle \\
-\eta_k\left(1 - \tfrac{\eta_k\|A\|^2}{2\gamma_{k+1}}\right)\|\hat{z}^{k+1}\|^2 - (1-\tau_k)\tfrac{\gamma_{k+1}}{2}\|\bar{u}^*_{k+1} - \hat{u}^{k+1}\|^2.
\end{aligned} \tag{4.58}$$

By (4.52) with the fact that $\varphi_\gamma(\lambda) := g^*_\gamma(A^\top\lambda)$, for any $\gamma_{k+1} > 0$ and $\gamma_k > 0$, we have

$$\varphi_{\gamma_{k+1}}(\bar{\lambda}^k) \le \varphi_{\gamma_k}(\bar{\lambda}^k) + (\gamma_k - \gamma_{k+1})b_{\mathcal{U}}(\bar{u}^*_{k+1}, \bar{u}_c).$$

Using this inequality and the fact that $d_\gamma := \varphi_\gamma + \psi$, we have

$$d_{\gamma_{k+1}}(\bar{\lambda}^k) \le d_{\gamma_k}(\bar{\lambda}^k) + (\gamma_k - \gamma_{k+1}) b_{\mathcal{U}}(\bar{u}^*_{k+1}, \bar{u}_c). \tag{4.59}$$

Next, using $\hat{v}^{k+1}$ from SAMA and its optimality condition, we can show that

$$\begin{aligned} h^*(B^\top\hat{\lambda}^k) - \tfrac{\eta_k}{2}\|A\hat{u}^{k+1} + B\hat{v}^*_k - c\|^2 &= \langle B^\top\hat{\lambda}^k, \hat{v}^*_k\rangle - h(\hat{v}^*_k) - \tfrac{\eta_k}{2}\|A\hat{u}^{k+1} + B\hat{v}^*_k - c\|^2 \\ &\le \langle B^\top\hat{\lambda}^k, \hat{v}^{k+1}\rangle - h(\hat{v}^{k+1}) - \tfrac{\eta_k}{2}\|A\hat{u}^{k+1} + B\hat{v}^{k+1} - c\|^2 - \tfrac{\eta_k}{2}\|B(\hat{v}^*_k - \hat{v}^{k+1})\|^2. \end{aligned}$$

Since $\psi(\lambda) := h^*(B^\top\lambda) - c^\top\lambda$, this inequality leads to

$$\begin{aligned} \psi(\hat{\lambda}^k) &\le \langle B^\top\hat{\lambda}^k, \hat{v}^{k+1}\rangle - \langle c, \hat{\lambda}^k\rangle - h(\hat{v}^{k+1}) - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 \\ &\quad - \tfrac{\eta_k}{2}\langle \hat{z}^{k+1}, A\hat{u}^{k+1} + B(2\hat{v}^*_k - \hat{v}^{k+1}) - c\rangle \\ &\le \langle \hat{\lambda}^k, B\hat{v}^{k+1} - c\rangle - h(\hat{v}^{k+1}) - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\| D_k. \end{aligned}$$

Now, by this estimate, $d_{\gamma_{k+1}} = \varphi_{\gamma_{k+1}} + \psi$ and SAMA, we can derive

$$\begin{aligned} d_{\gamma_{k+1}}(\hat{\lambda}^k) &\le \varphi_{\gamma_{k+1}}(\hat{\lambda}^k) - h(\hat{v}^{k+1}) + \langle \hat{\lambda}^k, B\hat{v}^{k+1} - c\rangle - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\| D_k \\ &= -f(\hat{x}^{k+1}) + \langle \hat{\lambda}^k, \hat{z}^{k+1}\rangle - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\| D_k - \gamma_{k+1} b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}^c). \end{aligned}$$

Combining this inequality, (4.58) and (4.59), we obtain

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) &\le (1-\tau_k) d_{\gamma_k}(\bar{\lambda}^k) - \tau_k f(\hat{x}^{k+1}) + \tau_k\langle \lambda^*_k, \hat{z}^{k+1}\rangle \\ &\quad - \eta_k\left(1 + \tfrac{\tau_k}{2} - \tfrac{\|A\|^2\eta_k}{2\gamma_{k+1}}\right)\|\hat{z}^{k+1}\|^2 \\ &\quad - \tau_k\gamma_{k+1} b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}^c) + (1-\tau_k)(\gamma_k - \gamma_{k+1}) b_{\mathcal{U}}(\bar{u}^*_{k+1}, \bar{u}_c) \\ &\quad - (1-\tau_k)\tfrac{\gamma_{k+1}}{2}\|\bar{u}^*_{k+1} - \hat{u}^{k+1}\|^2 + \tfrac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\| D_k. \end{aligned} \tag{4.60}$$

Now, using the definition G_k, we have

$$\begin{aligned} G_k(\bar{w}^k) := f_{\beta_k}(\bar{x}^k) + d_{\gamma_k}(\bar{\lambda}^k) &= f(\bar{x}^k) + d_{\gamma_k}(\bar{\lambda}^k) + \tfrac{1}{2\beta_k}\|A\bar{u}^k + B\bar{v}^k - c\|^2 \\ &= f(\bar{x}^k) + d_{\gamma_k}(\bar{\lambda}^k) + \tfrac{1}{2\beta_k}\|\bar{z}^k\|^2. \end{aligned}$$

Let us define $\Delta G_k := (1-\tau_k)G_k(\bar{w}^k) - G_{k+1}(\bar{w}^{k+1})$. Then, we can show that

$$\begin{aligned} \Delta G_k = (1-\tau_k)f(\bar{x}^k) &+ (1-\tau_k)d_{\gamma_k}(\bar{\lambda}^k) - f(\bar{x}^{k+1}) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \\ &+ \tfrac{(1-\tau_k)}{2\beta_k}\|\bar{z}^k\|^2 - \tfrac{1}{2\beta_{k+1}}\|\bar{z}^{k+1}\|^2. \end{aligned} \tag{4.61}$$

By (4.16), we have $\bar{z}^{k+1} = (1-\tau_k)\bar{z}^k + \tau_k\hat{z}^{k+1}$. Using this expression and the condition $\beta_{k+1} \ge (1-\tau_k)\beta_k$ in (4.17), we can easily show that

$$\frac{(1-\tau_k)}{2\beta_k}\|\bar{z}^k\|^2 - \frac{1}{2\beta_{k+1}}\|\bar{z}^{k+1}\|^2 \ge -\frac{\tau_k}{\beta_k}\langle \hat{z}^{k+1}, \bar{z}^k\rangle - \frac{\tau_k^2}{2\beta_k(1-\tau_k)}\|\hat{z}^{k+1}\|^2.$$

Substituting this inequality into (4.61), and using the convexity of f, we further get

$$\begin{aligned} \Delta G_k \geq (1-\tau_k) d_{\gamma_k}(\bar{\lambda}^k) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) - \tau_k f(\hat{x}^{k+1}) \\ - \tfrac{\tau_k}{\beta_k}\langle \hat{z}^{k+1}, \bar{z}^k\rangle - \tfrac{\tau_k^2}{2(1-\tau_k)\beta_k}\|\hat{z}^{k+1}\|^2. \end{aligned} \tag{4.62}$$

Substituting (4.60) into (4.62) and using $\lambda_k^* := \frac{1}{\beta_k}(c - A\bar{u}^k - B\bar{v}^k) = -\frac{1}{\beta_k}\bar{z}^k$, we obtain

$$\Delta G_k \geq \left[\eta_k\left(1 + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\gamma_{k+1}}\right) - \frac{\tau_k^2}{2(1-\tau_k)\beta_k}\right]\|\hat{z}^{k+1}\|^2 + R_k - \frac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\|D_k. \tag{4.63}$$

where

$$R_k := \tfrac{1-\tau_k}{2}\gamma_{k+1}\|\bar{u}_{k+1}^* - \hat{u}^{k+1}\|^2 + \tau_k\gamma_{k+1} b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}^c) - (1-\tau_k)(\gamma_k - \gamma_{k+1}) b_{\mathcal{U}}(\bar{u}_{k+1}^*, \bar{u}^c).$$

Furthermore, we have

$$\frac{\eta_k}{4}\|\hat{z}^{k+1}\|^2 - \frac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\|D_k = \frac{\eta_k}{4}\left[\|z^{k+1}\| - \tau_k D_k\right]^2 - \frac{\eta_k\tau_k^2 D_k^2}{4} \geq -\frac{\eta_k\tau_k^2 D_k^2}{4}.$$

Using this estimate into (4.63), we finally get

$$\Delta G_k \geq \left[\eta_k\left(\frac{3}{4} + \frac{\tau_k}{2} - \frac{\|A\|^2\eta_k}{2\gamma_{k+1}}\right) - \frac{\tau_k^2}{2(1-\tau_k)\beta_k}\right]\|\hat{z}^{k+1}\|^2 + R_k - \frac{\eta_k\tau_k^2 D_k^2}{4}. \tag{4.64}$$

Next step, we estimate R_k. Let $\bar{a}_k := \bar{u}_{k+1}^* - \bar{u}_c$, $\hat{a}_k := \hat{u}^{k+1} - \bar{u}_c$. Using the smoothness of $b_{\mathcal{U}}$, we can estimate R_k explicitly as

$$\begin{aligned} 2\gamma_{k+1}^{-1} R_k &\geq (1-\tau_k)\|\bar{a}_k - \hat{a}_k\|^2 - (1-\tau_k)(\gamma_{k+1}^{-1}\gamma_k - 1)L_b\|\bar{a}_k\|^2 + \tau_k\|\hat{a}_k\|^2 \\ &= \|\hat{a}^k - (1-\tau_k)\bar{a}_k\|^2 + (1-\tau_k)\left(\tau_k - (\gamma_{k+1}^{-1}\gamma_k - 1)L_b\right)\|\bar{a}_k\|^2. \end{aligned} \tag{4.65}$$

By the condition $(1 + L_b^{-1}\tau_k)\gamma_{k+1} \geq \gamma_k$ in (4.17), we have $\tau_k - (\gamma_{k+1}^{-1}\gamma_k - 1)L_b \geq 0$. Using this condition in (4.65), we obtain $R_k \geq 0$. Finally, by (4.9) we can show that $D_k \leq D_f$. Using this inequality, $R_k \geq 0$, and the second condition of (4.17), we can show from (4.63) that $\Delta G_k \geq -\frac{\eta_k\tau_k^2}{4}D_f^2$, which implies (4.18). □

Proof of Lemma 5: Parameter Updates

The tightest update for γ_k and β_k is $\gamma_{k+1} := \frac{\gamma_k}{\tau_k+1}$ and $\beta_{k+1} := (1-\tau_k)\beta_k$ due to (4.17). Using these updates in the third condition in (4.17) leads to $\frac{(1-\tau_{k+1})^2}{(1+\tau_{k+1})\tau_{k+1}^2} \geq$

$\frac{1-\tau_k}{\tau_k^2}$. By directly checking this condition, we can see that $\tau_k = \mathcal{O}(1/k)$ which is the optimal choice.

Clearly, if we choose $\tau_k := \frac{3}{k+4}$, then $0 < \tau_k < 1$ for $k \geq 0$ and $\tau_0 = 3/4$. Next, we choose $\gamma_{k+1} := \frac{\gamma_k}{1+\tau_k/3} \geq \frac{\gamma_k}{1+\tau_k}$. Substituting $\tau_k = \frac{3}{k+4}$ into this formula we have $\gamma_{k+1} = \left(\frac{k+4}{k+5}\right)\gamma_k$. By induction, we obtain $\gamma_{k+1} = \frac{5\gamma_1}{k+5}$. This implies $\eta_k = \frac{5\gamma_1}{2\|A\|^2(k+5)}$. With $\tau_k = \frac{3}{k+4}$ and $\gamma_{k+1} = \frac{5\gamma_1}{k+5}$, we choose β_k from the third condition of (4.17) as $\beta_k = \frac{2\|A\|^2\tau_k^2}{(1-\tau_k^2)\gamma_{k+1}} = \frac{18\|A\|^2(k+5)}{5\gamma_1(k+1)(k+7)}$ for $k \geq 1$. Using the value of τ_k and β_k, we need to check the second condition $\beta_{k+1} \geq (1-\tau_k)\beta_k$ of (4.17). Indeed, this condition is equivalent to $2k^2 + 28k + 88 \geq 0$, which is true for all $k \geq 0$. From the update rule of β_k, it is obvious that $\beta_k \leq \frac{18\|A\|^2}{5\gamma_1(k+1)}$. □

Proof of Theorem 1: Convergence of Algorithm 1

We estimate the term $\tau_k^2\eta_k$ in (4.18) as

$$\tau_k^2\eta_k = \frac{45\gamma_1}{2\|A\|^2(k+4)^2(k+5)} < \frac{45\gamma_1}{2\|A\|^2(k+4)(k+5)} - (1-\tau_k)\frac{45\gamma_1}{2\|A\|^2(k+3)(k+4)}.$$

Combing this estimate and (4.18), we get

$$G_{k+1}(\bar{w}^{k+1}) - \frac{45\gamma_1 D_f^2}{8\|A\|^2(k+4)(k+5)} \leq (1-\tau_k)\left[G_k(\bar{w}^k) - \frac{45\gamma_1 D_f^2}{8\|A\|^2(k+3)(k+4)}\right].$$

By induction, we have $G_k(\bar{w}^k) - \frac{45\gamma_1 D_f^2}{8\|A\|^2(k+3)(k+4)} \leq \omega_k[G_1(\bar{w}^1) - \frac{9\gamma_1}{32\|A\|^2}D_f^2] \leq 0$ whenever $G_1(\bar{w}^1) \leq \frac{3\gamma_1}{4\|A\|^2}D_f$, where $\omega_k := \prod_{i=1}^{k-1}(1-\tau_i)$. Hence, we finally get

$$G_k(\bar{w}^k) \leq \frac{45\gamma_1 D_f^2}{8\|A\|^2(k+3)(k+4)}. \tag{4.66}$$

Since $\eta_0 = \frac{\gamma_1}{2\|A\|^2}$, it satisfies the condition $5\gamma_1 > 2\eta_0\|A\|^2$ in Lemma 4. In addition, from Lemma 5, we have $\beta_1 = \frac{27\|A\|^2}{20\gamma_1} > \frac{\|A\|^2}{\gamma_1}$, which satisfies the second condition in Lemma 4. We also note that $\beta_k \leq \frac{18\|A\|^2}{5\gamma_1(k+1)}$. If we take $\hat{\lambda}^0 = \mathbf{0}^m$, then Lemma 4 shows that $G_{\gamma_1\beta_1}(\bar{w}^1) \leq \frac{\eta_0}{2}D_f^2 = \frac{\gamma_1}{4\|A\|^2}D_f^2 < \frac{9\gamma_1}{32\|A\|^2}D_f^2$. Using this estimate and (4.66) into Lemma 2, we obtain (4.23). Finally, if we choose $\gamma_1 := \|A\|$, then we obtain the worst-case iteration-complexity of Algorithm 1 is $\mathcal{O}(\varepsilon^{-1})$. □

Proof of Corollary 1: Strong Convexity of g

First, we show that if condition (4.24) hold, then (4.25) holds. Since $\nabla\varphi$ given by (4.5) is Lipschitz continuous with $L_{d_0^g} := \mu_g^{-1}\|A\|^2$, similar to the proof of Lemma 3, we have

$$\varDelta G_{\beta_k} \geq \left[\eta_k\left(\frac{3}{4}+\frac{\tau_k}{2}-\frac{\eta_k\|A\|^2}{2\mu_g}\right)-\frac{\tau_k^2}{2(1-\tau_k)\beta_k}\right]\|\hat{z}^{k+1}\|^2-\frac{\tau_k^2\eta_k}{4}D_f^2, \tag{4.67}$$

where $\varDelta G_{\beta_k} := (1-\tau_k)G_{\beta_k}(\bar{w}^k)-G_{\beta_{k+1}}(\bar{w}^{k+1})$. Under the condition (4.24), (4.67) implies (4.25).

The update rule (4.27) is in fact derived from (4.24). We finally prove the bounds (4.28). First, we consider the product $\tau_k^2\eta_k$. By (4.27) we have

$$\begin{aligned}\tau_k^2\eta_k &= \frac{9\mu_g}{2\|A\|^2(k+4)^2} < \frac{9\mu_g}{2\|A\|^2(k+3)(k+4)}\\ &= \frac{9\mu_g}{4\|A\|^2(k+4)} - (1-\tau_k)\frac{9\mu_g}{4\|A\|^2(k+3)}\end{aligned}$$

By induction, it follows from (4.25) and this last expression that:

$$G_{\beta_k}(\bar{w}^k) - \frac{9\mu_g D_f^2}{16\|A\|^2(k+3)} \leq \omega_k\left(G_{\beta_1}(\bar{w}^1)-\frac{9\mu_g D_f^2}{64\|A\|^2}\right) \leq 0, \tag{4.68}$$

whenever $G_{\beta_1}(\bar{w}^1) \leq \frac{9\mu_g D_f^2}{64\|A\|^2}$. Since $\bar{u}^1$ is given by (4.26), with the same argument as the proof of Lemma 4, we can show that if $\frac{1}{\beta_1} \leq \frac{5\eta_0}{2} - \frac{\|A\|^2\eta_0^2}{\mu_g}$, then $G_{\beta_1}(\bar{w}^1) \leq \frac{\eta_0}{4}D_f^2$. However, from the update rule (4.27), we can see that $\eta_0 = \frac{\mu_g}{2\|A\|^2}$ and $\beta_1 = \frac{18\|A\|^2}{16\mu_g}$. Using these quantities, we can clearly show that $\frac{1}{\beta_1} \leq \frac{5\eta_0}{2} - \frac{\|A\|^2\eta_0^2}{\mu_g} = \frac{\mu_g}{\|A\|^2}$. Moreover, $G_{\beta_1}(\bar{w}^1) \leq \frac{\eta_0}{4}D_f^2 < \frac{9\mu_g}{64\|A\|^2}D_f^2$. Hence, (4.68) holds. Finally, it remains to use Lemma 2 to obtain (4.28). The second part in (4.30) is proved similarly. The estimate (4.31) is a direct consequence of (4.68). □

Convergence Analysis of Algorithm 2

This appendix provides full proof of Lemmas and Theorems related to the convergence of Algorithm 2.

Proof of Lemma 6: Gap Reduction Condition

We first require the following key lemma to analyze the convergence of our SADMM scheme, whose proof is similar to (4.55) and we omit the details here.

Lemma 9 *Let $\bar{\lambda}^{k+1}$ be generated by SADMM. Then, for $\lambda \in \mathbb{R}^n$, one has*

$$d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq \tilde{\ell}_{\gamma_{k+1}}(\lambda) + \tfrac{1}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \hat{\lambda}^k\rangle - \tfrac{1}{\eta_k}\|\hat{\lambda}^k - \bar{\lambda}^{k+1}\|^2 + \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\tilde{\lambda}^k - \bar{\lambda}^{k+1}\|^2,$$

where $\tilde{\lambda}^k := \hat{\lambda}^k - \rho_k(A\hat{u}^{k+1} + B\hat{v}^k - c)$ and $\tilde{\ell}_\gamma(\lambda) := \varphi_\gamma(\tilde{\lambda}^k) + \langle \nabla\varphi_\gamma(\tilde{\lambda}^k), \lambda - \tilde{\lambda}^k\rangle + \psi(\lambda)$.

Now, we can prove Lemma 6. We still use the same notations as in the proof of Lemma 3. In addition, let us denote by $\hat{u}^*_{k+1} := u^*_{\gamma_{k+1}}(A^\top\hat{\lambda}^k)$ and $\bar{u}^*_{k+1} := u^*_{\gamma_{k+1}}(A^\top\bar{\lambda}^k)$ given in (4.12), $\tilde{z}^k := A\hat{u}^{k+1} + B\hat{v}^k - c$ and $\breve{D}_k := \|A\hat{u}^*_{k+1} + B\hat{v}^k - c\|$.

First, since $\varphi_\gamma(\tilde{\lambda}^k) + \langle \nabla\varphi_\gamma(\tilde{\lambda}^k), \lambda - \tilde{\lambda}^k\rangle \leq \varphi_\gamma(\lambda)$, it follows from Lemma 9 that

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq d_{\gamma_{k+1}}(\lambda) + \tfrac{1}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \lambda - \hat{\lambda}^k\rangle - \tfrac{1}{\eta_k}\|\hat{\lambda}^k - \bar{\lambda}^{k+1}\|^2 \\ + \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\tilde{\lambda}^k - \bar{\lambda}^{k+1}\|^2. \end{aligned} \tag{4.69}$$

Next, using [26, Theorem 2.1.5 (2.1.10)] with g^*_γ defined in (4.11) and $\lambda := (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k$ for any $\tau_k \in [0, 1]$, we have

$$\varphi_{\gamma_{k+1}}(\lambda) \leq (1 - \tau_k)\varphi_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k\varphi_{\gamma_{k+1}}(\hat{\lambda}^k) - \frac{\tau_k(1 - \tau_k)\gamma_{k+1}}{2}\|\hat{u}^*_{k+1} - \bar{u}^*_{k+1}\|^2. \tag{4.70}$$

Since ψ is convex, we also have $\psi(\lambda) \leq (1 - \tau_k)\psi(\bar{\lambda}^k) + \tau_k\psi(\hat{\lambda}^k)$ and $\lambda - \hat{\lambda}^k = (1 - \tau_k)\bar{\lambda}^k + \tau_k\hat{\lambda}^k - \hat{\lambda}^k = \tau_k(\hat{\lambda}^k - \lambda^*_k)$ due to (4.33). Combining these expressions, the definition $d_\gamma := \varphi_\gamma + \psi$, (4.69), and (4.70), we can derive

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq (1 - \tau_k)d_{\gamma_{k+1}}(\bar{\lambda}^k) + \tau_k d_{\gamma_{k+1}}(\hat{\lambda}^k) + \tfrac{\tau_k}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \hat{\lambda}^k - \lambda^*_k\rangle \\ - \tfrac{1}{\eta_k}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 \\ -(1 - \tau_k)\tau_k\tfrac{\gamma_{k+1}}{2}\|\bar{u}^*_{k+1} - \hat{u}^*_{k+1}\|^2. \end{aligned} \tag{4.71}$$

On the one hand, since $\hat{u}^{k+1}$ is the solution of the first convex subproblem in SADMM, using its optimality condition, we can show that

$$\begin{aligned} \varphi_{\gamma_{k+1}}(\hat{\lambda}^k) - \tfrac{\rho_k}{2}\breve{D}_k^2 &= \langle \hat{\lambda}^k, A\hat{u}^*_{k+1}\rangle - g(\hat{u}^*_{k+1}) - \gamma_{k+1}b_{\mathcal{U}}(\hat{u}^*_{k+1}, \bar{u}^c) - \tfrac{\rho_k}{2}\breve{D}_k^2 \\ &\leq \langle \hat{\lambda}^k, A\hat{u}^{k+1}\rangle - g(\hat{u}^{k+1}) - \tfrac{\rho_k}{2}\|\tilde{z}^k\|^2 - \gamma_{k+1}b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}_c) \\ &\quad - \tfrac{\rho_k}{2}\|A(\hat{u}^*_{k+1} - \hat{u}^{k+1})\|^2 - \tfrac{\gamma_{k+1}}{2}\|\hat{u}^*_{k+1} - \hat{u}^{k+1}\|^2. \end{aligned} \tag{4.72}$$

On the other hand, similar to the proof of Lemma 3, we can show that

$$\begin{aligned}\psi(\hat{\lambda}^k) &\leq \langle \hat{\lambda}^k, B\hat{v}^{k+1} - c\rangle - h(\hat{v}^{k+1}) - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 - \tfrac{\eta_k}{2}\langle \hat{z}^{k+1}, A\hat{u}^{k+1} + B(2\hat{v}_k^* - \hat{v}^{k+1}) - c\rangle \\ &\leq \langle \hat{\lambda}^k, B\hat{v}^{k+1} - c\rangle - h(\hat{v}^{k+1}) - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\| D_k. \end{aligned} \tag{4.73}$$

Combining (4.72) and (4.73) and noting that $d_\gamma := \varphi_\gamma + \psi$, we have

$$\begin{aligned} d_{\gamma_{k+1}}(\hat{\lambda}^k) \leq{}& \langle \hat{\lambda}^k, \hat{z}^{k+1}\rangle - f(\hat{x}^{k+1}) - \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\|^2 - \tfrac{\rho_k}{2}\|\tilde{z}^k\|^2 - \gamma_{k+1} b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}_c) \\ & - \tfrac{\rho_k}{2}\|A(\hat{u}^*_{k+1} - \hat{u}^{k+1})\|^2 - \tfrac{\gamma_{k+1}}{2}\|\hat{u}^*_{k+1} - \hat{u}^{k+1}\|^2 + \tfrac{\eta_k}{2}\|\hat{z}^{k+1}\| D_k + \tfrac{\rho_k}{2}\breve{D}_k^2. \end{aligned} \tag{4.74}$$

Next, using the strong convexity of $b_{\mathcal{U}}$ with $\mu_{b_{\mathcal{U}}} = 1$, we can show that

$$\tfrac{\gamma_{k+1}}{2}\|\hat{u}^*_{k+1} - \hat{u}^{k+1}\|^2 + \gamma_{k+1} b_{\mathcal{U}}(\hat{u}^{k+1}, \bar{u}_c) \geq \tfrac{\gamma_{k+1}}{4}\|\hat{u}^*_{k+1} - \bar{u}_c\|^2. \tag{4.75}$$

Combining (4.71), (4.59), (4.74) and (4.75), we can derive

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq{}& (1-\tau_k) d_{\gamma_k}(\bar{\lambda}^k) + \tfrac{\tau_k}{\eta_k}\langle \bar{\lambda}^{k+1} - \hat{\lambda}^k, \hat{\lambda}^k - \lambda_k^*\rangle \\ & - \tfrac{1}{\eta_k}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 \\ & - \tau_k f(\hat{x}^{k+1}) + \tau_k\langle \hat{\lambda}^k, \hat{z}^{k+1}\rangle - \tfrac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\|^2 - \tfrac{\tau_k\rho_k}{2}\|\tilde{z}^k\|^2 \\ & - \tfrac{\tau_k\gamma_{k+1}}{4}\|\hat{u}^*_{k+1} - \bar{u}_c\|^2 - (1-\tau_k)\tau_k\tfrac{\gamma_{k+1}}{2}\|\hat{u}^*_{k+1} - \bar{u}^*_{k+1}\|^2 \\ & + (1-\tau_k)(\gamma_k - \gamma_{k+1}) b_{\mathcal{U}}(\bar{u}^*_{k+1}, \bar{u}^c) + \tfrac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\| D_k + \tfrac{\tau_k\rho_k}{2}\breve{D}_k^2. \end{aligned} \tag{4.76}$$

$$\begin{aligned} \hat{R}_k :={}& \tfrac{\gamma_{k+1}}{2}(1-\tau_k)\tau_k\|\hat{u}^*_{k+1} - \bar{u}^*_{k+1}\|^2 + \tfrac{\gamma_{k+1}}{4}\tau_k\|\hat{u}^*_{k+1} - \bar{u}_c\|^2 \\ & - (1-\tau_k)(\gamma_k - \gamma_{k+1}) b_{\mathcal{U}}(\bar{u}^*_{k+1}, \bar{u}^c). \end{aligned} \tag{4.77}$$

From SADMM, we have $\bar{\lambda}^{k+1} - \hat{\lambda}^k = -\eta_k\hat{z}^{k+1}$ and $\tilde{\lambda}^k - \hat{\lambda}^k = -\rho_k\tilde{z}^k$. Plugging these expressions and (4.77) into (4.76) we can simplify this estimate as

$$\begin{aligned} d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq{}& (1-\tau_k) d_{\gamma_k}(\bar{\lambda}^k) + \tau_k\langle \hat{z}^{k+1}, \lambda_k^*\rangle - \tau_k f(\hat{x}^{k+1}) - \tfrac{(1+\tau_k)\eta_k}{2}\|\hat{z}^{k+1}\|^2 \\ & - \tfrac{1}{\eta_k}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \tfrac{\|A\|^2}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 - \tfrac{\tau_k}{2\rho_k}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 - \hat{R}_k \\ & + \tfrac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\| D_k + \tfrac{\tau_k\rho_k}{2}\breve{D}_k^2. \end{aligned} \tag{4.78}$$

Using again the elementary inequality $\nu\|a\|^2 + \kappa\|b\|^2 \geq \frac{\nu\kappa}{\nu+\kappa}\|a-b\|^2$, under the condition $\gamma_{k+1} \geq \|A\|^2\left(\eta_k + \frac{\rho_k}{\tau_k}\right)$ in (4.34), we can show that

$$\frac{1}{2\eta_k}\|\bar{\lambda}^{k+1} - \hat{\lambda}^k\|^2 + \frac{\tau_k}{2\rho_k}\|\tilde{\lambda}^k - \hat{\lambda}^k\|^2 - \frac{\|A\|^2}{2\gamma_{k+1}}\|\bar{\lambda}^{k+1} - \tilde{\lambda}^k\|^2 \geq 0. \tag{4.79}$$

On the other hand, similar to the proof of Lemma 3, we can show that $\frac{\eta_k}{4}\|\hat{z}^{k+1}\|^2 - \frac{\tau_k\eta_k}{2}\|\hat{z}^{k+1}\|D_k \geq -\frac{\eta_k\tau_k^2}{4}D_k^2$. Using this inequality, (4.79), and $\lambda_k^* = -\frac{1}{\beta_k}\bar{z}^k$, we can simplify (4.78) as

$$d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) \leq (1-\tau_k)d_{\gamma_k}(\bar{\lambda}^k) - \frac{\tau_k}{\beta_k}\langle\hat{z}^{k+1}, \bar{z}^k\rangle - \tau_k f(\hat{x}^{k+1}) - \eta_k\left(\frac{1}{4}+\frac{\tau_k}{2}\right)\|\hat{z}^{k+1}\|^2 \\ - \hat{R}_k + \left(\frac{\eta_k\tau_k^2}{4}D_k^2 + \frac{\tau_k\rho_k}{2}\breve{D}_k^2\right). \tag{4.80}$$

Since $\beta_{k+1} \geq (1-\tau_k)\beta_k$ due to (4.34), similar to the proof of (4.62) we have

$$\begin{aligned} \Delta G_k \geq{} & (1-\tau_k)d_{\gamma_k}(\bar{\lambda}^k) - d_{\gamma_{k+1}}(\bar{\lambda}^{k+1}) - \tau_k f(\hat{x}^{k+1}) \\ & - \frac{\tau_k}{\beta_k}\langle\hat{z}^{k+1}, \bar{z}^k\rangle - \frac{\tau_k^2}{2(1-\tau_k)\beta_k}\|\hat{z}^{k+1}\|^2. \end{aligned} \tag{4.81}$$

Combining (4.80) and (4.81), we get

$$\Delta G_k \geq \frac{1}{2}\left[\left(\frac{1}{2}+\tau_k\right)\eta_k - \frac{\tau_k^2}{(1-\tau_k)\beta_k}\right]\|\hat{z}^{k+1}\|^2 + \hat{R}_k - \left(\frac{\eta_k\tau_k^2}{4}D_k^2 + \frac{\tau_k\rho_k}{2}\breve{D}_k^2\right). \tag{4.82}$$

Next, we estimate $\hat{R}_k$ defined by (4.77) as follows. We define $\bar{a}_k := \bar{u}_{k+1}^* - \bar{u}_c$, $\hat{a}_k := \hat{u}_{k+1}^* - \bar{u}_c$. Using $b_{\mathcal{U}}(\bar{u}_{k+1}^*, \bar{u}^c) \leq \frac{L_b}{2}\|\bar{u}_{k+1}^* - \bar{u}^c\|^2$, we can write $\hat{R}_k$ explicitly as

$$\begin{aligned} \frac{2\hat{R}_k}{\gamma_{k+1}} &= (1-\tau_k)\tau_k\|\bar{a}_k - \hat{a}_k\|^2 + \frac{\tau_k}{2}\|\hat{a}_k\|^2 - (1-\tau_k)\big(\frac{\gamma_k}{\gamma_{k+1}} - 1\big)L_b\|\bar{a}_k\|^2 \\ &= \tau_k\left(\frac{3}{2}-\tau_k\right)\left\|\hat{a}_k - \frac{(1-\tau)}{(3/2-\tau_k)}\bar{a}_k\right\|^2 + (1-\tau_k)\left[\frac{\tau_k}{3-2\tau_k} + \left(1-\frac{\gamma_k}{\gamma_{k+1}}\right)L_b\right]\|\bar{a}\|^2. \end{aligned}$$

Since $\gamma_{k+1} \geq \left(\frac{3-2\tau_k}{3-(2-L_b^{-1})\tau_k}\right)\gamma_k$ due to (4.34), it is easy to show that $\hat{R}_k \geq 0$. In addition, by (4.34), we also have $(1+2\tau_k)\eta_k - \frac{2\tau_k^2}{(1-\tau_k)\beta_k} \geq 0$. Using these conditions, we can show from (4.82) that $\Delta G_k \geq -\frac{\eta_k\tau_k^2}{4}D_k^2 - \frac{\tau_k\rho_k}{2}\breve{D}_k^2 \geq -\left(\frac{\tau_k^2\eta_k}{4} + \frac{\tau_k\rho_k}{2}\right)D_f^2$, which is indeed the gap reduction condition (4.35). □

Proof of Lemma 7: Parameter Updates

Similar to the proof of Lemma 5, we can show that the optimal rate of $\{\tau_k\}$ is $\mathcal{O}(1/k)$. From the conditions (4.34), it is clear that if we choose $\tau_k := \frac{3}{k+4}$ then $0 < \tau_k \leq \frac{3}{4} < 1$ for $k \geq 0$. Next, we choose $\gamma_{k+1} := \left(\frac{3-2\tau_k}{3-\tau_k}\right)\gamma_k$. Then γ_k

satisfies (4.34). Substituting $\tau_k = \frac{3}{k+4}$ into this formula we have $\gamma_{k+1} = \left(\frac{k+2}{k+3}\right)\gamma_k$. By induction, we obtain $\gamma_{k+1} = \frac{3\gamma_1}{k+3}$. Now, we choose $\eta_k := \frac{\gamma_{k+1}}{2\|A\|^2} = \frac{3\gamma_1}{2\|A\|^2(k+3)}$. Then, from the last condition of (4.34), we choose $\rho_k := \frac{\tau_k\gamma_{k+1}}{2\|A\|^2} = \frac{9\gamma_1}{2\|A\|^2(k+3)(k+4)}$.

To derive an update for β_k, from the third condition of (4.34) with equality, we can derive $\beta_k = \frac{2\tau_k^2}{(1-\tau_k)(1+2\tau_k)\eta_k} = \frac{6\|A\|^2(k+3)}{\gamma_1(k+1)(k+10)} < \frac{9\|A\|^2}{5\gamma_1(k+1)}$. We need to check the second condition $\beta_{k+1} \geq (1-\tau_k)\beta_k$ in (4.34). Indeed, we have $\beta_{k+1} = \frac{6\|A\|^2(k+4)}{\gamma_1(k+2)(k+11)} \geq (1-\tau_k)\beta_k = \frac{6\|A\|^2(k+3)}{\gamma_1(k+1)(k+10)}$, which is true for all $k \geq 0$. Hence, the second condition of (4.34) holds. □

Proof of Theorem 2: Convergence of Algorithm 2

First, we check the conditions of Lemma 4. From the update rule (4.36), we have $\eta_0 = \frac{\gamma_1}{2\|A\|^2}$ and $\beta_1 = \frac{12\|A\|^2}{11\gamma_1}$. Hence, $5\gamma_1 = 10\|A\|^2\eta_0 > 2\|A\|^2\eta_0$, which satisfies the first condition of Lemma 4. Now, $\frac{2\gamma_1}{(5\gamma_1 - 2\eta_0\|A\|^2)\eta_0} = \frac{\|A\|^2}{\gamma_1} < \frac{12\|A\|^2}{11\gamma_1} = \beta_1$. Hence, the second condition of Lemma 4 holds.

Next, since $\tau_k = \frac{3}{k+4}$, $\rho_k = \frac{9\gamma_1}{2\|A\|^2(k+3)(k+4)}$ and $\eta_k = \frac{3\gamma_1}{2\|A\|^2(k+3)}$, we can derive

$$\begin{aligned}\frac{\tau_k^2\eta_k}{4} + \frac{\tau_k\rho_k}{2} &= \frac{81\gamma_1}{8\|A\|^2(k+3)(k+4)^2}\\ &< \frac{81\gamma_1}{8\|A\|^2(k+3)(k+4)} - (1-\tau_k)\,\frac{81\gamma_1}{8\|A\|^2(k+2)(k+3)}.\end{aligned}$$

Substituting this inequality into (4.35) and rearrange the result we obtain

$$G_{k+1}(\bar{w}^{k+1}) - \frac{81\gamma_1 D_f^2}{8\|A\|^2(k+3)(k+4)} \leq (1-\tau_k)\Big[G_k(\bar{w}^k) - \frac{81\gamma_1 D_f^2}{8\|A\|^2(k+2)(k+3)}\Big].$$

By induction, we obtain $G_k(\bar{w}^k) - \frac{81\gamma_1 D_f^2}{8\|A\|^2(k+2)(k+3)} \leq \omega_k\Big[G_0(\bar{w}^0) - \frac{27\gamma_1 D_f^2}{16\|A\|^2}\Big] \leq 0$ as long as $G_0(\bar{w}^0) \leq \frac{27\gamma_1 D_f^2}{16\|A\|^2}$. Now using Lemma 4, we have $G_0(\bar{w}^0) \leq \frac{\eta_0}{4}D_f^2 = \frac{\gamma_1}{8\|A\|^2}D_f^2 < \frac{27\gamma_1 D_f^2}{16\|A\|^2}$. Hence, $G_k(\bar{w}^k) \leq \frac{27\gamma_1 D_f^2}{16\|A\|^2(k+2)(k+3)}$.

Finally, by using Lemma 2 with $\beta_k := \frac{6\|A\|^2(k+3)}{\gamma_1(k+1)(k+10)}$ and $\beta_k \leq \frac{9\|A\|^2}{5\gamma_1(k+1)}$, and simplifying the results, we obtain the bounds in (4.37). If we choose $\gamma_1 := \|A\|$ then, we obtain the worst-case iteration-complexity of Algorithm 2 is $\mathcal{O}(\varepsilon^{-1})$. □

References

1. A. Alotaibi, P.L. Combettes, N. Shahzad, Best approximation from the Kuhn-Tucker set of composite monotone inclusions. Numer. Funct. Anal. Optim. **36**(12), 1513–1532 (2015)
2. H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, Berlin, 2011)
3. A. Beck, M. Teboulle, A fast dual proximal gradient algorithm for convex minimization and applications. Oper. Res. Lett. **42**(1), 1–6 (2014)
4. J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for non-convex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
5. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
6. R.S. Burachik, V. Martín-Márquez, An approach for the convex feasibility problem via monotropic programming. J. Math. Anal. Appl. **453**(2), 746–760 (2017)
7. X. Cai, D. Han, X. Yuan, 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**(1), 39–73 (2017)
8. E. Candès, B. Recht, Exact matrix completion via convex optimization. Commun. ACM **55**(6), 111–119 (2012)
9. V. Cevher, S. Becker, M. Schmidt, Convex optimization for big data: scalable, randomized, and parallel algorithms for big data analytics. IEEE Signal Process. Mag. **31**(5), 32–43 (2014)
10. A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
11. D. Davis, W. Yin, Convergence rate analysis of several splitting schemes, in *Splitting Methods in Communication, Imaging, Science, and Engineering* (Springer, Cham, 2016), pp. 115–163
12. D. Davis, W. Yin, Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. Math. Oper. Res. **42**(3), 783–805 (2017)
13. D. Davis, W. Yin, A three-operator splitting scheme and its optimization applications. Tech. Report. (2015)
14. W. Deng, W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
15. J. Eckstein, D. Bertsekas, On the Douglas Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
16. E. Ghadimi, A. Teixeira, I. Shames, M. Johansson, Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems. IEEE Trans. Autom. Control **60**(3), 644–658 (2015)
17. T. Goldstein, B. ODonoghue, S. Setzer, Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**(3), 1588–1623 (2012)
18. B. He, X. Yuan, On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. Numer. Math. **130**(3), 567–577 (2012)
19. B. He, X. Yuan, On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
20. T. Lin, S. Ma, S. Zhang, On the global linear convergence of the ADMM with multi- block variables. SIAM J. Optim. **25**(3), 1478–1497 (2015)
21. T. Lin, S. Ma, S. Zhang, Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. J. Sci. Comput. **69**(1), 52–81 (2016)
22. T. Lin, S. Ma, S. Zhang, An extragradient-based alternating direction method for convex minimization. Found. Comput. Math. **17**(1), 35–59 (2017)
23. I. Necoara, J. Suykens, Applications of a smoothing technique to decomposition in convex optimization. IEEE Trans. Autom. Control **53**(11), 2674–2679 (2008)

24. A. Nemirovskii, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. **15**(1), 229–251 (2004)
25. A. Nemirovskii, D. Yudin, *Problem Complexity and Method Efficiency in Optimization* (Wiley Interscience, New York, 1983)
26. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization, vol. 87 (Kluwer Academic Publishers, Norwell, 2004)
27. Y. Nesterov, Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
28. Y. Ouyang, Y. Chen, G. Lan, E.J. Pasiliao, An accelerated linearized alternating direction method of multiplier. SIAM J. Imaging Sci. **8**(1), 644–681 (2015)
29. N. Parikh, S. Boyd, Proximal algorithms. Found. Trends Optim. **1**(3), 123–231 (2013)
30. R.T. Rockafellar, *Convex Analysis*. Princeton Mathematics Series, vol. 28 (Princeton University Press, Princeton, 1970)
31. R. Shefi, M. Teboulle, Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**(1), 269–297 (2014)
32. R. Shefi, M. Teboulle, On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. EURO J. Comput. Optim. **4**(1), 27–46 (2016)
33. F. Simon, R. Holger, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013)
34. M. Tao, X. Yuan, On the $O(1/t)$-convergence rate of alternating direction method with logarithmic-quadratic proximal regularization. SIAM J. Optim. **22**(4), 1431–1448 (2012)
35. Q. Tran-Dinh, V. Cevher, Constrained convex minimization via model-based excessive gap, in *Proceedings of the Neural Information Processing Systems (NIPS)*, Montreal, vol. 27 Dec. 2014, pp. 721–729
36. Q. Tran-Dinh, O. Fercoq, V. Cevher, A smooth primal-dual optimization framework for nonsmooth composite convex minimization. SIAM J. Optim. **28**, 96–134 (2018)
37. P. Tseng, D. Bertsekas, Relaxation methods for problems with strictly convex cost and linear constraints. Math. Oper. Res. **16**(3), 462–481 (1991)
38. W. Wang, A. Banerjee, Bregman alternating direction method of multipliers, in *Advances in Neural Information Processing Systems* 27 (NIPS 2014), pp. 1–9
39. E. Wei, A. Ozdaglar, On the $O(1/k)$-convergence of asynchronous distributed alternating direction method of multipliers, in *Global Conference on Signal and Information Processing (GlobalSIP)* (IEEE, Piscataway, 2013), pp. 551–554